Logic and Interactive RAtionality

Yearbook 2011

The printing of this book was supported by the UP fund of Johan van Benthem.
Cover design by Nina Gierasimczuk.

Foreword

It is our pleasure to present you with the 2011 edition of the "Logic and Interactive Rationality" (LIRa) Yearbook. Already in its fourth year, the Yearbook mirrors the activities of the LIRa seminar held regularly at the ILLC in Amsterdam. It also includes contributions stemming from an on-going cooperation with colleagues at the Universities of Groningen, Tilburg and Utrecht, where several special LIRa sessions have been organized and hosted in the past year. Moreover, the Yearbook features a number of invited papers by colleagues working in China, Japan and the US, thus reflecting the global nature of research into interactive rationality and logical dynamics.

We want to thank all the contributors for allowing us to include the fruits of their work. We also would like to express our gratitude to Nina Gierasimczuk—her beautiful cover designs have become a trademark of the Yearbook over the years. We thank Fernando Velázquez-Quesada for help with various LaTeX-related issues. Finally, we would like to thank Johan van Benthem, sponsor of and source of inspiration for the seminar's activities.

Amsterdam
July 23, 2012

Alexandru Baltag
Davide Grossi
Alexandru Marcoci
Ben Rodenhäuser
Sonja Smets
(eds.)

Contents

Preface

Johan van Benthem

Research lines are like those enticing hiking trails that lure us far and wide into Nature, always across the next ridge into the next valley. Often they do not run in the direction that we had planned, and often also, they present us with surprising encounters. Every year, the Dynamics Yearbook shows where the trails in our community have taken us, and whom we met on the way. In this volume, you will find a variety of fresh perspectives when walking along major map lines in logical information dynamics for rational agents. Topics including the nature of explicit non-omniscient beliefs, the creation of strategies in parallel games, general patterns in interactive rationality, agents' attention horizons in games, information-based coalitional influence in epistemic models, and richer views of choice, preference, obligation in social scenarios all the way up to the cathartic finale of voting. But you also get to read about new developments of a more fundamental nature. Logical tools do not get blunt in the course of applications, they often get shinier and even sharper with new edges. Examples of such new gloss and cutting edge show in this Yearbook with papers on new substitution-closed versions of public announcement logic, frame correspondence analysis of update postulates, generalized algebraic views of logical dynamics, and new interfaces with categorial and linear logics as a resource-conscious paradigm of information. So much for the trails. Which strangers did we meet on the way? This Yearbook contains several interesting encounters with colleagues in the empirical domain of natural language, with papers

ranging from categorial grammar to the semantics of quantifiers and speech act theory. And also, we find contributions from the computational worlds of epistemic planning and the situation calculus in AI, drawing attention to the congenial research in communities a few valleys further down the trail.

It is a pleasure to see the wide range of contributions to this Yearbook, which is fast becoming a document linking various active research sites in Europe, the US and Asia. Currently, discussions are underway to turn the next edition of the Yearbook into a truly international venture, with an editorial team spanning three continents. The websites loriweb.org, www.illc.uva.nl/dg and stanford.edu/~thoshi/ldl/Home.html give further information about the topography, the events, and the inhabitants that make this landscape such an enticing destination.

Johan van Benthem

July 2012

On the Power of Knowledge

Thomas Ågotnes, Wiebe van der Hoek and Michael Wooldridge

University of Bergen, University of Liverpool, Oxford University
thomas.agotnes@infomedia.uib.no, Wiebe.Van-Der-Hoek@liverpool.ac.uk,
mjw@cs.ox.ac.uk

Abstract

In epistemic logic, Kripke structures are used to model the distribution of information in a multi-agent system. In this paper, we present an approach to *quantifying* how much information each particular agent in a system has, or how important the agent is, with respect to some fact represented as a goal formula. It is typically the case that the goal formula is distributed knowledge in the system, but that no individual agent alone knows it. It might be that several different groups of agents can get to know the goal formula together by combining their individual knowledge. By using power indices developed in voting theory, such as the Banzhaf index, we get a measure of how important an agent is in such groups. We analyse the properties of this notion of information-based power in detail, and characterise the corresponding class of voting games. Although we mainly focus on distributed knowledge, we also look at variants of this analysis using other notions of group knowledge. An advantage of our framework is that power indices and other power properties can be expressed in standard epistemic logic. This allows, e.g., standard model checkers to be used to quantitatively analyse the distribution of information in a given Kripke structure.

1 Introduction

Epistemic logic is widely used in the multi-agent systems community to reason about the knowledge and ignorance of agents in terms of the information they possess Fagin et al. (1995). In many situations, it would be useful to be able to *quantify* how information is distributed in a system, or to reason about the *relative importance* of the information that different agents have. In general, it is difficult to answer the question of whether an agent has more information than another agent except for in special cases, such as when one agent knows everything another agent knows Van Ditmarsch et al. (2009). In this paper, we quantify the distribution of information in a system in a specific sense satisfying two assumptions. The first is that we are interested in who knows more *about* some given fact. The second is that we are interested in situations where information can be *communicated* between agents, and it is not always possible or desirable to communicate with every other agent in the system.

Consider the following situation. M knows that if sales are up this quarter, the stock price will increase ($p \implies q$). T knows that if the new CEO has signed the contract, the stock price will increase ($r \implies q$). W knows that sales are up this quarter and that the new CEO has signed the contract ($p \wedge r$). Assume that this describes all (relevant) facts that the three agents know. Who knows more? We are here interested in a more specific type of question: who has the most *important* or *valuable* information *about* whether or not the stock price will increase (q), *in* a social setting where communication is possible? None of the agents alone knows q, but they can *combine* their knowledge to find out that q is in fact true. And here the importance of the knowledge of the three agents differ: M and W can together find out q, as can T and W. M and T cannot. It can thus be argued that W knows more about q in this social setting, since he can combine his knowledge in several different ways with others' knowledge – and, indeed, it is not hard to see that W's knowledge is *necessary* for any group to be able to find out q, unlike that of M or T. If it is important for each individual agent to find out q, and since no agent already knows q, the only possibility is to communicate with someone else; in which case clearly W would be considered the most *important* agent.

In this paper we analyse the relative importance of the knowledge each agent has in a system where information about some fact or objective (q in our example above) is distributed throughout the system. To this end, we employ *power indices* such as the Banzhaf index, known from voting theory. The starting

point is a pointed Kripke structure. It is typically the case that the objective is distributed knowledge in the system, but that no individual agent knows it. It might be that several different groups of agents can get to know the objective by combining their knowledge. Our approach measures the importance of an agent in an arbitrary group of agents wrt. deriving the objective. We consider an agent to be powerful, or to have important information, if the probability of changing the distributed knowledge in the group from ignorance to knowledge about the objective by joining some arbitrary group, is high. This concept of *information based power* can, e.g., be used to identify agents that are crucial to the functioning of the multi-agent system.

The question of "who knows more" in epistemic logic has recently been studied in Van Ditmarsch et al. (2009). The notion of information based power we introduce in this paper is a more fine-grained generalisation: if an agent knows more in the sense of Van Ditmarsch et al. (2009) then she has a higher power index, but not necessarily the other way around. Solution concepts for coalitional games have recently been used to measure the degree of inconsistency in databases Hunter and Konieczny (2010). In Ågotnes et al. (2009) power indices are used to analyse the relative importance of agents when in terms of complying or not complying with a *normative system* defined over a Kripke-like structure Shoham and Tennenholtz (1992), Ågotnes et al. (2007). However, we are not aware of any approaches using power indices to measure relative importance of agents in terms of their knowledge/information as described by a Kripke structure.

The paper is organised as follows. In the two next sections we briefly review some background material about epistemic logic and power indices that we will use. In Section 4 we define power indices for agents, given a pointed Kripke structure and a goal formula. We give a complete characterisation of the power indices that can be obtained in this way, study their properties in detail, and show how standard epistemic logic can be used to express power properties. Since these power properties can be expressed in epistemic logic, we can also use epistemic logic to reason about agents' *knowledge* about such properties. In Section 5 we study what agents know about the distribution of information-based power in the system. In most of the paper we use distributed knowledge to define power, but in Section 6 we discuss other types of group knowledge as well. We conclude in Section 7.

2 Epistemic Logic

Assume a finite set of agents $Ag = \{1, \ldots, n\}$ and a countably infinite set of atomic propositions Θ. The language $\mathcal{L}_K$ of the epistemic logic $S5_n$ is defined by the following grammar:

$$\phi ::= \top \mid p \mid K_i\phi \mid \neg\phi \mid \phi_1 \wedge \phi_2$$

where $p \in \Theta$ and $i \in Ag$. An *epistemic (Kripke) structure*, M, (over Ag, Θ) is an $(n+2)$-tuple Fagin et al. (1995):

$$M = \langle W, \sim_1, \ldots, \sim_n, \pi \rangle, \quad \text{where}$$

- W is a finite, non-empty set of *states*;
- $\sim_i \subseteq W \times W$ is an *epistemic accessibility relation* for each agent $i \in Ag$, where each $\sim_i$ is an equivalence relation; and
- $\pi : W \to 2^{\Theta}$ is a Kripke valuation function, which gives the set of primitive propositions satisfied in each state.

Formulae are interpreted in a *pointed structure*, a pair M, s, where M is a model and s is a state in M, as follows.

- $M, s \models \top$
- $M, s \models p$ iff $p \in \pi(s)$ (where $p \in \Theta$)
- $M, s \models \neg\phi$ iff $M, s \not\models \phi$
- $M, s \models \phi \ \& \ \psi$ iff $M, s \models \phi$ and $M, s \models \psi$
- $M, s \models K_i\phi$ iff for all t such that $s \sim_i t$, $M, t \models \phi$.

We will make use of extensions of $S5_n$ with *group knowledge*. To this end, when $G \subseteq Ag$, we denote the union of G's accessibility relations by $\sim_G^E$, so $\sim_G^E = (\bigcup_{i \in G} \sim_i)$. We use $\sim_G^C$ to denote the transitive closure of $\sim_G^E$. Finally, $\sim_G^D$ denotes the intersection of G's accessibility relations (cf. (Fagin et al. 1995, p.66–70)). The logics $S5_n^D$, $S5_n^C$ and $S5_n^{CD}$ are obtained as follows. The respective languages, $\mathcal{L}_D$, $\mathcal{L}_C$, and $\mathcal{L}_{CD}$, are obtained by adding the clause $D_G\phi$, $C_G\phi$, and both, respectively, where $G \subseteq Ag$, to the definition of $\mathcal{L}_K$. The interpretation of the two group operators:

- $M, s \models D_G\phi$ iff for all t such that $s \sim_G^D t$, $M, t \models \phi$
- $M, s \models C_G\phi$ iff for all t such that $s \sim_G^C t$, $M, t \models \phi$

We use the same notation for the satisfaction relation for all these logics; it will be clear from context which logic we are working in. As usual, we write $M \models \phi$ if $M, s \models \phi$ for all s in M, and $\models \phi$ if $M \models \phi$ for all M; in this latter case, we say that ϕ is *valid*. A formula is *satisfied* in a pointed model if it is true. When Φ is a set of formulae, $\Phi \models \phi$, Φ *entails* ϕ, means that any pointed model that satisfies Φ also satisfies ϕ. A formula is *satisfiable* if there exists a pointed model that satisfies it. A formula or set of formulae is satisfiable in a *set* of pointed models if it is satisfied by *at least one* pointed model in that set. The usual propositional abbreviations are used, in addition to $E_G\phi$ ($G \subseteq Ag$) for $\bigwedge_{i \in G} K_i\phi$; $\hat{K}_i\phi$ for $\neg K_i \neg \phi$; $\hat{D}_G\phi$ for $\neg D_G \neg \phi$ and $\hat{C}_G\phi$ for $\neg C_G \neg \phi$. We will often abuse notation and write singleton sets of agents $\{i\}$ as i.

$E_G\varphi$ means that all individuals in the group G know φ. $D_G\varphi$ means that φ is distributed knowledge among G. Roughly speaking, this knowledge would come about if all members of G were to share their information (but see also Section 4.2). $C_G\varphi$, that φ is common knowledge in G, means that $E_G\varphi \wedge E_G E_G\varphi \wedge E_G E_G E_G\varphi \wedge \ldots$. These concepts of group and individual knowledge are related as follows (with $i \in G$):

$$\models (C_G\varphi \rightarrow E_G\varphi) \wedge (E_G\varphi \rightarrow K_i\varphi) \wedge (K_i\varphi \rightarrow D_G\varphi) \wedge (D_G\varphi \rightarrow \varphi)$$

The above implications express that common knowledge is the strongest property, and truth the weakest. However, since $C_G\varphi$ is such a strong notion, this often means it will only be obtained for 'weak' φ. Or Fagin et al. (1995), common knowledge can be paraphrased as what 'any fool knows', while distributed knowledge corresponds to what 'a wise man knows'.

Finally, the *knowledge set* of $G \subseteq Ag$ in M, s is:

$$\mathcal{K}_G(M, s) = \{\phi \in \mathcal{L}_K : M, s \models K_i\phi \text{ for some } i \in G\}$$

3 Coalitional Games and Power

We briefly review some key concepts from the area of cooperative game theory Osborne and Rubinstein (1994) and the theory of voting power Felsenthal

and Machover (1998) that we will use in the following. A *cooperative* (or *coalitional*) *game* is a pair $\Gamma = \langle Ag, v \rangle$, where $Ag = \{1, \ldots, n\}$ is a set of *players*, or *agents*, and $v : 2^{Ag} \rightarrow \mathbb{R}$ is the *characteristic function* of the game, which assigns to every set of agents a numeric value, which is conventionally interpreted as the value that this group of agents could obtain if they chose to cooperate. A cooperative game is said to be *simple* if the range of v is $\{0, 1\}$; in simple games we say that G are *winning* if $v(G) = 1$, while if $v(G) = 0$, we say G are *losing*. A simple cooperative game is said to be *monotonic* if $v(G) = 1$ implies that $v(H) = 1$, whenever $G \subseteq H$. A monotonic simple cooperative game is sometimes called a *simple voting game* Felsenthal and Machover (1998). For simple games, a number of *power indices* attempt to characterise in a systematic way the *influence* that a given agent has, by measuring how effective this agent is at turning a losing coalition into a winning coalition Felsenthal and Machover (1998). The best-known of these is perhaps the *Banzhaf index* and its relatives, the Banzhaf score and Banzhaf measure Banzhaf III (1965).

Agent i is said to be a *swing player* for G if G is not winning but $G \cup \{i\}$ is. We define a function $swing(G, i)$ so that this function returns 1 if i is a swing player for G, and 0 otherwise, i.e.,

$$swing(G, i) = \begin{cases} 1 & \text{if } v(G) = 0 \text{ and } v(G \cup \{i\}) = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Now, we define the *Banzhaf score* for agent i, denoted σ_i, to be the number of coalitions for which i is a swing player:

$$\sigma_i = \sum_{G \subseteq Ag \setminus \{i\}} swing(G, i). \tag{1}$$

The *Banzhaf measure* μ_i, is the probability that i would be a swing player for a coalition chosen at random from $2^{Ag \setminus \{i\}}$:

$$\mu_i = \frac{\sigma_i}{2^{n-1}} \tag{2}$$

The *Banzhaf index* for a player $i \in Ag$, denoted by β_i, is the proportion of coalitions for which i is a swing to the total number of swings in the game – thus the Banzhaf index is a measure of relative power, since it takes into account the Banzhaf score of other agents:

$$\beta_i = \frac{\sigma_i}{\sum_{j \in Ag} \sigma_j} \tag{3}$$

Finally, we define the *Shapley-Shubik index*; here the *order* in which agents join a coalition plays a role. Let $P(Ag)$ denote the set of all permutations of Ag, with typical members ω, ω', etc. If $\omega \in P(Ag)$ and $i \in Ag$, then let $prec(i, \omega)$ denote the members of Ag that precede i in the ordering ω. Given this, let ς_i denote the Shapley-Shubik index of i, defined as follows:

$$\varsigma_i = \frac{1}{|Ag|!} \sum_{\omega \in P(Ag)} swing(prec(i, \omega), i) \tag{4}$$

Thus the Shapley-Shubik index is essentially the Shapley value (Osborne and Rubinstein 1994, p.291) applied to simple ($\{0, 1\}$-valued) cooperative games.

We say that a player is a *veto player* if it is included in all winning coalitions, a *dictator* if $\mu_i = 1$, and a *dummy* if $\mu_i = 0$.

4 Power of Distributed Knowledge

We define the power of agents given a pointed Kripke structure, and an objective specified as a *goal formula*. Intuitively, an agent is maximally powerful if she already knows the goal formula, and is completely powerless if she does not know anything needed in combination with others' knowledge to be able to conclude that the goal formula is true. In between these two extremes are potentially many intermediate levels of power: the more sub-groups the agent can join in order for the group to have shared knowledge of the objective, the more powerful the agent is.

In order to formalise the fact that information about the goal formula is shared in a group, we use the concept of distributed knowledge. We define a simple coalitional game where a coalition is winning iff it has distributed knowledge about the goal formula.

Formally, a *goal structure* is a tuple $S = \langle M, s, \chi \rangle$, where M, s is a pointed model over agents Ag and $\chi \in \mathcal{L}_D$ is a goal formula. Given a goal structure we define the simple game $\langle Ag, v_S^D \rangle$:

$$v_S^D(G) = \begin{cases} 1 & M, s \models D_G \chi \\ 0 & \text{otherwise.} \end{cases}$$

Example 1. Figure 1 shows a model M_{MTW} of the situation described in the introduction. Observe that $M_{MTW}, s \models K_M(p \rightarrow q) \wedge K_T(r \rightarrow q) \wedge K_W(p \wedge r)$, and

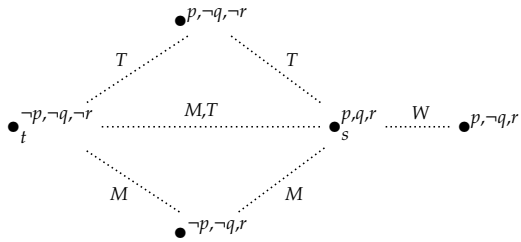


Figure 1: The model M_{MTW}. Reflexive loops are omitted.

also that these formulae represent "private" knowledge of the respective agents; i.e., we have that $M_{MTW}, s \models \neg K_M(r \to q) \wedge \neg K_M(p \wedge r) \wedge \neg K_T(p \to q) \wedge \neg K_T(p \wedge r) \wedge \neg K_W(p \to q) \wedge \neg K_W(r \to q)$. Furthermore observe that $M_{MTW}, s \models \neg D_x q$ for all $x \in \{M, T, W\}$, and that $M_{MTW}, s \models \neg D_{\{M,T\}} q \wedge D_{\{M,W\}} q \wedge D_{\{T,W\}} q$. We thus get that M is swing for exactly $\{W\}$, that T is swing for exactly $\{W\}$, that W is swing for exactly $\{M\}$, $\{T\}$ and $\{M, T\}$, and thus that:

$$\begin{array}{ll} \sigma_M = \sigma_T = 1, \sigma_W = 3 & \mu_M = \mu_T = \frac{1}{4}, \mu_W = \frac{3}{4} \\ \beta_M = \beta_T = \frac{1}{5}, \beta_W = \frac{3}{5} & \varsigma_M = \varsigma_T = \frac{1}{6}, \varsigma_W = \frac{2}{3}. \end{array}$$

What are the properties of v_S^D? From the fact that $D_G\chi$ implies $D_H\chi$ when $G \subseteq H$ it follows that v_S^D is always *monotonic*. In fact, monotonicity completely characterise the (simple) games induced in this way: every monotonic (voting) game is induced by some Kripke structure and goal formula via the definition above.

Theorem 1. *For any simple cooperative game $\Gamma = \langle Ag, v \rangle$, there exists a goal structure S such that $v_S^D = v$ iff Γ is monotonic.*

Proof. The implication to the right is immediate (as already mentioned), so assume that v is monotonic. Let $p \in \Theta$. We construct a goal structure $S = \langle M, s, \chi \rangle$ such that $v_S^D = v$ as follows: $W = \{s_0\} \cup \{s_H : v(H) = 0\}$; $s = s_0$; $V(p) = \{s_0\}$; $\chi = p$. $\sim_i$ is defined by the following equivalence classes: $[s_0]_{\sim_i} = \{s_0\} \cup \{s_H : i \in H\}$ and for every H' such that $i \notin H'$, $[s_{H'}]_{\sim_i} = \{s_{H'}\}$. Informally: for each H such that $v(H) = 0$ there is a designated state s_H where p is false, which no agent in H can discern from s_0.

Let $v(G) = 1$. We must show that $M, s_0 \models D_G p$, so let t be such that $(s_0, t) \in \bigcap_{i \in G} \sim_i$. It suffices to show that $t = s_0$. Assume otherwise: that $t = s_H$ for

some H such that $v(H) = 0$. For every $i \in G$, $s_0 \sim_i s_H$, and by the definition of $\sim_i$ it follows that $i \in H$. Thus, $G \subseteq H$. But since $v(G) = 1$ and $v(H) = 0$, that contradicts monotonicity.

Conversely, let $v(G) = 0$. We have that $s_0 \sim_i s_G$ for every $i \in G$ and $M, s_G \models \neg p$. Thus $M, s_0 \not\models D_G p$. □

4.1 Expressing Power

Epistemic logic can be used to express and reason about power in Kripke structures. The following expressions can, e.g., be used together with a standard model checker, to determine the power distribution in a given structure.

- i is swing for G when the goal is χ:

$$Swing(G, i, \chi) \equiv \neg D_G \chi \wedge D_{G \cup \{i\}} \chi$$

- The Banzhaf score of i wrt. goal χ is at least k:

$$BAL(i, k, \chi) \equiv \bigvee_{G_1 \neq \cdots \neq G_k \subseteq Ag \setminus \{i\}} \bigwedge_{G \in \{G_1, \ldots, G_k\}} Swing(G, i, \chi)$$

- The Banzhaf score of i wrt. goal χ is k:

$$B(i, k, \chi) \equiv BAL(i, k, \chi) \wedge \neg BAL(i, k+1, \chi)$$

- Of potential interest is checking whether or not one agent has more information/power than another. Note that the maximal Banzhaf score is determined by the maximum number of coalitions not containing the agent; 2^{n-1}. The Banzhaf score of agent i is at least as high as that of agent j:

$$BNoLower(i, j, \chi) \equiv \bigvee_{k \in [0, 2^{n-1}]} BAL(i, k, \chi) \wedge \neg BAL(j, k, \chi)$$

- i is a veto player wrt. goal χ:

$$Veto(i, \chi) \equiv \neg D_{Ag \setminus \{i\}} \chi$$

  i is a veto player iff it is included in all winning coalitions, iff all coalitions without i are losing, iff $\neg D_G \chi$ holds for all G without i. By monotonicity this holds iff $Veto(i, \chi)$ holds.

- i is a dictator wrt. goal χ:

$$Dictator(i, \chi) \equiv Veto(i, \chi) \wedge K_i\chi$$

  i is a dictator iff all coalitions containing i are winning, and no coalition without i is winning. This holds iff $Dictator(i, \chi)$ holds, by monotonicity.

- i is a dummy wrt. goal χ:

$$Dummy(i, \chi) \equiv \bigwedge_{G \in 2^{Ag}} D_{G\cup\{i\}}\chi \rightarrow D_G\chi$$

  i is a dummy iff $\forall G : M, s \models \neg(\neg D_G\chi \wedge D_{G\cup\{i\}}\chi)$ which is equivalent to $\forall G : M, s \models D_{G\cup\{i\}}\chi \rightarrow D_G\chi$.

4.2 Full Communication

Implicit in the idea of information-based power is that groups of agents should somehow be able to *realise* the knowledge distributed among them in order to jointly find out that the goal formula is true. However, while distributed knowledge is the most popular concept in the literature aiming to capture the "sum" of the knowledge in a group, it has the following property, as first pointed out in van der Hoek et al. (1999). It might be that G has distributed knowledge of the goal, but it is still not possible for the group to establish χ through communication in the following sense: it might not be the case that there exists a formula ϕ_i for each $i \in G$ such that $M, s \models \bigwedge_{i\in G} K_i\phi_i$ and $\models \bigwedge \phi_i \rightarrow \chi$. This (possibly lacking) communication property is equivalent van der Hoek et al. (1999) to:

$$M, s \models D_G\chi \quad \Rightarrow \quad \bigcup_{i\in G} \mathcal{K}_i(M, s) \models \chi \tag{5}$$

and van der Hoek et al. (1999) calls this the *principle of full communication* (the other direction of (5), $\bigcup_{i\in G} \mathcal{K}_i(M, s) \models \chi \Rightarrow M, s \models D_G\chi$, holds on any model). As an example, consider the model M_1 in Figure 2. In this model p is distributed knowledge among agents 1 and 2 in state s, but p is not entailed from the individual knowledge of 1 and 2 in s and the model does not satisfy the principle of full communication.

So, if we take the p as the goal formula, agent 1 is swing for $\{2\}$ in state s in the model M_1 above, but it is not possible for agents 1 and 2 to actually infer

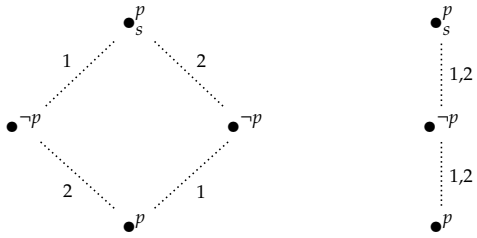


Figure 2: Models M_1 (left) and M_2 (right). Reflexive and transitive edges omitted.

p together by communicating using the epistemic language. Our information-based power measures make particular sense in models that satisfy the principle of full communication, because in such models whatever is distributed knowledge can be obtained by communication in the sense that it follows from individual knowledge that the involved agents can specify and communicate as logical formulas. So which models satisfy the principle of full communication? There are two particularly relevant model properties here (generalisations of propositions given in van der Hoek et al. (1999)). A model $M = \langle W, \sim_1, \ldots, \sim_n, \pi \rangle$ is a:

- *full model* Gerbrandy (1999) iff for all $s \in W$, $G \subseteq Ag$, and $\Phi \subseteq \mathcal{L}_D$: if $\Phi \cup \mathcal{K}_G(M, s)$ is satisfiable then Φ is satisfiable in $\{t : (s, t) \in \sim_G^D\}$.
- *full communication model* Roelofsen (2007) iff for all $s \in W$, $G \subseteq Ag$, and $\phi \in \mathcal{L}_K$: if $\{\phi\} \cup \mathcal{K}_G(M, s)$ is satisfiable then ϕ is satisfiable in $\{t : (s, t) \in \sim_G^D\}$.

Clearly, full models are full communication models. Gerbrandy (1999) shows that fullness is sufficient for the principle of full communication to hold, while Roelofsen (2007) shows that a model satisfies the principle of full communication *if and only if* the model is a full communication model.

While this definition of full communication models may seem somewhat technical, note that the principle of full communication is often violated by the existence of bisimilar states in the model (such as in the model above). Indeed, bisimulation contractions of finite models are full communication models (they are *distinguishing* in the sense of van der Hoek et al. (1999), due to the existence of characteristic formulae). Models that are finite and do not contain bisimilar states (and thus are their own bisimulation contractions) are very common.

Thus, on full communication models we get an alternative, equivalent, definition of power. We have that:

$$v_S^D(G) = 1 \quad \Leftrightarrow \quad \bigcup_{i \in G} \mathcal{K}_i(M, s) \models \chi \tag{6}$$

when M is a full communication model.

4.3 Properties of Power

The relationship between power properties and epistemic properties is of natural interest, not the least in order to validate that our definition of power is reasonable. The relationship properties in the following lemma are discussed below.

Lemma 1. *Let the goal structure $S = \langle M, s, \chi \rangle$ be given.*

1. *If $M, s \models \neg D_{Ag}\chi$, then $x_i = 0$ for all i and $x \in \{\sigma, \mu, \beta, \varsigma\}$.*
2. *If $M, s \models \neg\chi$, then $x_i = 0$ for all i and $x \in \{\sigma, \mu, \beta, \varsigma\}$.*
3. *If $M, s \models K_i\chi$, then $x_i \geq x_j$ for all j and $x \in \{\sigma, \mu, \beta, \varsigma\}$.*
4. *If $M, s \models \neg D_{Ag\setminus\{i\}}\chi$, $x_i \geq x_j$ for all j and $x \in \{\sigma, \mu, \beta, \varsigma\}$.*
5. *If $M, s \models K_i\chi \wedge \neg K_j\chi$, then $x_i > x_j$ for all $x \in \{\sigma, \mu, \beta, \varsigma\}$.*
6. *On full communication models, if $\mathcal{K}_i(M, s) \subseteq \mathcal{K}_j(M, s)$ then $x_i \leq x_j$, for any power measure $x \in \{\sigma, \mu, \beta, \varsigma\}$.*

The first property says that if not enough information to infer the goal formula is distributed throughout the complete system, then every agent has *no power*. The second property is a special case of the first – the goal *cannot* be derived because it is not true. The third and fourth properties represent the other extreme: *maximum power*. The agent has maximum power (at least as much power as anyone else) if she already knows the goal, or if the rest of the system does not have enough information to derive the goal (i.e, if the agent is a veto player). The fifth and sixth properties are about *relative power*. The fifth says that an agent who already knows χ is always strictly more powerful than an agent who does not know χ. The sixth property says that if one agent knows at

least as much as another agent, then the first agent is at least as powerful. This relates our definition of power to a more classical notion of "knowing more" in a reasonable way. Our notion is more fine grained; the implication does not hold in the other direction. The sixth property holds for full communication models, which, again, is a natural class of models in which to interpret our power measures since they come with a natural mechanism for distribution of information.

Proof (of Lemma 1).
1. Follows immediately from monotonicity: if i is swing for G, then $M, s \models D_{G\cup\{i\}}\chi$.

2. Immediate from $\models \neg\chi \rightarrow D_{Ag}\neg\chi$ and the first item.

3. If suffices to show that i is swing for any coalition any agent j is swing for. So assume that $M, s \models \neg D_G\chi \wedge D_{G\cup\{j\}}\chi$. From $M, s \models K_i\chi$ it follows that $M, s \models D_{G\cup\{i\}}\chi$, and thus i is also swing for G.

4. Assume that j is swing for G. From $M, s \models D_{G\cup\{j\}}\chi$, the assumption that $M, s \models \neg D_{Ag\setminus\{i\}}\chi$ and monotonicity, it follows that $i \in G$. Thus it also follows that i is swing for $(G \setminus \{i\}) \cup \{j\}$. Because $i \in G$ and $j \notin G$ for coalitions G for which j is swing, $(G_1 \setminus \{i\}) \cup \{j\} \neq (G_2 \setminus \{i\}) \cup \{j\}$ for any two different coalitions G_1, G_2 for which j is swing, and thus there are at least as many swings for i.

5. If j is swing for G, $M, s \models \neg D_G\phi$ so G cannot contain i and i is also swing for G. In addition, i is swing for $\emptyset$, unlike j.

6. Let M be a full communication model and assume that i is swing for G, i.e., that $M, s \models D_G\chi \wedge D_{G\cup\{i\}}\chi$. From the fact that M is a full communication model and eq. (6) above, we get that $\bigcup_{l\in G\cup\{i\}} \mathcal{K}_l(M, s) \models \chi$. From $\mathcal{K}_i(M, s) \subset \mathcal{K}_j(M, s)$ it follows that $\bigcup_{l\in G\cup\{j\}} \mathcal{K}_k(M, s) \models \chi$ which again means that $M, s \models D_{G\cup\{j\}}\chi$. Thus, j is swing for G. □

In the following lemma we look at power measures in "similar" models. The proper notion of bisimulation for distributed knowledge, and hence our power measures, is given in the second point.

Lemma 2.

1. *The power measures are not invariant under (standard) bisimulation. That is, bisimilar pointed models may have different power measures.*
2. *The power measures are invariant under* collective bisimulation *Roelofsen (2007).*

3. *On full models, the power measures are invariant under (standard) bisimulation.*

Proof. 1. A counter-example is found in Figure 2, which contains two bisimilar models with two agents. It is easy to see that by taking $\chi = p$, we get $\sigma_1 = 1$ in M_1 but $\sigma_1 = 0$ in M_2.

2. follows immediately from the fact that satisfaction in $\mathcal{L}_D$ is invariant under collective bisimulation (Roelofsen 2007, Prop. 19).

3. For full models the notions of collective bisimulation and bisimulation coincide (Roelofsen 2007, Prop. 20). □

Finally, let us look at the relationship between power properties and the structure of the goal formula. We will make use of the logical expressions of power properties from Section 4.1.

Starting with tautologies and contradictions:

$$\begin{array}{ll} \models \neg Swing(G, i, \top) & \models \neg Swing(G, i, \bot) \\ \models Veto(i, \bot) & \models \neg Veto(i, \top) \\ \models \neg Dictator(i, \bot) & \models \neg Dictator(i, \top) \end{array}$$

With such goal formulae, no agent can be swing for any coalition. Every agent is a veto player for $\bot$, while no agent is a veto player for $\top$. No agent can be a dictator for $\bot$ nor $\top$.

The case of conjunction:

$$\models (Swing(G, i, \chi_1) \wedge Swing(G, i, \chi_2)) \rightarrow Swing(G, i, \chi_1 \wedge \chi_2)$$

Swings are closed under the operation of taking conjunction of goal formulae. The converse does not hold, but this does:

$$\models Swing(G, i, \chi_1 \wedge \chi_2) \rightarrow (Swing(G, i, \chi_1) \vee Swing(G, i, \chi_2))$$

– if i is swing wrt. a conjunction, she is swing wrt. at least one of the conjuncts (but not necessarily both).

For negation we have that (but not the other way around):

$$\models Swing(G, i, \neg\chi) \rightarrow \neg Swing(G, i, \chi)$$

Moving on to the case that the goal formula is epistemic, first observe the following properties of distributed $S5$ knowledge: $\models D_G D_{G'} \phi \rightarrow D_G \phi$ for any

G, G', and $\models D_G D_{G'}\phi \leftrightarrow D_G\phi$ when $G \subseteq G'$. From these properties it follows that:

$$\models Swing(H, i, D_G\chi) \rightarrow Swing(H, i, \chi) \quad \text{when } H \subseteq G$$
$$\models Swing(H, i, D_G\chi) \leftrightarrow Swing(H, i, \chi) \quad \text{when } H \cup \{i\} \subseteq G$$

In particular, using a goal formula $D_G\phi$ is equivalent to using ϕ when it comes to counting swings within G.

If we take $G = \{j\}$ in the expressions above, we get the case where the goal formula describes individual knowledge. It follows that:

$$\models Swing(\emptyset, i, K_j\chi) \rightarrow Swing(\emptyset, i, \chi) \quad \text{for any } j$$
$$\models Swing(\{j\}, i, K_j\chi) \rightarrow Swing(\{j\}, i, \chi) \quad \text{for any } j$$
$$\models Swing(\emptyset, i, K_i\chi) \leftrightarrow Swing(\emptyset, i, \chi)$$

5 Knowledge of Power

We have thus associated power indices with states of Kripke structures, by assuming that they are defined by agents' knowledge. But epistemic logic allows us to reason about agents' knowledge *about* state-properties – so we can go from analysing the power of knowledge to analysing knowledge of power: what do the agents in the system know about the distribution of power?

The formula $K_j Swing(G, i, \chi)$, where $Swing(G, i, \chi) = \neg D_G\chi \wedge D_{G\cup\{i\}}\chi$, denotes the fact that agent j knows that i is swing for G. If we look first at the more general case of *distributed* knowledge of that fact, we have the following (we formally prove this and the following validities in Theorem 2 below):

$$\models Swing(G, i, \chi) \rightarrow D_{G\cup\{i\}} Swing(G, i, \chi) \tag{7}$$

– if i is swing for G, then this is distributed knowledge in $G \cup \{i\}$.

However, this does not carry over to individual knowledge. It turns out that $Swing(G, i, \chi) \wedge \neg K_j Swing(G, i, \chi)$ is satisfiable, for any j including $j = i$. Thus, an agent can be swing for a coalition, without neither the agent nor the agents in the coalition knowing it. When, then, *does* an agent know that she is swing? The answer is: *almost never*. The following holds:

$$\models K_j \neg Dummy(i, \chi) \rightarrow K_j\chi \tag{8}$$

for any i, j (including $i = j$). In other words, an agent can only know that any agent (including herself) is swing for any coalition if she (the first agent) already

knows the goal formula! In the typical case that χ is distributed information throughout the system, but no individual agent alone knows χ, *no* agent knows that *any* agent can swing *any* coalition from ignorance to knowledge about χ. It follows that

$$\models K_j \neg Dummy(i, \chi) \rightarrow K_j \bigwedge_{k \in Ag} BNoLower(j, k, \chi) \tag{9}$$

– only agents that are maximally powerful (at least as powerful as any other agent), and know that they are, can know that anyone (including themselves) are not a dummy player.

It also holds that

$$\models K_j Swing(G, i, \chi) \rightarrow K_j Swing(G, j, \chi) \tag{10}$$

– if an agent knows that another agent is swing for some coalition, then the first agent must be swing for the same coalition. In particular: $\models K_j \neg Dummy(i, \chi) \rightarrow K_j \neg Dummy(j, \chi)$.

However, *no* agent *in* a coalition can know that someone is swing for that coalition:

$$\models \bigwedge_{j \in G} \neg K_j Swing(G, i, \chi) \tag{11}$$

For veto players, we have that

$$\models K_i Veto(j, \chi) \rightarrow \neg K_i \neg Dummy(i, \chi) \quad i \neq j \tag{12}$$

– the only agents that can know that someone else is a veto player are agents that consider it possible that they are dummies themselves.

For dictators, we have that

$$\models \neg K_j Dictator(i, \chi) \quad i \neq j \tag{13}$$

– the only agent that can know who the dictator is, is the dictator.

Turning to knowledge about the values of power indices, we have

$$\models K_j B(i, k, \chi) \rightarrow BNoLower(j, i, \chi) \tag{14}$$

– no agent can know the Banzhaf score of any agent with a lower score than herself.

We can conclude that the distribution of power is generally not known *in* the system. We emphasise that this does not pose any problem for our interpretation of the power indices as measures of the distribution of information in the system, as we discuss further in Section 7.

Theorem 2. *Properties (7)–(14) hold.*

Proof. We make use of the fact that distributed knowledge satisfies the $S5$ properties Blackburn et al. (2001), which follows from the fact that the intersection of equivalence relations is an equivalence relation, as well as the monotonicity property ($D_G\phi \rightarrow D_H\phi$ when $G \subseteq H$).

(7): from $\neg D_G\chi$ it follows that $D_G\neg D_G\chi$ by negative introspection, and $D_{G\cup\{i\}}\neg D_G\chi$ follows by monotonicity. $D_{G\cup\{i\}}D_{G\cup\{i\}}\chi$ follows from $D_{G\cup\{i\}}\chi$ by positive introspection. $D_{G\cup\{i\}}Swing(G,i)$ follows by knowledge distribution.

(8): $K_j\neg Dummy(i,\chi)$ is equal to $K_j\bigvee_G(D_{G\cup\{i\}}\chi \wedge \neg D_G\chi)$. By reflexivity $D_{G\cup\{i\}}\chi$ implies χ, and thus $\bigvee_G(D_{G\cup\{i\}}\chi \wedge \neg D_G\chi)$ implies that χ. By knowledge distribution, $K_j\chi$ holds.

(9): let $K_j\neg Dummy(i,\chi)$ be true. By (8), $K_j\chi$ and from positive introspection $K_jK_j\chi$. From Lemma 1.3 it follows that $K_jBNoLower(j,k,\chi)$ for any k.

(10): from $K_jSwing(G,i,\chi)$ it follows that $K_j\neg D_G\chi$. By (8) it also follows that $K_j\chi$. By knowledge distribution, $K_j(\neg D_G\chi \wedge K_j\chi)$, which by monotonicity implies that $K_j(\neg D_G\chi \wedge D_{G\cup\{j\}}\chi)$.

(11): if $K_jSwing(G,i,\chi)$ is true for some $j \in G$, then $K_jSwing(G,j,\chi)$ by (10), and $Swing(G,j,\chi)$ by reflexivity. But this is a contradiction.

(12): from $K_iVeto(j,\chi)$ it follows that $K_i\neg K_i\chi$ when $i \neq j$, from which it follows that $\neg K_i\chi$. If $K_i\neg Dummy(i,\chi)$ is true, then $K_i\chi$ by (8); a contradiction.

(13): $K_jDictator(i,\chi)$ is equivalent to $K_j(Veto(i,\chi) \wedge K_i\chi)$, which implies that $K_j\chi$ and $Veto(i,\chi)$. From the latter it follows that $\neg D_{Ag\setminus\{i\}}\chi$, and from monotonicity it follows that $\neg K_j\chi$ – a contradiction.

(14): if $\sigma_i = 0$, the formula holds trivially. If $\sigma_i > 0$, $K_jB(i,k,\chi)$ implies that there is a G such that $K_j(\neg D_G\chi \wedge D_{G\cup\{i\}}\chi)$ is true. It follows that $K_j\chi$, and by Lemma 1.3 that $\sigma_j \geq \sigma_i$. □

6 Other types of group knowledge

We have so far used the notion of distributed knowledge to measure power. Can other notions of group knowledge be used? Note that both everybody-knows and common knowledge are anti-monotonic, in the sense that $C_G\phi$ implies $C_{G'}\phi$ when $G' \subseteq G$, while distributed knowledge is monotonic ($D_{G'}\phi$ implies $D_G\phi$). This means that simply "replacing" distributed knowledge in the definition of the game by any of these notions would not make sense (e.g., $\neg C_G\phi \wedge C_{G\cup\{i\}}\phi$ is not satisfiable). However, there is another way in which we can look at an agent's power with respect to common knowledge (and similarly with everybody-knows). An agent has "negative" power if he can swing a coalition from *having* common knowledge of the goal, to *not* having it. In other words, this would correspond to an agent's power to spoil, rather than to achieve, the goal. Using this definition of the power measures, a high value means that the agent has *little* information, and including it in a group is likely to, e.g., break common knowledge needed for coordination.

Let us start with everybody-knows. Given $S = \langle M, s, \chi\rangle$, let:

$$v_S^E(G) = \begin{cases} 1 & M, s \models \neg E_G\chi \\ 0 & \text{otherwise} \end{cases}$$

We say that a simple cooperative game is *determined* if there is a set of agents $Winners \subseteq Ag$ such that $v(G) = 1$ iff $G \cap Winners \neq \emptyset$. Note that determined games are monotonic.

Theorem 3. *For any simple cooperative game $\Gamma = \langle Ag, v\rangle$, there exists a goal structure S such that $v_S^E = v$ iff Γ is determined.*

Proof. For the implication to the right, given S let $Winners = \{i : M, s \models \neg K_i\chi\}$. It is easy to see that $v_S^E(G) = 1$ iff $G \cap Winners \neq \emptyset$. For the implication to the right, we define $S = \langle M, s, \chi\rangle$ as follows. Let $p \in \Theta$. Let $W = \{s, t\}$; $s_0 = s$; $V(p) = \{s\}$, $V(q) = \emptyset$ for $q \neq p$; $s \sim_i t \Leftrightarrow i \in Winners$; $\chi = p$. Let $v(G) = 1$. That means that there is an agent i such that $i \in G \cap Winners$. From $i \in Winners$ it follows that $M, s_0 \models \neg K_i p$, and since $i \in G$ we get that $M, s_0 \models \neg E_G\chi$. For the other direction, let $M, s_0 \models \neg E_G p$. That means that $M, s_0 \models \neg K_i p$ for some $i \in G$. But the only possibility then is that also $i \in Winners$. Thus, $i \in G \cap Winners$, and thus $v(G) = 1$. □

It is easy to see that for determined games, the Banzhaf score is the same for all winners, as well as the same (0) for all non-winners:

Lemma 3. *For any determined game and any agent i,*

$$\sigma_i = \begin{cases} 2^{|Ag \setminus Winners|} & i \in Winners \\ 0 & otherwise \end{cases}$$

It follows that it is easy to compute the power measures:

Theorem 4. *Given a goal structure $S = \langle M, s, \chi \rangle$ and an agent i in M, the Banzhaf score σ_i for i in the game $\langle Ag, v_S^E \rangle$ can be computed in polynomial time.*

Proof. By Theorem 3 the game is determined. The winners can be computed in polynomial time: for every state t, check whether $M, t \models \neg\chi$, and if it does add i to *Winners* if there is an i-transition from t to s. σ_i is computed from the size of *Winners* according to Lemma 3. □

Moving on to common knowledge, given $S = \langle M, s, \chi \rangle$, let:

$$v_S^C(G) = \begin{cases} 1 & M, s \models \neg C_G \chi \\ 0 & \text{otherwise} \end{cases}$$

Example 2. The following two examples are inspired by (van Ditmarsch et al. 2007, Section 2.3). In the first setting, the set of agents Ag is the set of participants of a conference, and $a \in Ag$ represents our hero Alco. During one afternoon, while all other participants are attending a joint session, Alco spends his time in the bar of the conference hotel. The session chair announces χ: 'tomorrow, sessions start at 9:00 rather than 10:00'. Everybody (i.e., Ag) at the conference feels very responsible for the well-being of the participants, and only if $C_{Ag}\chi$ holds, people will stop informing each other of χ. If s is the situation immediately after the chair's announcement, we obviously have $M, s \models Swing(Ag \setminus \{a\}, a, \chi)$, where *Swing* is now defined for common knowledge: $Swing(G, i, \chi) = C_G\chi \wedge \neg C_{G \cup \{i\}}\chi$. Now consider a new state s_1, in which Alco leaves the bar to get some fresh air, and which leads to a state s_2 where at the general session a friend f of Alco makes the chair (publicly) aware that Alco was in the bar during the announcement χ. At this moment it is common knowledge among $Ag \setminus \{a\}$ that $Swing(Ag \setminus \{a\}, a, \chi)$, but then the chair replies to f by saying that there is an intercom in the bar that is directly connected

to the conference room. Note that a is now still a veto player wrt. Ag and χ, since Alco does not know about the discussion regarding his absence during the announcement of χ. In other words, although in s_2 we have $E_{Ag}\chi$, we also have $\neg K_a K_f K_a \chi$: Alco knows that his friend f may have concerns about Alco not knowing χ (this concern is justified, since f notified the chair), and Alco does not know that f has been properly informed (that $K_a\chi$) by the chair, so one may expect that a will make at some time an effort to make publicly known that he knows χ, so people can stop worrying about a's time-table tomorrow.

Swing players for common knowledge in a coalition G often come with delicate protocols for the communication in G. An example here is the celebrations of Santa Claus in certain cultures, where it is common knowledge among those over a certain age that Santa Claus is in fact not responsible for the presents at the evening (this is χ), while χ is not known among the participants under a certain age. Now, even when everybody at the Christmas party knows that χ, there may be several swing players for several coalitions, which explains that conversations have to be participated in carefully. To be more precise, suppose that $E_G E_G \chi \wedge \neg K_i K_j K_i \chi$ (with $i, j \in G$). Since i knows that everybody in G knows χ already, he might chose not to look childish to j and reveal to j that $K_i\chi$, indicating he is not a fool. But i might also chose to exploit $\neg K_i K_j K_i \chi$, and challenge j into a 'dangerous conversation', where j may think he needs to be careful not to reveal χ to i.

These examples also suggest that power is in fact an interesting issue in dynamic contexts, after enough communication has taken place for instance, Alco may seize to be a swing player. Dynamic Epistemic Logic (van Ditmarsch et al. (2007)) paves the right formal framework to study these phenomena, like the fact that some true formulas can never be known no matter how often they are announced: they would always have a veto player (Moore sentences like $(p \wedge \neg K_a p)$ being the most prominent examples).

Like for the case of distributed knowledge, the class of games obtained in this way is exactly the monotonic games.

Theorem 5. *For any simple cooperative game $\Gamma = \langle Ag, v \rangle$, there exists a goal structure S such that $v_S^C = v$ iff Γ is monotonic.*

Proof. It is easy to see that v_S^C is monotonic.

For the other direction, let v be monotonic. If there is no coalition G with

$\nu(G) = 1$, let M consist of only one state s with $V(p) = \{s\}$ and $\sim_a = W \times W$ for every $a \in Ag$. It is easily seen that $v^C_{M,s,p}(G) = 0$ for all coalitions G.

Otherwise put first of all $s \in W \cap V(p)$ and add (s, s) to each $\sim_a$. Let $H_1, \ldots H_k$ be the coalitions with the property that $\nu(H_i) = 1$ and for no proper subset of H_i, it holds that $\nu(H) = 1$. For each such H_i, do the following. Let $H_i = \{a^i_1, a^i_2, \ldots a^i_{m(i)}\}$. Add new states $W_i = \{s^i_1, s^i_2, \ldots s^i_{m(i)}\}$ to W in such a way that (s, s^i_1) and (s^i_1, s) become members of $\sim_{a^i_1}$ and furthermore add (s^i_j, s^i_{j+1}), (s^i_{j+1}, s^i_j) to $\sim_{a^i_{j+1}}$ with $1 \leq j < m(i)$. Add $(s^i_{j'}, s^i_j)$ to each $\sim_a$ $(1 \leq m(i))$. Finally, add $W_i \setminus \{s^i_{m(i)}\}$ to $V(p)$. When this process has finished for all H_i, take the transitive symmetric reflexive closure of every $\sim_a$ so far defined. The effect of this last step is that for every agent a and every two states s^i_1 and s^j_1 with (s, s^i_1) and $(s, s^j_1) \in \sim_a$, we also add (s^i_1, s^j_1) and (s^j_1, s^i_1) to $\sim_a$.

A straight path π in the model is a sequence of state-agent alterations $\langle x_1, a_1, x_2, a_2 \ldots x_n \rangle$, with each $x_i \in W, a_i \in Ag$, and $(x_i, x_{i+1}) \in \sim_{a_i}$ such that $x_i \neq x_j$ if $i \neq j$. It is a straight s-path if $x_1 = s$. Let $Ag(\pi)$ be the set of agents occurring in π. Note that a straight s-path that ends in state s_n denotes a 'shortest' route in the model from s to s_n, since the states in a straight path are different. A straight path $x_1, a_1, x_2, a_2 \ldots x_n$ leads to φ if x_n is the only-φ world in it. The following is an important property of our model: there is a straight path π leading to $\neg p$ iff for some H_i, we have $\nu(H_i) = 1$ and $Ag(\pi) = H_i$.

We now prove that $\forall G \subseteq Ag$ ($\nu(G) = 1$ iff $M, s \models \neg C_G p$). First, if $\nu(G) = 1$, there is a smallest set $H_i = \{a^i_1, \ldots, a^i_{m(i)}\} \subseteq G$ such that $\nu(H_i) = 1$. For this H_i, we have constructed a straight s-path π leading to $\neg p$ and for which $Ag(\pi) = H_i$. So, we have $M, s \models \neg C_{H_i} p$, and hence $M, s \models \neg C_G p$, i.e., $v^C_S(G) = 1$. Secondly, suppose $M, s \models \neg C_G p$, it means for our model that there is a straight s-path π leading to $\neg p$ for which $Ag(\pi) \subseteq G$ (indeed, there may be agents $a \in G \setminus Ag(\pi)$). But the only such paths we have in M are paths that use a minimal set of agents H_i for which $\nu(H_i) = 1$, so $\nu(Ag(\pi)) = 1$. By monotonicity, $\nu(G) = 1$. □

7 Discussion

We have shown that our information-based notion of power has reasonable properties, at least on full communication models – which come with a natural mechanism for distribution of information. We have also shown that it is easy

to compute such power indices using a standard model checker for epistemic logic.

It is natural to define swings using distributed knowledge. A high power index here means that the agent's knowledge is important for an arbitrary group jointly getting to know the goal formula by sharing their information. We also gave alternative definitions of "negative" power in terms of swinging a group from a situation where every member knows the goal, or the goal is common knowledge. Here, a high power index means that the agent knows little: if it is important to have common knowledge in a group (e.g., for coordination), then it is likely that including a high-power agent will lead to failure. The everybody-knows case is computationally tractable, but the price is a lower "resolution": the agents divide into only two classes, with agents in the same class having the same power. It is interesting that the common knowledge case and the distributed knowledge correspond to the same class of voting games (Theorems 1 and 5). If this seems counter-intuitive, keep in mind that the two theorems express that there is a connection between distributed knowledge and *lack* of common knowledge: conceiving distributed knowledge as a game where a coalition wins if it implicitly knows the goal formula, is structurally similar to conceiving common knowledge as a game where a coalition wins if it does *not* commonly know the goal.

Van Ditmarsch et al. (2009) studies a particular notion of "knowing more". Their concept "i knows at least what j knows" is defined by $R_i(s) \subseteq R_j(s)$ where s is a state and $R_x(s) = \{t : (s,t) \in R_x\}$ and R_x is an indistinguishability relation for agent x. Our power measures for distributed knowledge agree: if $R_i(s) \subseteq R_j(s)$ then $Swing(G,j,\chi)$ implies that $Swing(G,i,\chi)$ for any χ, and thus $\sigma_i \geq \sigma_j$. The implication does not hold in the other direction; our notion of "knowing more" is more fine grained. Van Ditmarsch et al. (2009) also introduces a modal operator $\geq$ where, for agents i and j, the formula $i \geq j$ expresses that whatever state is an alternative for j, is also an alternative for i. This provides a way to locally express that $K_i\varphi \rightarrow K_j\varphi$ for all φ. There is one sense in which such an operator allows one also to express properties of the power of knowledge in a compact way. For distributed knowledge for instance, the formula $i \geq j$ implies that $(Swing(G,i,\chi) \rightarrow Swing(G,j,\chi))$ and $\neg Swing(G \cup \{i\},j,\chi)$ – for *any* χ. When reasoning about the power in the context of everybody knows, "opposite" properties derive: $\models (i \geq j) \rightarrow (Swing(G,j,\chi) \rightarrow Swing(G,i,\chi))$ and $\models (i \geq j) \rightarrow \neg Swing(G \cup \{j\},i,\chi)$. Note that such properties cannot be expressed in modal logic without such an operator: for instance in $\models (K_i\varphi \rightarrow K_j\varphi) \rightarrow (Swing(G,i,\chi) \rightarrow Swing(G,j,\chi))$ the formula φ is a specific formula

(not a scheme), and $\models (K_i\varphi \rightarrow K_j\varphi) \Rightarrow \models (Swing(G,i,\chi) \rightarrow Swing(G,j,\chi))$ is obviously true, but much weaker: the antecedent is false (if $i \neq j$).

In Section 5 we saw that agents *in* the system generally know very little about the distribution of information-based power in the system. For example, an agent with a high power index typically does not know which coalitions she needs to join in order to derive the goal formula (or indeed *that* she is a high-power agent). We emphasise that this is not in any way a problem for the interpretation of our power indices. A high Banzhaf index means, in our setting, that the probability of changing some arbitrary coalition from ignorance to knowledge about the goal is high – in the same way that it is interpreted as the probability of changing an outcome in voting theory. In fact, that an agent does not know which coalitions it is swing for makes the probability of being swing for an *arbitrary* coalition more interesting. Furthermore, in many distributed and multi-agent systems, such as sensor networks, agents are restricted to communication with some arbitrary sub-group of all agents at any given time. We think of these power measures as a tool for external analysis of the information distribution in a system, to find out, e.g., whether information is evenly distributed or whether there are some agents that are particularly crucial to the functioning of the system in the sense that the information they have is difficult to obtain elsewhere in the system. The negative results about knowledge of power properties can also be seen as a *barrier against strategic behaviour*: it is almost never possible for an agent to know that it suffices to share information with only some particular subgroup of the grand coalition.

An interesting direction for future work is to associate formulae of the form $D_G D_H \phi$ with *composite* voting games (Felsenthal and Machover 1998, p. 27). In this paper we have studied a semantic notion of power, associated with a point in a Kripke structure. Another direction for future work is to develop a *syntactic* notion of power, based on a set of epistemic formulae. For such an approach it would be necessary to syntactically describe that agents know "this and nothing more", and extensions of epistemic logic with *only knowing* Levesque (1990) seem like a promising starting point.

Acknowledgements We thank the AAMAS program committee, Pål Grønås Drange, Alexandru Baltag, Johan van Benthem and Fenrong Liu for comments that helped us improve the paper.

References

T. Ågotnes, W. van der Hoek, J. A. Rodriguez-Aguilar, C. Sierra, and M. Wooldridge. On the logic of normative systems. In M. M. Veloso, editor, *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI 2007)*, pages 1175–1180, California, 2007. AAAI Press.

T. Ågotnes, W. van der Hoek, M. Tennenholtz, and M. Wooldridge. Power in normative systems. In Decker, Sichman, Sierra, and Castelfranchi, editors, *Proceedings of the Eighth International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, pages 145–152, Budapest, Hungary, May 2009. IFAMAAS.

J. F. Banzhaf III. Weighted voting doesn't work: A mathematical analysis. *Rutgers Law Review*, 19(2):317–343, 1965.

P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Cambridge University Press: Cambridge, England, 2001.

R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning About Knowledge*. The MIT Press: Cambridge, MA, 1995.

D. S. Felsenthal and M. Machover. *The Measurement of Voting Power*. Edward Elgar: Cheltenham, UK, 1998.

J. Gerbrandy. *Bisimulations on Planet Kripke*. Ph.D. thesis, University of Amsterdam, 1999.

A. Hunter and S. Konieczny. On the measure of conflicts: Shapley inconsistency values. *Artificial Intelligence*, 174(14):1007 – 1026, 2010. ISSN 0004-3702. doi: DOI:10.1016/j.artint.2010.06.001.

H. J. Levesque. All I know: a study in autoepistemic logic. *Artificial Intelligence*, 42:263–309, 1990.

M. J. Osborne and A. Rubinstein. *A Course in Game Theory*. The MIT Press: Cambridge, MA, 1994.

F. Roelofsen. Distributed knowledge. *Journal of Applied Non-Classical Logics*, 17(2):255–273, 2007.

Y. Shoham and M. Tennenholtz. On the synthesis of useful social laws for artificial agent societies. In *Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI-92)*, San Diego, CA, 1992.

W. van der Hoek, B. van Linder, and J.-J. Meyer. Group knowledge is not always distributed (neither is it always implicit). *Mathematical Social Sciences*, 38:215–240, 1999.

H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*. Springer, Berlin, 2007.

H. Van Ditmarsch, W. Van Der Hoek, and B. Kooi. Knowing more: from global to local correspondence. In *IJCAI'09: Proceedings of the 21st international jont conference on Artifical intelligence*, pages 955–960, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

Two Logical Faces of Belief Revision

Johan van Benthem

University of Amsterdam and Stanford University
j.vanbenthem@uva.nl

Abstract

This piece proposes a style of thinking using modal frame correspondence that puts Segerberg's dynamic doxastic logic and 'Dutch' dynamic-epistemic logic for belief change in one setting. While our technical results are elementary, they do suggest new lines of thought.[1]

1 Two modal logics for belief change

Belief revision theory is a small corner of the world of philosophy and computer science, and modal logic is a small corner of the world of logic. When two specialized topics come together, surely, there can be only one way of doing that? The dynamic-doxastic logic *DDL* of Segerberg (1995; 1999)[2] has abstract modal operators describing transitions in abstract universes of models to describe changes in belief, and then encodes basic postulates on belief change in modal axioms that can be studied by familiar techniques. But there is also another line in the logical literature, started in van Benthem (2007), Baltag and Smets (2008)[3] that works differently. Here belief changes are modeled in the

[1] This paper will appear in a volume celebrating the work of Krister Segerberg.

[2] See also Leitgeb and Segerberg (2007) for extensive discussion of the research program.

[3] Relevant predecessors to this work are Aucher (2004), van Ditmarsch (2005).

framework of dynamic-epistemic logic (*DEL*) as acts of changing a plausibility ordering in a current model, and the update rule for doing that is made explicit, while its properties are axiomatized completely in modal terms. The contrast may be stated as follows. Segerberg follows *AGM* belief revision theory (Gärdenfors 1988) in its *postulational* approach constraining spaces of all possible belief changes, while the *DEL* approach is *constructive*, studying specific update rules and the complete logics of their corresponding dynamic model-changing modalities.

Stated this way, there need not be any conflict between the two approaches – and in fact, there is not. Still, there are many differences in their subsequent technical agenda.[4] One could spend much time analyzing these differences, but my aim in this paper is modest. I want to suggest that, for colleagues from modal logic, *DDL* and *DEL* fit very well, if we use the method of *frame correspondence*. This suggestion occurs in van Benthem (2011), but I will pursue it more systematically here. My results are simple technically, but they suggest new perspectives. I start with knowledge in Section 2, exploring frame correspondences for 'public announcement logic' *PAL*. Many general methodological points can be made at this level, as they are not specific to belief. Next, I give modal correspondence for logics of belief change in Section 3. In Section 4, I discuss two generalizations: full dynamic-epistemic logic with product update over event models, and an extension of correspondence analysis to neighborhood models, using the DEL treatment in van Benthem and Pacuit (2011). Section 5 lists new general issues coming to light in my analysis, all of them 'to be explored'. Section 6 states the conclusion of this paper, though it will already be clear right here at the start: the two existing styles of modal logic for belief revision live well together, and analyzing their connections actually reveals some interesting issues that will unfold in due course.

[4] *DEL*-style logics of belief revision depart from the *AGM*-format in a number of ways. (i) The content of new beliefs need not be factual, but it can itself consist of complex statements about beliefs. (ii) What changes in acts of revision is not just beliefs, but crucially also conditional beliefs. (iii) Infinitely many types of triggering event can be analyzed structurally in the logic by mechanisms like 'event models' or 'model-change programs'. (iv) The setting is essentially multi-agent, making, in principle, social acts of belief merge as crucial to the logical system as individual acts of revision (cf. the logics for merging in Girard (2008), Liu (2011)).

2 Correspondence for information update and knowledge

We start with a phenomenon that is not very interesting in the AGM style, though it becomes wildly exciting when we study it in a constructive setting: update with new hard information that shrinks agents' current ranges of options for the actual situation.

2.1 Hard information, knowledge, and public announcement logic

Basic epistemic logic We start by recalling some basics. Standard epistemic logic *EL* describes semantic information encoded in agents' ranges of uncertainty. The language extends propositional logic with modal operators $K_i\phi$ (i knows that ϕ), for agents i, and $C_G\phi$ (ϕ is common knowledge in group G). Epistemic models $\mathbf{M} = (W, \{\sim_i\}_{i\in I}, V)$ have a set of worlds W, accessibility relations $\sim_i$ for agents i in some total group I, and a valuation V for proposition letters. Pointed models $(\mathbf{M}, s)$ mark an actual world s.[5] The key truth condition is that $\mathbf{M}, s \models K_i\phi$ iff for all worlds t with $s \sim_i t$: $\mathbf{M}, t \models \phi$.[6][7] Complete logics capturing epistemic reasoning about oneself and others are known (Fagin et al. 1995). The base system is a minimal modal logic. A restriction to equivalence relations adds S5 axioms of positive and negative introspection, while the complete logic of common knowledge can be axiomatized with *PDL*-techniques.

Information update by elimination Now for the logical dynamics of information flow. An event $!\phi$ yielding the information that ϕ is true shrinks the current model to just those worlds that satisfy ϕ. This is the well-known notion of *public hard information*. More precisely, for any epistemic model $\mathbf{M}$, world s, and formula ϕ true at s, the new $(\mathbf{M}|\phi, s)$ ($\mathbf{M}$ *relativized to* ϕ *at* s) is the sub-model

[5] Further relational conditions on $\sim_i$ encode special assumptions about agents' powers of observation and introspection: very common is the special case of equivalence relations.

[6] As for common knowledge, $\mathbf{M}, s \models C_G\phi$ iff for all worlds t that are reachable from s by some finite sequence of arbitrary $\sim_i$ steps ($i \in G$): $\mathbf{M}, t \models \phi$.

[7] In what follows, for convenience, we mostly suppress agent indices, and use standard modal notation for the epistemic modality of one accessibility relation R. Also for convenience, we will work mostly with existential modalities $\Diamond$ instead of universal boxes $\Box$.

of $\mathbf{M}$ whose domain is the set $\{t \in \mathbf{M} \mid \mathbf{M}, t \models \phi\}$. This mechanism models public communication, but also public observation. There is much more to this dynamics than meets the eye in standard views of 'mere update' with factual formulas. For instance, crucially, truth values of complex epistemic formulas may change after update: agents who did not know that ϕ now do. Therefore, it makes sense to get clear on the exact dynamic logic behind this.

Public announcement logic The language of *public announcement logic PAL* adds action expressions to *EL*, plus matching modalities, defined by the syntax rules:

Formulas	$F:$	$p \mid \neg\phi \mid \phi \vee \phi \mid K_i\phi \mid C_G\phi \mid \langle A\rangle\phi$
Action Expressions	$A:$	$!F$

The semantic clause for the dynamic action modality looks ahead between models:

$$\mathbf{M}, s \models \langle !\phi\rangle\psi \quad \text{iff} \quad \mathbf{M}, s \models \phi \text{ and } \mathbf{M}|P, s \models \psi$$

PAL is axiomatized by any complete logic over static models plus the following crucial *recursion axioms*:

$$\begin{array}{lcll}
\langle !\phi\rangle q & \leftrightarrow & (\phi \wedge q) & \text{for proposition letters } q \\
\langle !\phi\rangle(\psi \vee \chi) & \leftrightarrow & (\langle !\phi\rangle\psi \vee \langle !\phi\rangle\chi) & \\
\langle !\phi\rangle\neg\psi & \leftrightarrow & (\phi \wedge \neg\langle !\phi\rangle\psi) & \\
\langle !\phi\rangle\Diamond\psi & \leftrightarrow & (\phi \wedge \Diamond\langle !\phi\rangle\psi) &
\end{array}$$

Intuitively, the final recursion axiom for knowledge captures the essence of getting hard information. We will see in just which sense this is true in our further analysis. For further theory and applications of *PAL* and related systems, cf. Baltag et al. (1998), van Ditmarsch et al. (2007), van Benthem (2011).

2.2 Switching directions: from valid axioms to constraints

PAL is about one constructive way of taking incoming hard information: elimination of incompatible worlds. Now we reverse the perspective. Let us ask

which postulates look plausible for hard update, of course, always keeping in mind that our intuitions need to be valid for arbitrary propositions, bringing the logic in harmony.[8] Having done that, we can see which transformations of models validate them. This sounds grand. In what follows, however, I take a simple approach, investigating the recursion axioms of *PAL* themselves as postulates, since they have a lot of general appeal. To make this work, we need a suitably abstract setting – close to the models of *DDL*.[9]

Update universe and update relations Consider any family M of pointed epistemic models $(\mathbf{M}, s)$, viewed as an 'update universe' where model changes can take place. Possible changes are given as a family of update relations $R_P(\mathbf{M}, s)(\mathbf{N}, t)$ relating pointed models, where the index set P is a subset of $\mathbf{M}$: intuitively, the proposition triggering the update. One can think of the R as recording the action of some update operation $\heartsuit$ occurring in the syntax of our language that depends on the proposition P. Here different operations can have different effects: from our hard updates $!\phi$ to the soft updates $\Uparrow \phi$ to be discussed below. As just said, this is essentially the semantic setting of Krister Segerberg's dynamic doxastic logic, where each transition relation has a matching modality.[10] Now, for each formula ϕ, let $[[\phi]]$ be the set of worlds in $\mathbf{M}$ satisfying ϕ. We set, for the update modality matching R:

$$\mathbf{M}, s \models \langle \heartsuit \phi \rangle \psi$$

iff

there exists a model $(\mathbf{N}, t)$ in M with $R_{[[\phi]]}(\mathbf{M}, s)(\mathbf{N}, t)$ and $(\mathbf{N}, t) \models \psi$

[8] It is a curiously overlooked mismatch that modal logics for philosophical notions are often based on philosophers' intuitions about factual statements only, whereas the logic itself also deals with complex assertions that make good sense, for which the philosophers' intuitions might have to be different. Other imbalances of this sort occur in logics for non-standard consequence relations, and accounts of knowledge proposed in formal epistemology.

[9] The setting chosen here is more abstract and flexible than that used in the correspondence analaysis of van Benthem (2011), and it removes some infelicities in that earlier treatment.

[10] This is not the only possible format, and one can experiment with others. In particular, making the relational transition depend on just an extensional set of worlds reflects the valid *PAL* rule of *Replacement of Provable Equivalents*. Stated as one axiom in a language extended with a universal modality U ranging over the whole universe, this is the following implication making announced propositions 'extensional': $U(\phi \leftrightarrow \psi) \rightarrow (\langle !\phi \rangle \alpha \leftrightarrow \langle !\psi \rangle \alpha)$.

Remark To be yet more precise, we are really interpreting our language in a *three-index format* $\mathsf{M}, \mathbf{M}, s$, and for the accessibility relations R in this update universe M, we have that $(\mathbf{M}, s)R(\mathbf{M}, t)$ iff Rst in $\mathbf{M}$, without any jumps out of the model $\mathbf{M}$. This precision can be ignored for most of what follows, but it will come up occasionally.

2.3 A correspondence theorem for eliminative update

In what follows, the reader is supposed to know how modal frame correspondence works: cf. the textbooks (Blackburn et al. 2000, van Benthem 2010). We will analyze the *PAL* recursion axioms one by one in this style to see what they say, as a way of determining their total content as a correspondence constraint on update operations. But before doing so, we need to address a subtlety.

Substitution closure Correspondence arguments use frame truth of modal formulas, i.e., truth under all possible valuations for the proposition letters. Thus, if a formula is true, so are all its substitution instances: proposition letters are schematic variables for arbitrary propositions. But this sits badly with the system *PAL*, whose valid principles are not closed under substitution. In particular, the base axiom $\langle !\phi\rangle q \leftrightarrow (\phi \wedge q)$ is only valid for proposition letters q. Substituting to the general form $\langle !\phi\rangle\psi \leftrightarrow (\phi \wedge \psi)$ yields obviously invalid instances for epistemic assertions ψ. Much can be said about this phenomenon (Holliday et al. 2011), but in this paper, we take a simple line. We will first analyze the substitution-closed principles of *PAL*, and then return to the correspondence status of the base axiom. Thus, for the moment, we only look at the following obviously substitution-closed special case:

$$\langle !\phi\rangle \top \leftrightarrow \phi$$

In our correspondence setting, substitution failures relate to the semantics of atomic propositions p. Inside one epistemic model $\mathbf{M}$, the obvious choice seems to be sets of worlds. But in an update universe M as above, propositions range over all *pairs* $(\mathbf{M}, s)$, and hence one p could have different truth values at pairs $(\mathbf{M}, s)$, $(\mathbf{N}, s)$. We will view Greek letters in axioms as standing for such general context-dependent propositions in what follows, returning to the original view of *PAL*-atoms as sets of worlds later on. Finally, here is one more important convention in what follows:

Remark Throughout, we will fix announced formulas ϕ in contexts $\langle !\phi\rangle\psi$, refraining from varying these in correspondence. Think of distinguished fixed propositions.

Now we are ready to go through the crucial axioms that make *PAL* tick:

Base axiom The axiom $\langle !\phi\rangle\top \leftrightarrow \phi$ says that, given any model **M**, the domain of the transition relation $R[[\phi]]$ is the set of worlds satisfying ϕ in **M**. In other words, our abstract update action has the truth of ϕ as a necessary and sufficient precondition.

Disjunction axiom There is no special constraint expressed by the modal formula $\langle !\phi\rangle(\psi \vee \chi) \leftrightarrow \langle !\phi\rangle\psi \vee \langle !\phi\rangle\chi$, since this law holds for any transition relation.

Negation axiom One direction of this axiom expresses no constraint on the update operation: $(\phi \wedge \neg\langle !\phi\rangle\psi) \rightarrow \langle !\phi\rangle\neg\psi$ is valid, given that is equivalent to $\langle !\phi\rangle\top$. But the converse $\langle !\phi\rangle\neg\psi \rightarrow (\phi \wedge \neg\langle !\phi\rangle\psi)$, even just $\langle !\phi\rangle\neg\psi \rightarrow \neg\langle !\phi\rangle\psi$, says by a standard correspondence argument that the transition relation is a partial function:[11]

$$\text{if } (M,s)R[[\phi]](N,t) \text{ and } (M,s)R[[\phi]](K,u), \text{ then } (N,t) = (K,u).$$

Using this observation, we now simplify the original transition relations R_P in the update universe to *partial functions* F_P on pointed models. In particular, given any model **M** with a subset P, we can meaningfully talk about its image $F_P[\mathbf{M}]$.

Knowledge axiom So far, we were just doing preliminaries. The heart of the matter is evidently the recursion axiom for knowledge: $\langle !\phi\rangle\Diamond\psi \leftrightarrow (\phi \wedge \Diamond\langle !\phi\rangle\psi)$. The two directions of this clearly express two constraints on the update function – and together, they enforce a well-known model-theoretic notion from modal logic (Segerberg 1971):

[11] The above comment on interpreting propositions is crucial here: in the argument, we use the singleton set of the pointed model $(\mathbf{N}, t)$ as the denotation of ψ in the update universe M.

Fact. *The update function satisfies frame truth of* $\langle !\phi\rangle\Diamond\psi \leftrightarrow (\phi \wedge \Diamond\langle !\phi\rangle\psi)$ *iff every map* F_P *is a* p-morphism *between* $\mathbf{M}$ *and* $F_P[\mathbf{M}]$.

Proof. We do this first proof in a bit of detail, mainly to show how simple correspondence arguments for update functions are. Consider any model $\mathbf{M}$, with $[[\phi]] = P$. First we show that F_P is a homomorphism. Suppose that Rst in $\mathbf{M}$, with s, t both in the domain of F_P. Now set $V(\psi) = \{F_P(t)\}$. Then $(\mathbf{M}, s) \models \phi \wedge \Diamond\langle !\phi\rangle\psi$, and therefore also, $(\mathbf{M}, s) \models \langle !\phi\rangle\Diamond\psi$. By the definition of $V(\psi)$, this implies that $RF_P(s)F_P(t)$. Next, for the backward clause of being a p-morphism, suppose that $RF_P(s)u$, and now set $V(\psi) = \{u\}$. Then we have $(\mathbf{M}, s) \models \langle !\phi\rangle\Diamond\psi$. It follows from the truth of our axiom that $(\mathbf{M}, s) \models \phi \wedge \Diamond\langle !\phi\rangle\psi$, and hence there exists a t in $\mathbf{M}$ with Rst and $F_P(t) = u$. □

Collecting all our observations so far, we have the following result:

Theorem. *An update universe satisfies the substitution-closed principles of* PAL *iff its transition relations* F_P *are partial p-morphisms defined on the sets* P.

Discussion This is not quite the formation of *submodels* in standard elimination. Here is why. First, having a p-morphism is enough for validity of the *PAL* axioms, so we found a generalization of the standard semantics that may be of independent interest. Also, contracting several worlds into one during update occurs naturally in the setting of *PAL*: cf. van Benthem (2011).[12]

The base axiom once more Still, the above outputs enforced by our update mechanism are relational subframes, rather than submodels. What about the atomic propositions? *PAL* update assumes that these stay the same when a world does not change. Here is how we can think of this. Consider the usual proposition letters of epistemic logic as distinguished atomic propositions. The base axiom tells us that these special propositions have a special behavior: if they hold for an pointed model $(\mathbf{M}, s)$, they also hold for any of its update images under a map F_P, and vice versa:

$$(\mathbf{M}, s) \models p \text{ iff } F_P(\mathbf{M}, s) \models p$$

[12] If one insists on making the maps one-to-one, this can be done by enriching the modal language, and enforcing one more reduction axiom for public announcement, namely, for the *difference modality* $D\psi$ saying that ψ holds in a least one different world.

This might be the only content to the base axiom: update maps respect distinguished atomic propositions. But we can say a bit more in correspondence style. We assumed that proposition letters ranged over all sets of pointed models in the update universe. Now introduce special 'context-independent' proposition letters q ranging only over special sets of pointed models, with the property that they only depend on worlds:

$$(\mathbf{M}, s) \models q \text{ iff } (\mathbf{N}, s) \models q, \quad \text{for all models } \mathbf{M}, \mathbf{N} \text{ in } \mathsf{M}$$

Fact. *An update universe satisfies the base axiom* $\langle !\phi\rangle q \leftrightarrow (\phi \wedge q)$ *for all context-independent* q *iff the update maps are the identity on worlds:* $F_P(\mathbf{M}, s) = (\mathbf{N}, s)$ *for some model* $\mathbf{N}$.

Proof. Consider a pointed model $(\mathbf{M}, s)$ in the domain of F_P. Now set $V(q) = \{(\mathbf{N}, s) \mid (\mathbf{N}, s) \text{ is in } \mathsf{M}\}$. This is clearly a context-independent predicate. Taking this as $V(q)$, the true implication $(\phi \wedge q) \rightarrow \langle !\phi\rangle q$ says that $F_P(\mathbf{M}, s) = (\mathbf{N}, s)$ for some model $\mathbf{N}$. □

Even so, models $\mathbf{N}$ occurring in F_P-values for pointed models $(\mathbf{M}, t)$ with the same $\mathbf{M}$ could still differ. We will soon see a further recursion law making this uniform.[13]

This concludes our discussion of the correspondence content of the *PAL* axioms.[14]

2.4 Variations, extensions, and a provocation

Recursion axioms as general postulates We have determined the update content of one specific axiom for update. But there is more to this. Dynamic-epistemic recursion axioms are not just 'any sort of principle'. They have several features that make them candidates for general postulates on informa-

[13] For an analogy, think of correspondence theory for intuitionistic logic (Rodenburg 1986), where axioms are only valid for all 'hereditary propositions'.

[14] Readers who like open problems may ponder this: how should the above analysis be modified to allow *factual change*, as in van Benthem et al. (2006)?

tion update.[15] In particular, our analysis says that the *PAL* recursion axiom for knowledge expresses a sort of *partial bisimulation* between the original model and the output of an update rule applied to it. I find abstract simulation behavior very appealing as a general semantic constraint on update functions, though I am not sure how to define it in its proper generality.[16]

Protocols Update universes also suggest a different setting, that has been proposed in dynamic-epistemic logic for independent reasons. So far, we had that $\langle !\phi\rangle\top \leftrightarrow \phi$. This says that executing an action $!\phi$ requires truth of the precondition ϕ, but also, whenever ϕ is true, $!\phi$ can be executed. But in civilized conversation or regimented inquiry, the latter assumption is often untenable. To represent this, 'protocol models' make restrictions on propositions that can be announced or observed. Hoshi (2009) shows how *PAL* changes in this setting, since the earlier recursion axioms will now be valid only with $\langle !\phi\rangle\top$ in the place of ϕ on their right-hand sides. This move has many technical repercussions, though the system remains axiomatizable and decidable. From our correspondence perspective, nothing much changes: the only new thing is that the domain of an update map F_P will now be a subset of P, but not necessarily all of P. Our analysis of the modified recursion axioms remains essentially as before.

Language extensions We analyzed update axioms for the epistemic base language. But *PAL* also has a complete version for the full epistemic language with common knowledge. The recursion axiom then requires a new notion of 'conditional common knowledge' (van Benthem et al. 2006). Since the axiom for single-agent knowledge already fixed the *PAL* update rule, as we have seen, no further constraints arise. We will return later to what this 'passive behavior' of common knowledge vis-à-vis single-agent knowledge means in terms of definability or derivability.[17] A useful language extension whose recursion axiom

[15] The commutation of action and knowledge in the key PAL recursion axiom has an appealing interpretation in terms of desirable features of logically well-endowed agents. It expresses notions of *Perfect Recall* and *No Miracles* in the sense of Halpern and Vardi (1989).

[16] A relevant analogy here may be with the modal logic of a *bisimulation* Z itself, viewed as a relation on a universe whose worlds are models. The key back-and-forth clause of bisimulation is precisely a commutation axiom $\langle Z\rangle\Diamond\psi \leftrightarrow \Diamond\langle Z\rangle\psi$.

[17] There is also the question whether the recursion axiom for conditional common knowledge by itself fixes world elimination as the update rule – but we will consider this issue only with an analogous case in the dynamic logic of belief change.

does add to our correspondence analysis introduces an *existential modality* $E\psi$ saying that ψ is true in some world in the current model, accessible or not. In update universes M, we interpret this as saying, at a pointed model $(\mathbf{M}, s)$, that there is some t in $\mathbf{M}$ with ψ true at $(\mathbf{M}, t)$.

Fact. *On update universes* M *satisfying the earlier* PAL *update conditions, the axiom* $\langle !\phi\rangle E\psi \leftrightarrow (\phi \wedge E\langle !\phi\rangle\psi)$ *is frame-true iff, for every model* $\mathbf{M}$*, the update images of worlds in* $\mathbf{M}$ *have the same model* $\mathbf{N}$ *throughout.*

Proof. First, the axiom is clearly valid in the intended update universes. Conversely, its right to left direction implies the stated property. Consider any two worlds $(\mathbf{N}, t)$, $(\mathbf{K}, u)$ in the image $F_P(\mathbf{M})$. Set $V(\psi) = \{(K, u)\}$. Then the F_P-original of $(\mathbf{N}, t)$ in $\mathbf{M}$ satisfies $\phi \wedge E\langle !\phi\rangle\psi$. It follows that $\langle !\phi\rangle E\psi$, and by the preceding definition, this only happens when $(\mathbf{N}, t)$ and $(\mathbf{K}, u)$ share the same model component. □

Finally, update universes suggest yet further language extensions. For instance, there is also a natural relation $(\mathbf{M}, s) \sim (\mathbf{N}, s)$ holding between different models sharing the same distinguished world. Its modality would make sense, even though it does not make sense inside single epistemic models, the way basic epistemic logic works.

What is the right version of *PAL*? We conclude with a more provocative feature of our analysis. We started by analyzing what standard public announcement says about update, and then determined its force in update universes. But doing so involved a natural distinction between the substitution-closed principles of *PAL* and the more 'accidental' base axiom holding only for a restricted class of valuations. So, what is 'public announcement logic' after all? Is its base semantics perhaps the one on update universes with context-dependent propositions and substitution-closed validities? And if so, is what we call the 'standard version' perhaps an accident of formulation?

2.5 Other natural operations: link cutting

Update with hard information that ϕ does show variety beyond the above elimination. In a well-known *link-cutting* variant, the operation $|\phi$ performed announces *whether* ϕ is the case. This means that the domain of worlds stays

the same, but all epistemic links get cut between ϕ-worlds and $\neg\phi$-worlds in the current model – an operation used by many authors. The changes induced in the *PAL* axioms are mainly these:

$$\langle|\phi\rangle q \leftrightarrow q \text{ (this implies the substitution-closed instance } \langle|\phi\rangle\top)$$
$$\langle|\phi\rangle\Diamond\psi \leftrightarrow ((\phi \wedge \Diamond(\phi \wedge \langle|\phi\rangle\psi)) \vee (\neg\phi \wedge \Diamond(\neg\phi \wedge \langle|\phi\rangle\psi)))$$

The following result can be proved in the same correspondence style as before:

Fact. *Link cutting is the only model-changing operation that satisfies the reduction axioms for the dynamic modality* $\langle|\phi\rangle$.

Proof. We merely give a sketch of the substitution-closed part. Start from any pointed model $\mathbf{M}, s$. The modified base axiom tells us that the update map is now total on the whole domain of $\mathbf{M}$. Next, the recursion axiom for knowledge, read from left to right, says that the only links in the image come from already existing links between either ϕ-worlds, or $\neg\phi$-worlds. Finally, from right to left, the axiom says that all links of the two mentioned types existing in $\mathbf{M}$ get preserved into the image. □

3 Correspondence analysis of modal logics for belief change

Now that we have seen how to analyze principles of knowledge update by changing domains or accessibility relations, an extension to belief revision is straightforward. We mainly need to decide what models we will be working with.

3.1 Soft information and belief

Doxastic models are structures $\mathbf{M} = (W, \{\leq_i\}_{i\in I}, V)$ where the $\leq_i$ are binary comparison relations $\leq_i xy$ saying that agent i considers x at least as plausible as y. As before, for convenience, we drop agent indices henceforth. These plausibility relations are usually taken to be reflexive and transitive, making the modal base logic *S4* – or also connected, like the 'Grove models' of belief revision

theory, making the logic *S4.3*. Such options are important in practice, but they do not affect the analysis to follow.

These models encode varieties of information. While the whole domain represents our current hard information in the earlier sense, the most plausible worlds in the ordering $\leq$ represent our *soft information* about the actual world. This soft information is the basis of our beliefs and actions based on these, but it is defeasible: the actual world may lie outside of the most plausible area, and we may learn this as a scenario unfolds. In this setting, belief is commonly interpreted as truth in all most plausible worlds:[18]

$$\mathbf{M}, s \models B\phi \text{ iff } \mathbf{M}, t \models \phi \text{ for all worlds } t \text{ that are minimal in the ordering } \leq$$

But absolute belief does not suffice for most purposes. We also need the notion of *conditional belief*:[19]

$$\mathbf{M}, s \models B^{\psi}\phi \text{ iff } \mathbf{M}, t \models \phi \text{for all } \leq\text{-minimal worlds in } \{u \mid \mathbf{M}, u \models \psi\}$$

This point returns with recursion axioms for belief change. From a systematic logical perspective, we should not analyze changes in beliefs only (the usual practice in belief revision theory), but also changes in conditional belief.

Conditional logic Complete logics for conditional belief can be found in close analogy with *conditional logic* based on similarity semantics (Lewis 1973). One difference is that conditional models usually involve a *ternary* comparison ordering $\leq zxy$: world x is closer to world z then world y. A generalization from binary to ternary relation also makes sense for plausibility semantics of belief, but we forego this here. [20]

Safe belief While the preceding belief modalities are interesting, it has become clear recently that the plain base modality of plausibility models has independent interest.

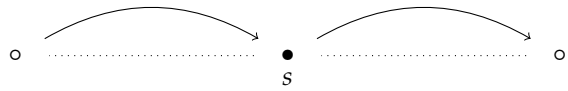
$$\mathbf{M}, s \models \langle\leq\rangle\phi \text{ iff there exists a } t \geq s \text{ with } \mathbf{M}, t \models \phi$$

[18] We disregard some modifications of truth clauses needed with infinite models.

[19] Absolute belief can be retrieved as the special case of $\psi = \top$.

[20] Another natural generalization are *epistemic-doxastic models* $\mathbf{M} = (W, \{\sim_i\}_{i\in I}, \{\leq_i\}_{i\in I}, V)$ allowing for both knowledge update and belief revision. Our methods also work there.

The corresponding universal modality offers an interesting doxastic notion in between knowledge and belief. Consider this picture with the actual world s in the middle:



$K\phi$ describes what we know: ϕ must be true in all three worlds in the range, less or more plausible than the current one. $B\phi$ describes beliefs, which have to be true in the right-most world only. Now $[\leq]\phi$ describes our safe beliefs, referring to the actual s plus the right-most world. These cannot be refuted by any future correct observations. Technically, safe belief can also define the other kinds of belief (Boutilier 1994):

$$\text{on finite pre-orders, } B^{\psi}\phi \text{ is defined by } U(\psi \rightarrow \langle\leq\rangle(\psi \wedge [\leq](\psi \rightarrow \phi)))$$

with U the universal modality, or in epistemic-doxastic models, a knowledge modality. Thus, an analysis of belief change might focus on safe belief without losing much.

3.2 Dynamic logics of belief change

Now we can write complete logics for belief change. Indeed, there are several systems for this, depending on what kind of new information triggers the relevant change in the structure of the current model.[21]

Hard information For hard information, the complete dynamic logic is as follows:

Theorem. *The logic of conditional belief under public announcements is axiomatized completely by*

- *any complete static logic for the model class chosen,*
- *the* PAL *recursion axioms for atomic facts and Boolean operations,*

[21] The results cited in this subsection and the next are from van Benthem (2007).

- *an axiom for conditional belief:* $\langle !\phi\rangle B^{\alpha}\psi \leftrightarrow (\phi \wedge B^{\langle !\phi\rangle\psi}\langle !\phi\rangle\psi)$.

A similar analysis can be given for safe belief, with a simpler key recursion axiom

$$\langle !\phi\rangle\langle\leq\rangle\psi \leftrightarrow (\phi \wedge \langle\leq\rangle\langle !\phi\rangle\psi)$$

Formally, this is just the earlier recursion axiom for a modality $\Diamond$.

Soft information and plausibility change Now comes a major further step. Triggers for belief change can be of many kinds, and we do not always expect the same model changes. In particular, incoming new information may be soft rather than hard, which means that it does not eliminate worlds, but merely rearranges the current plausibility order. A common example is a radical upgrade $\Uparrow \phi$ changing the current ordering $\leq$ between worlds in a model $(\mathbf{M}, s)$ to a new model $(\mathbf{M} \Uparrow \phi, s)$ as follows:

> all ϕ-worlds in the current model become better than all $\neg\phi$-worlds,
> while, within those two zones, the old plausibility ordering remains.

Like for public announcement, we introduce an upgrade modality into our language:

$$\mathbf{M}, s \models \langle\Uparrow \phi\rangle\psi \text{ iff } \mathbf{M} \Uparrow \phi, s \models \psi$$

The earlier techniques extend. Again there is a complete set of recursion axioms:

Theorem. *The dynamic logic of lexicographic upgrade is axiomatized by*

- *any complete static logic for the model class chosen,*
- *the following recursion axioms:*

$$\begin{array}{lcl}
\langle\Uparrow \phi\rangle q & \leftrightarrow & q \qquad \textit{for all atomic proposition letters } q \\
\langle\Uparrow \phi\rangle\neg\psi & \leftrightarrow & \neg\langle\Uparrow \phi\rangle\psi \\
\langle\Uparrow \phi\rangle(\psi \vee \chi) & \leftrightarrow & \langle\Uparrow \phi\rangle\psi \vee \langle\Uparrow \phi\rangle\chi \\
\langle\Uparrow \phi\rangle B^{\alpha}\psi & \leftrightarrow & (E(\phi \wedge \langle\Uparrow \phi\rangle\alpha) \wedge B^{\phi\wedge\langle\Uparrow\phi\rangle\alpha}\langle\Uparrow \phi\rangle\psi) \\
 & & \vee (\neg(E(\phi \wedge \langle\Uparrow \phi\rangle\alpha)) \wedge B^{\langle\Uparrow\phi\rangle\alpha}\langle\Uparrow \phi\rangle\psi)
\end{array}$$

Again, there is also an evident valid recursion axiom that governs the induced changes in safe belief:

$$\langle\Uparrow \phi\rangle\langle\leq\rangle\psi \leftrightarrow E(\phi \wedge \langle\Uparrow \phi\rangle\psi) \vee (\neg\phi \wedge \langle\leq\rangle\langle\Uparrow \phi\rangle\psi)$$

Given the earlier modal definition of absolute and conditional belief in terms of safe belief, one can even derive the preceding recursion axioms from this one. Other belief change policies can be treated in the same style, using the relation transformers of van Benthem and Liu (2007) or the priority product update of Baltag and Smets (2008).

3.3 Correspondence for axioms of belief change

As before with knowledge, we can now invert the preceding results and use the key recursion axioms as constraints to determine the space of possible update operations. For update operations transforming plausibility relations only, leaving domains of models the same, a more complex correspondence proof than earlier ones shows:

Theorem. *The recursion axioms of the dynamic logic of radical upgrade hold universally for an update operation on a universe of pointed plausibility models iff that operation is in fact radical upgrade.*[22]

It is important to realize what is going on here. *AGM*-style postulates on changes in beliefs will not fix the relational transformation: we need to constrain the changes in conditional beliefs, since the new plausibility order encodes all of these. A similar analysis works for other revision policies, such as 'conservative' belief change. But actually, there is an easier road to such results, closer to earlier arguments.

Theorem. *Radical upgrade is the only update operation validating the given recursion axioms for atoms, Booleans plus safe belief.*

Proof. Suppose that the axiom is valid on a universe of plausibility models. The axiom for atoms tells us in particular that our update function is defined everywhere. Now consider any model $(\mathbf{M}, s)$. From left to right, taking ψ to

[22] Here as before, we work with the substitution-closed version of the logic. In particular, the atomic case simplifies to just $\langle\Uparrow \phi\rangle\top$: radical upgrade is defined everywhere.

denote just one world $(\mathbf{N}, t)$ with $F_P(\mathbf{M}, s) \leq (\mathbf{N}, t)$, it follows that $(\mathbf{N}, t)$ was either the image of some ϕ-world in M, or $s \leq u$ in $\mathbf{M}$ for some world u mapped to $(\mathbf{N}, t)$, i.e., the new $\leq$-link came from an old one originating in a $\neg\phi$-world. This means that each new relational link comes from the set defined by radical upgrade. That in fact all such links occur in the F_P-image of $\mathbf{M}$ follows by similar unpacking of the reverse implication of the axiom. □

Given this last correspondence result, the earlier more complex ones seem less urgent, since safe belief defines absolute and conditional belief. Indeed, *AGM*-style postulates on 'safe-belief change' might be easier conceptually than those for regular belief.[23]

3.4 Discussion: generality of the analysis

We have seen how recursion laws in constructive logics of belief change can serve as general postulates to constrain, and almost uniquely fix, possible updates. As before, this relates the *DDL* and *DEL* approaches to modal logics of belief change, softening a contrast that we started out with. Also as before, issues of generality arise. Are the recursion axioms too specific for belief change postulates? Here we repeat our earlier intuition of 'simulation' between input and output models of the transformation. One might add that a recursive postulate may itself be philosophically attractive as providing the core 'dynamic equation' driving the process of update or revision. Finally, here is an issue more specific to belief. Given the overwhelming variety of belief revision policies, what is the general thrust of correspondence results like ours? We will return to this issue in Section 5, when discussing product update and other general mechanisms replacing separate revision rules by one master rule plus richer input.[24]

[23] Still, it is interesting that recursion axioms for conditional belief fix radical upgrade, too. This might imply further definability and proof-theoretic connections between the various doxastic notions mentioned. If one recursion axiom fixes update, it looks as if others should be derivable in some way. We cannot explore this technical line here.

[24] Here is a more technical issue. We have only analyzed single update mechanisms so far. But some *AGM*-postulates mix update and revision. Can we use modal versions of such postulates to get correspondence results for axioms with two update modalities simultaneously?

4 Richer formats as a test case

The style of analysis proposed here works on richer semantic formats for update than modal relational models. In this brief digression, we sketch two examples. These will also raise some issues about the scope and limitations of our earlier analysis.

4.1 Event models and product update

While public announcement logic *PAL* is a good pilot system, its restriction to public information makes it unsuitable for analyzing individual differences in observation and communication. A much richer dynamic-epistemic logic for the latter tasks is true *DEL* (Gerbrandy 1999, Baltag et al. 1998). It uses *action models* **E** that collect events with attached 'preconditions', with epistemic uncertainty links between events representing agents' observational access to what actually happens. Action models have been used to represent a wide variety of triggers for information change. Next, by performing *product update* of an action model **E** with the current epistemic or doxastic model **M** one obtains a new updated information model $\mathbf{M} \times \mathbf{E}$ displaying the right information for all agents involved after the event has taken place.

We assume that the reader knows how *DEL* update works, including its complete set of recursion axioms. We display two of these for later reference – suppressing agent indices as before, and using the letter R to denote the agent's accessibility relation:

$$\langle E, e\rangle \top \leftrightarrow Pre_e$$

$$\langle E, e\rangle \Diamond \psi \leftrightarrow (Pre_e \wedge \bigvee_{eRf \text{ in } \mathbf{E}} \Diamond \langle E, f\rangle \psi)$$

This mechanism changes epistemic or doxastic models much more drastically than the earlier world elimination or relation change. In particular, the set $\{(s, e) \mid s \in \mathbf{M}, e \in \mathbf{E}, \mathbf{M}, s \models Pre_e\}$, of worlds in $\mathbf{M} \times \mathbf{E}$ may grow beyond the size of the initial model **M**.

Theorem. *The recursion axioms for the dynamic modality* $\langle \mathbf{E}, e\rangle \phi$ *of* DEL *determine product update uniquely modulo* p*-morphism.*

The precise sense in which this fact is true will emerge from the following discussion.

Proof. (Sketch) As in our study of *PAL*, we analyze the impact of the *DEL* recursion axioms on an update universe of epistemic models with an abstract transition relation for the update for the pointed event model $(\mathbf{E}, e)$. The negation axiom of *DEL* tells us that this is a partial function $F_{\mathbf{E},e}$. This functionality means that we can think of values $F_{\mathbf{E},e}(\mathbf{M}, s)$ as pairs (s, e) without loss of information. Next, the substitution-closed base axiom tells us that $F_{\mathbf{E},e}$ is defined on those models $(\mathbf{M}, s)$ whose s satisfies the precondition of e in $\mathbf{E}$. Finally, also as before, the recursion axiom for individual knowledge puts constraints on the function $F_{\mathbf{E},e}$. First, if sRt in $\mathbf{M}$, and eRf in $\mathbf{E}$, while $F_{\mathbf{E},e}(\mathbf{M}, s)$, $F_{\mathbf{E},e}(\mathbf{M}, t)$ are both defined, then $(s, e)R(t, f)$ holds by the direction from right to left in the axiom. Vice versa, any link in the image of the model $\mathbf{M}$ must also arise in this way, if we unpack the left-to-right direction of the axiom.[25] □

One update logic to bind them all? The preceding analysis may still be too piecemeal, ignoring a key innovation of *DEL* in the area of constructive update logics. An earlier trend had been to define specific model changes for particular kinds of informational event: 'announcements that', link cutting 'announcements whether', or more complex types of private information flow, such as sending a *bcc* message over email. One gets different complete logics for each case. But *DEL* changed the game. All relevant structure triggering different updates is put in matching event models $\mathbf{E}$, and the logic for the special case is then a direct instance of the above 'mother logic' of $\langle \mathbf{E}, e \rangle \phi$. In this light, characterizing specific update functions may have some value, but the real logical insight is the general product update mechanism. Is this, then, the best constructive counterpart to a postulational approach to update?

Belief and priority update Similar points can be made about belief revision. One can capture complete logics for specific revision policies, as we have shown. But one can also work at the level of product update with 'plausibility event models', where agents now may think it more plausible that one event occurred rather than another. Update works with the priority rule that strict event

[25] This argument still ignores some key features of product update, like its use of ordered pairs (s, e) of worlds and events by themselves without marking the context *s in* $\mathbf{M}$, *e in* $\mathbf{E}$.

plausibility overrides prior plausibility:[26]

$$(s, e) \leq (t, f) \text{ iff } (s \leq t \wedge e \leq f) \vee e < f$$

The key recursion axiom for the 'mother logic' of priority update is given in Baltag and Smets (2008):[27]

$$\langle \mathbf{E}, e\rangle\langle\leq\rangle\phi \leftrightarrow (Pre_e \wedge (\bigvee_{e\leq f \text{ in } \mathbf{E}} \langle\leq\rangle\langle \mathbf{E}, f\rangle\phi \vee (\bigvee_{e<f \text{ in } \mathbf{E}} E\langle \mathbf{E}, f\rangle\phi))$$

We will not analyze this approach further, but this seems the most general dynamic-epistemic counterpart to the postulational approach of dynamic doxastic logic.[28]

4.2 Updating neighborhood models for evidence

It is hard to roam for long in modal logic without finding Krister Segerberg's traces. Another long-standing interest of his are neighborhood models (Segerberg 1971) that have been used recently as a model for the epistemological notion of evidence and its dynamics (cf. van Benthem and Pacuit (2011) for technical details of what follows).

Static neighborhood logic An epistemic accessibility relation encodes an agent's current range of worlds after some history of informational events. If we want to retain some of the latter 'evidence', a set of neighborhoods (sets of worlds) does well – where we think of the current range as the intersection of all evidence sets.[29] The simplest neighborhood models, and all that we consider here, have just one family N of sets on a domain of worlds. We then interpret an evidence modality as follows:

[26] As an illustration, an event model with two signals $!\phi$, $!\neg\phi$, with the first more plausible than the second, generalizes the above radical upgrade $\Uparrow \phi$, that typically also had this over-ruling character for worlds that satisfied the distinguished triggering proposition ϕ.

[27] Here E is the earlier existential modality over all worlds in the model, accessible or not.

[28] Other ways of achieving generality in constructive update logics include the *PDL*-style *program format* of van Benthem and Liu (2007), specifying intended relation changes in models. Girard et al. (2011) define a merge of action models and programs that represents realistic social scenarios. We leave a correspondence analysis to another occasion.

[29] If not all given sets overlap, we need more subtle views of conflicting evidence.

$$\mathbf{M}, s \models \Box\phi \text{ iff there is a set } X \text{ in } \mathsf{N} \text{ with } \mathbf{M}, t \models \phi \text{ for all } t \in X$$

The base logic of this notion is that of a monotone modality that does not necessarily distributive over either disjunction or conjunction. This generalization of modal logic supports correspondence analysis.[30] Neighborhood models support many epistemic notions. At least in finite models, one can define (cautious evidence-based) belief as what is true in all intersections of maximally overlapping families of evidence.[31]

Evidence dynamics: two samples In this setting, our pilot system *PAL* for information update can be seen as mixing different update actions into its public announcements $!\phi$. The first is evidence addition $+\phi$, adding the denotation $[[\phi]]$ in the current model as one more piece of evidence to the current evidence family N. The dynamic logic of this action can be determined completely. Here is one key recursion axiom:

$$\langle +\phi \rangle \Box\psi \leftrightarrow (\Box\langle +\rangle\phi \vee U(\phi \to \psi))$$

Again, the content of this principle can be determined by a straightforward correspondence argument:

Fact. *An abstract update function on a universe of neighborhood models satisfies the recursion axiom for evidence addition iff each new evidence set is a superset of either some old evidence set or of the set* $[[\phi]]$.[32]

A second aspect of a public announcement $!\phi$ that now gets into its own is removal of the evidence for $\neg\phi$. The general new operation $-\psi$ removes all evidence sets from the current family N that are included in $[[\psi]]$. Complete recursion axioms are known for removal and the evidence modality, as well as belief, though a considerable extension of the standard static modal base

[30] For instance, the *K*-axiom $\Box \bigwedge_i \psi_i \leftrightarrow \bigwedge_i \Box\psi_i$ forces N to be generated from a binary accessibility relation – provided we read it with an infinitary conjunction.

[31] There are links with modeling beliefs in relational plausibility models here that we ignore.

[32] Recursion axioms for new beliefs under evidence addition extend the base language for evidence models to *conditional belief* in two basic varieties that had not surfaced so far.

languages over evidence models is required. Here is one such principle, using a notion of evidence conditional on $\neg\phi$ being true:[33]

$$\langle -\phi \rangle \Box \psi \leftrightarrow (E\neg\phi \rightarrow \Box_{\neg\phi} \langle -\phi \rangle \psi)$$

We leave a correspondence analysis of recursion axioms for removal to future work. Clearly, we have only scratched the surface here, but hopefully, the reader has seen that our analysis still makes sense when the semantic modeling of dynamic epistemic logic undergoes a drastic neighborhood extension of a sort that Krister Segerberg has long ago proposed for dynamic doxastic logic (Segerberg 1995, Girard 2008).

5 Further directions

We have shown how modal correspondence brings together the postulational format of AGM theory and dynamic doxastic logic with the constructive model transformation style of dynamic-epistemic logic. Our technical illustrations were very simple, and we opened up more new problems than closing old ones. Several technical and conceptual issues were already raised in the text. In this section we briefly mention a few more.

Extended semantic formats We have worked with binary accessibility relations for knowledge and belief. This analysis should be extended to ternary relational models, where plausibility can be world-dependent. Likewise, the analysis needs to be taken to the realm of neighborhood models, a natural finer modeling for belief and evidence.

Group knowledge and belief At the start of this paper, we said that a multi-agent perspective is crucial to *DEL*-style logics, but soon this social aspect vanished. One should also analyze update postulates for common knowledge or belief in our style.[34]

[33] This is remarkable, since dealing with operations of contraction or removal has long been considered a stumbling block to constructive update logics. The reason why it works in the neighborhood setting after all is the richer model structure one is working on.

[34] No complete dynamic logic has been given yet for changes in common belief produced by radical upgrade. Technical difficulties here might require a redesign of the base language to an

'Dancing with the stars': propositional dynamic logic Common knowledge or belief go beyond the modal base language, being iterated modalities as found in dynamic logic *PDL*: another lifelong interest of Krister Segerberg. Iteration occurs naturally in dynamic-epistemic logic, also in the dynamic action component, as with repeated announcement or measurement. The resulting logical systems can be highly complex: cf. Miller and Moss (2005) on *PAL* with iteration, and Baltag and Smets (2009) on limit phenomena with iterated radical update. Still *PDL* is no obstacle to our analysis. There have been some striking advances in the treatment of modal frame correspondence for non-first-order principles like Löb's Axiom for provability logic or Segerberg's Axiom for dynamic logic, making them fall under an extended Sahlqvist syntax matching the system *LFP+FO*, first-order logic with added fixed-point operators. New results and references are found in van Benthem et al. (to appear).

Temporal setting and procedural information Both dynamic doxastic logic and *DEL* focus on single update steps. But equally essential is the temporal horizon. We make sense of local event in terms of global scenarios: a conversation, a process of inquiry, or a game. This 'procedural information' (Hoshi 2009) suggests interfacing dynamic logics with temporal logics of knowledge and belief (Parikh and Ramanujam 2003, Belnap et al. 2001, Bonanno 2007). Existing results at this interface take the form of representation theorems for 'update evolution': cf. van Benthem et al. (2009). One obvious question is how our correspondence results relate to representation theorems in the area of logics of belief (Dégrémont 2010).

General model theory The proofs in this paper were very simple. The recursion axioms all had Sahlqvist syntax (Blackburn et al. 2000). One would like a correspondence analysis of axioms for belief change at the latter level of generality. Moreover, correspondence is not the only abstract analysis of concrete modal logics. The mechanism of model change behind the dynamic-epistemic logics in this paper invites reflection on their general features as modal logics. In an earlier book for Krister Segerberg, I gave a Lindström Theorem capturing basic modal logic in terms of bisimulation invariance and compactness. It would be of interest to take this further to capture the essentials of dynamic modal logics of model change.

analogue of the 'epistemic *PDL*' of van Benthem et al. (2006), a system defined for the purpose of stating recursion axioms for common knowledge with product update.

Coda: have we really dealt with all logics of belief change? Do our two protagonists of dynamic-doxastic and dynamic-epistemic logic exhaust the field? My first attempt at doing modal logic of belief revision in van Benthem (1987) worked over a universe of information stages in the style of Beth or Kripke models for intuitionistic logic. An update with hard information was defined as a minimal upward move to a stage where the new information holds, while revision involved backtracking to the past and then going forward again to incorporate new information in conflict with what we thought so far. I am not sure how this third view relates to either *DDL* or *DEL*, though it, too, offers abstract spaces for a wide array of update actions.

6 Conclusion

We have shown how the two main logic approaches to belief change, Segerberg's dynamic doxastic logic and the *DEL* tradition, co-exist in the perspective of modal frame correspondence. Indeed, 'modal logic of belief revision' has two dual aspects that belong together. This much was our contribution to translatability and interaction between frameworks. Our evidence was a set of very simple technical observations – but around these, many new problems came to light. To me, this agenda of unknowns seems a virtue of the proposed analysis. Krister and I have our work cut out for us. Finally, a confession is in order. In starting this study, I thought the main beneficiary would be *DDL*, as it could now import new ideas from the pressure-cooker of *DEL*. But as will be clear at various places in the paper, I now feel that a correspondence perspective also raises serious issues about best design for dynamic-epistemic logics, rethinking their striking deviant feature of being non-substitution-closed. And so, I submit that both sides benefit from the style of analysis presented here.

References

G. Aucher. A combined system for update logic and belief revision. Master of logic thesis, ILLC, University of Amsterdam, 2004.

A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. In G. Bonanno, W. van der Hoek, and M. Wooldridge, editors, *Texts in Logic and Games Vol. 3*, pages 9–58. Amsterdam University Press, 2008.

A. Baltag and S. Smets. Group belief dynamics under iterated revision: Fixed points and cycles of joint upgrades. In *Proceedings of TARK XII, Stanford*, pages 41–50. Morgan Kaufmann Publishers, 2009.

A. Baltag, L. Moss, and S. Solecki. The logic of public announcements, common knowledge and private suspicions. In *Proceedings of TARK, Los Altos*, pages 43–56. Morgan Kaufmann Publishers, 1998.

N. Belnap, M. Perloff, and M. Xu. *Facing the Future. Agents and Choice in Our Indeterminist World*. Oxford University Press, 2001.

P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Cambridge University Press, Cambridge, 2000.

G. Bonanno. Axiomatic characterization of the AGM theory of belief revision in a temporal logic. *Artificial Intelligence*, 171:144–160, 2007.

C. Boutilier. Conditional logics of normality: a modal approach. *Artificial Intelligence*, 68:87–154, 1994.

C. Dégrémont. *The Temporal Mind. Observations on the logic of belief change in interactive systems*. Dissertation, ILLC, University of Amsterdam, 2010.

R. Fagin, J. Halpern, Y. Moses, and M. Vardi. *Reasoning about Knowledge*. The MIT Press, Cambridge (Mass.), 1995.

P. Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. The MIT Press, Cambridge (Mass.), 1988.

J. Gerbrandy. *Bisimulations on Planet Kripke*. Dissertation, ILLC, University of Amsterdam, 1999.

P. Girard. *Modal Logics for Belief and Preference Change*. Dissertation, Department of Philosophy, Stanford University (ILLC-DS-2008-04), 2008.

P. Girard, F. Liu, and J. Seligman. A product model construction for PDL. Departments of Philosophy, Auckland University, and Tsinghua University, 2011.

J. Halpern and M. Vardi. The complexity of reasoning about knowledge and time, I: Lower bounds. *Journal of Computer and Systems Science*, 38:195–237, 1989.

W. Holliday, T. Hoshi, and T. Icard. Schematic validity in dynamic epistemic logic: Decidability. In H. van Ditmarsch, J. Lang, and S. Ju, editors, *Proc's LORI-III, Guangzhou*, number 6953 in Springer Lecture Notes in Computer Science, pages 87–96, 2011.

T. Hoshi. *Epistemic Dynamics and Protocol Information*. Ph.d. thesis, Department of Philosophy, Stanford University (ILLC-DS-2009-08), 2009.

H. Leitgeb and K. Segerberg. Dynamic doxastic logic: why, how, and where to? *Synthese*, 155:167–190, 2007.

D. Lewis. *Counterfactuals*. Blackwell, Oxford, 1973.

F. Liu. *Reasoning about Preference Dynamics*. Synthese Library. Springer, 2011.

J. Miller and L. Moss. The undecidability of iterated modal relativization. *Studia Logica*, (97):373–407, 2005.

R. Parikh and R. Ramanujam. A knowledge based semantics of messages. *Journal of Logic, Language and Information*, 12:453–467, 2003.

P. Rodenburg. *Intuistionistic Correspondence Theory*. Dissertation, Mathematical Institute, University of Amsterdam, 1986.

H. Rott. Information structures in belief revision. In J. van Benthem and P. Adriaans, editors, *Handbook of the Philosophy of Information*, pages 457–482. Elsevier Science Publishers, Amsterdam, 2007.

K. Segerberg. An essay in classical modal logic. Philosophical Institute, University of Uppsala, 1971.

K. Segerberg. Belief revision from the point of view of doxastic logic. *Bulletin of the IGPL*, 3:534–553, 1995.

K. Segerberg. Default logic as dynamic doxastic logic. *Erkenntnis*, 50:333–352, 1999.

J. van Benthem. Semantic parallels in natural language and computation. In *Logic Colloquium, Granada*, pages 331–375, Amsterdam, 1987. North-Holland.

J. van Benthem. Dynamic logic of belief revision. *Journal of Applied Non-Classical Logics*, 17:129–155, 2007.

J. van Benthem. *Modal Logic for Open Minds*. CSLI Publications, Stanford, 2010.

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

J. van Benthem and F. Liu. Dynamic logic of preference upgrade. *Journal of Applied Non-Classical Logics*, 17:157–182, 2007.

J. van Benthem and E. Pacuit. Dynamic logic of evidence-based belief. *Studia Logica*, 99(1):61–92, 2011.

J. van Benthem, J. van Eijck, and B. Kooi. Logics of communication and change. *Information and Computation*, 204:1620–1662, 2006.

J. van Benthem, J. Gerbrandy, T. Hoshi, and E. Pacuit. Merging frameworks for interaction. *Journal of Philosophical Logic*, 38:491–526, 2009.

J. van Benthem, G. Bezhanisvili, and I. Hodkinson. Sahlqvist correspondence for modal μ-calculus. *Studia Logica*, to appear.

H. van Ditmarsch. Prolegomena to dynamic logic for belief revision. *Synthese*, 147:229–275, 2005.

H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*. Cambridge University Press, Cambridge, 2007.

Epistemic planning for single- and multi-agent systems

Thomas Bolander and Mikkel Birkegaard Andersen

DTU Informatics, Technical University of Denmark
Richard Petersens Plads, building 321, DK-2800 Lyngby (Denmark)
tb@imm.dtu.dk, mibi@imm.dtu.dk

Abstract

In this paper, we investigate the use of event models for automated planning. Event models are the action defining structures used to define a semantics for dynamic epistemic logic. Using event models, two issues in planning can be addressed: Partial observability of the environment and knowledge. In planning, partial observability gives rise to an uncertainty about the world. For single-agent domains, this uncertainty can come from incomplete knowledge of the starting situation and from the nondeterminism of actions. In multi-agent domains, an additional uncertainty arises from the fact that other agents can act in the world, causing changes that are not instigated by the agent itself. For an agent to successfully construct and execute plans in an uncertain environment, the most widely used formalism in the literature on automated planning is "belief states": sets of different alternatives for the current state of the world. Epistemic logic is a significantly more expressive and theoretically better founded method for representing knowledge and ignorance about the world. Further, epistemic logic allows for planning according to the knowledge (and iterated knowledge) of other agents, allowing the specification of a more complex class of planning domains, than those simply concerned with simple facts about the world. We show how to model multi-agent planning problems using

Kripke-models for representing world states, and event models for representing actions. Our mechanism makes use of slight modifications to these concepts, in order to model the internal view of agents, rather than that of an external observer. We define a type of planning domain called epistemic planning domains, a generalisation of classical planning domains, and show how epistemic planning can successfully deal with partial observability, nondeterminism, knowledge and multiple agents. Finally, we show epistemic planning to be decidable in the single-agent case, but only semi-decidable in the multi-agent case.

1 Introduction

For most of its early life in the 60's and 70's, the field of automated planning was concerned with ways in which the problem of creating long-term plans for achieving goals could be formulated, such that solving problems of non-trivial size, would be computationally feasible. The types of planning that arose from this early work, is what is known today as Classical Planning. Classical Planning, as defined by Ghallab et al. (2004), imposes a number simplifying restrictions on the planning problem, namely that it be *finite*, *fully observable*, *deterministic* and *static*.

While there certainly are computational benefits to the above restrictions, it is also clear that such planning domains are much easier to construct theoretically sound planning algorithms for. In other words, the reason that Classical Planning became so dominant was not only due to limited computational resources, but also a limited understanding of the intricacies of how, for instance, to take the actions of other agents into account when planning, or how to naturally represent incomplete information about the world state – the complexity of automated planning is not solely computational.

In this paper, we examine a new method of planning, with which the full observability and determinism requirement can be lifted. Getting partial observability comes from the use of epistemic Kripke-models to represent knowledge about the world, recognising that partial observability and knowledge (or a lack thereof) are two sides of the same coin. Event models, taken from dynamic epistemic logic, are used in defining the ways in which actions change epistemic models, whether they are factual – changing propositional facts about the world – or epistemic – changing knowledge of the facts, but not the facts

themselves – or a combination thereof. In addition, event models provide a natural way to handle nondeterminism. Epistemic planning, as we name this new approach, will be considered in both single- and multi-agent versions.

Consider the similarities between belief states, the most widely used method in the literature on automated planning for dealing with the incomplete knowledge that arises from partial observability, and Kripke-models for epistemic logic. Belief states are sets of propositions about the world, each of which represents an alternative version of the world. In epistemic modal logic, each world also represents an alternative, but with the addition of a notion of indistinguishability of these alternatives by particular agents. Even without going into details about models of epistemic logic, it is immediately obvious that epistemic logic is at least as expressive as belief states when it comes to planning, and, as the reader will learn, they are actually much more so. With the combination of epistemic logic and event models, we gain the ability to plan in nondeterministic, partially observable, multi-agent domains with knowledge, where belief states only affords us the ability to deal with the first two. Further, epistemic planning, as we call this new paradigm, internalises nondeterminism and observability in the planning language, rather than dealing with it at an algorithmic level. We find this to be a much more satisfying approach.

The remainder of this paper is organised as follows. Section 2 introduces the the well known notions of epistemic models from the literature on modal logic, and shows how they, with rather elegant modifications, can be used to model the internal view of an agent involved in the situation being modeled. Section 3 introduces our version of event models, which are largely similar to those of dynamic epistemic logic, with minor modifications to facilitate the internal view. In section 4 we show definitions of classical planning problems, epistemic planning domains and their correspondences. With epistemic planning defined, section 5 examines properties of different types of actions based on event models, and establishes a nomenclature for these. Finally, sections 6 and 7 deals with the single- and multi-agent versions of epistemic planning domains respectively, and gives decidability results for both.

2 Epistemic logic and epistemic states

In this section we present the notions from (dynamic) epistemic logic required for the remaining article. First of all, we define a language of epistemic logic.

Let P be a finite set of atomic propositions (propositional symbols), and $\mathcal{A}$ a finite set of agents. We will most often use symbols $p, q, r, s, \ldots$ for atomic propositions and $i, j, k, l, \ldots$ for agents. The language $\mathcal{L}_K(P, Ag)$, the language of multi-agent epistemic logic on $(P, \mathcal{A})$, is generated by the following BNF:

$$\phi ::= \top \mid \bot \mid p \mid \neg\phi \mid \phi \wedge \phi \mid K_i\phi,$$

where $p \in P$ and $i \in \mathcal{A}$. As usual, the intended interpretation of a formula $K_i\phi$ is "agent i knows ϕ". We also consider an extended language $\mathcal{L}_{KC}(P, Ag)$ obtained by adding formulas of the type $C\phi$ intended to express common knowledge of ϕ. The semantics of $\mathcal{L}_K(P, Ag)$ and $\mathcal{L}_{KC}(P, Ag)$ is defined as usual through Kripke structures, here called *epistemic models*.

Definition 2.1 (Epistemic models)**.** An *epistemic model* of the languages $\mathcal{L}_K(P, Ag)$ and $\mathcal{L}_{KC}(P, Ag)$ is a triple $\mathsf{M} = (W, R, V)$, where

- W is the *domain*, a finite set of *worlds* (often called *states* in the literature, but we will use the word "state" for a different purpose in this paper).
- $R : \mathcal{A} \to 2^{W \times W}$ assigns an *accessibility relation* (or *indistinguishability relation*) R_i to each agent $i \in \mathcal{A}$. All accessibility relations are equivalence relations.
- $V : P \to 2^W$ assigns a set of worlds to each atomic proposition; this is the *valuation* of that variable.

The domain W of an epistemic model $\mathsf{M} = (W, R, V)$ is often denoted $D(\mathsf{M})$. The requirement of the accessibility relations being equivalence relations ensures that the modal operators K_i capture knowledge. Most of what we do in this paper would work equally well with weaker or no conditions on the accessibility relations, e.g."belief" or even weaker notions, but for simplicity we stick to knowledge in this paper.

Definition 2.2 (Truth in an epistemic model)**.** Let an epistemic model $\mathsf{M} =$

(W, R, V) of $\mathcal{L}_{KC}(P, Ag)$ be given. Let $i \in \mathcal{A}$, $w \in W$ and $\phi, \psi \in \mathcal{L}_{KC}(P, Ag)$.

$\mathsf{M}, w \models \top$		always
$\mathsf{M}, w \models \bot$		never
$\mathsf{M}, w \models p$	iff	$w \in V(p)$
$\mathsf{M}, w \models \neg\phi$	iff	$\mathsf{M}, w \not\models \phi$
$\mathsf{M}, w \models \phi \wedge \psi$	iff	$\mathsf{M}, w \models \phi$ and $\mathsf{M}, w \models \psi$
$\mathsf{M}, w \models K_i\phi$	iff	for all $v \in W$, if wR_iv then $\mathsf{M}, v \models \phi$
$\mathsf{M}, w \models C\phi$	iff	for all $v \in W$, if $w(\cup_{j\in\mathcal{A}} R_j)^* v$ then $\mathsf{M}, v \models \phi$

where R^* is the transitive closure of R.

If $\mathsf{M}, v \models \phi$ holds for all epistemic models $\mathsf{M} = (W, R, V)$ and all $w \in W$, the formula ϕ is said to be *valid*, denoted $\models \phi$.

A pair (M, w) consisting of an epistemic model M and a world $w \in D(\mathsf{M})$ is often called an *epistemic state* (or *pointed epistemic model*). In an epistemic state (M, w), w denotes the *actual world*. Epistemic states provide a model of the world from an external point of view, where the modeler is assumed to be an omniscient and external observer of the epistemic situation Aucher (2010). Thus, the modeler knows which is the actual world. In this paper, we also wish to be able to represent an internal point of view, where the modeler is one of the agents represented in the epistemic model. For this purpose, we distinguish between *global (epistemic) states* representing the external view of the world and pointing out the actual state of affairs, and *local (epistemic) states* representing individual agents' view of the world. This is related to the distinction made in e.g. Aucher (2010), Fagin et al. (1995).

Definition 2.3 (Local and global (epistemic) states)**.** A pair (M, W_d) consisting of an epistemic model $\mathsf{M} = (W, R, V)$ of $\mathcal{L}_{KC}(P, Ag)$ and a non-empty set of *designated* worlds $W_d \subseteq W$ is called an *epistemic state* or simply a *state* (of $\mathcal{L}_{KC}(P, Ag)$). If W_d is a singleton, the state is called *global*. If W_d is closed under R_i, where $i \in \mathcal{A}$, it is called a *local state* for agent i. In general, a *local state* is any pair (M, W_d) which is the local state of some agent. Given a global state $(\mathsf{M}, \{w\})$, the *associated local state* of agent i is $(\mathsf{M}, \{v \mid wR_iv\})$. States (M, W_d) in which the domain of M is a singleton are called *atomic states*.

Note that in a local state (M, W_d) for agent i, it is possible to have a pair of nodes $w, v \in W_d$ with $(w, v) \notin R_i$. We will later use this to model "plan-time indistinguishability" whereas if $(w, v) \in R_i$ it models "run-time indistinguishability".

Consider a local state $s = (\mathsf{M}, W_d)$ containing only w and v, and that these are plan-time indistinguishable. This means, that while the agent is planning, it does not know whether the actual world is w or v. However, when the plan is being executed and the agent actually achieves the state of the world represented by s, it will know which of w and v is the actual world. In other words, the agent knows, that while it does not yet know the actual world, it will come to know this once the plan is being carried out. If w and v are run-time indistinguishable, then the agent is unable to distinguish them, both while planning and when carrying out the plan. These concepts will be elaborated on later, particularly in Section 6.

An alternative way to define (local) states would be to introduce an additional accessibility relation R_d and an additional world w_0 s.t. $w_0 R_d w$ iff w belongs to the set of designated worlds. In this way (local) states would become ordinary pointed models of the form (M, w_0). However, we stick to the definition above, as it makes some of the following definitions and constructions simpler. The only disadvantage is that one has to be a bit more careful in defining bisimulations on states. Here is the definition.

Definition 2.4 (Bisimulations between (epistemic) states)**.** A *bisimulation* between states $((W, R, V), W_d)$ and $((W', R', V'), W'_d)$ is a non-empty binary relation $B \subseteq W \times W'$ which is an ordinary bisimulation between (W, R, V) and (W', R', V') and which furthermore satisfies that the domain of B extends W_d and the image of W_d under B is W'_d.

Note than when W_d and W'_d are singletons, this definition reduces to the ordinary definition of a bisimulation between pointed models. We can then, as usual, define the *bisimulation contraction* of a state as the quotient structure of the union of all autobisimulations (see e.g. Blackburn et al. (2001) for details).

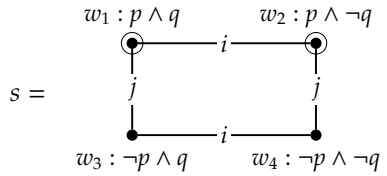
Definition 2.5 (Truth in an (epistemic) state)**.** Let (M, W_d) be an epistemic state of $\mathcal{L}_{KC}(P, Ag)$ and ϕ a formula of $\mathcal{L}_{KC}(P, Ag)$. Then truth of ϕ in (M, W_d) is defined as follows:

$$(\mathsf{M}, W_d) \models \phi \quad \text{iff} \quad \mathsf{M}, w \models \phi \text{ for all } w \in W_d$$

Note that for all local states (M, W_d) of some agent i and all formulas ϕ, the following holds: $(\mathsf{M}, W_d) \models K_i\phi \Leftrightarrow (\mathsf{M}, W_d) \models \phi$, as W_d is closed under R_i, and since R_i is reflexive. The property reflects the fact that a local state of some agent gives that agent's internal view of the world. When $(\mathsf{M}, W_d) \models \phi$

– or, equivalently, $(\mathsf{M}, W_d) \models K_i\phi$ – we say that agent *i knows* ϕ in (M, W_d). If $(\mathsf{M}, W_d) \models K_i\phi \vee K_i\neg\phi$, we say that agent *i knows whether* ϕ holds.

Example 1. Consider the following state:



The reflexive loops at each of the worlds have been left out for visual simplicity, which will be the case in the remainder of this paper. More generally, we will always only the show the *reflexive transitive reduction* of a state, that is, the one in which each accessibility relation R_i has been replaced by the minimal relation R_i' having the same reflexive transitive closure as R_i. The symbol $\odot$ marks designated worlds. Here, the designated worlds are w_1 and w_2. As the set of designated worlds is closed under R_i but not R_j, the state is a local state of agent i but not agent j, thus enabling its interpretation as i's view of the world. That both w_1 and w_2 are designated is due to agent i's inability to recognise which of these is the actual world. In the state, agent i knows that p holds, but doesn't know whether q holds. i does, however, know that j knows whether q holds.

3 Event models and epistemic actions

From Dynamic Epistemic Logic (DEL), we take the concept *event model* (or *update model* or *action model*), see e.g. van Ditmarsch and Kooi (2008), for modeling the changes to epistemic states, brought about by the execution of actions. The exact relationship between event models and epistemic states in relation to planning will be clarified later.

Definition 3.1 (Event models)**.** An *event model* for $\mathcal{L}_{KC}(P, Ag)$ is a quadruple $\mathcal{E} = (E, Q, pre, post)$, where

- E, the *domain*, is a finite non-empty set of *events*.
- $Q : \mathcal{A} \to 2^{E \times E}$ assigns an *accessibility relation* (or *indistinguishability relation*) to each agent $i \in \mathcal{A}$. All accessibility relations are equivalence relations.

- $pre : E \to \mathcal{L}_{KC}(P, Ag)$ assigns to each event a *precondition*.
- $post : E \to \mathcal{L}_{KC}(P, Ag)$ assigns to each event a *postcondition*. Postconditions are conjunctions of propositional literals, that is, conjunctions of atomic propositions and their negations (including $\top$ and $\bot$).

The domain E of an event model $\mathcal{E} = (E, Q, pre, post)$ is denoted $D(\mathcal{E})$. The postcondition mapping is defined in a slightly non-standard way here. Usually, it is defined as a mapping $post' : E \to (P \to \mathcal{L}_{KC}(P, Ag))$. As shown in van Ditmarsch and Kooi (2008), one can without loss of generality restrict to mappings of this type where $post'(e)(p)$ is always either $\top$, $\bot$ or p itself. Any such mapping gives rise to a mapping $post : E \to \mathcal{L}_{KC}(P, Ag)$ of the type defined above by letting:

$$post(e) = \left(\bigwedge_{post'(e)(p)=\top} p \right) \wedge \left(\bigwedge_{post'(e)(p)=\bot} \neg p \right).$$

One of the advantages of this formulation of the postcondition mapping is that it links more naturally with classical planning, as will be seen further below.

Definition 3.2 (Local and global (epistemic) actions)**.** A pair $(\mathcal{E}, E_d)$ consisting of an event model $\mathcal{E} = (E, Q, pre, post)$ of $\mathcal{L}_{KC}(P, Ag)$ and a non-empty set of *designated* events $E_d \subseteq E$ is called an *epistemic action* or simply an *action* (of $\mathcal{L}_{KC}(P, Ag)$). If E_d is a singleton, the action is called *global*. If E_d is closed under R_i, where $i \in \mathcal{A}$, it is called a *local action* for agent i. In general, a *local action* is any pair $(\mathcal{E}, E_d)$ which is the local state of some agent. Given a global action $(\mathcal{E}, \{e\})$, the *associated local action* of agent i is $(\mathcal{E}, \{f \mid eQ_i f\}$. Actions $(\mathcal{E}, E_d)$ in which the domain of $\mathcal{E}$ is a singleton are called *atomic actions*.

The literature sometimes refer to our global actions as *pointed updates* and our non-global actions as *multi-pointed updates*, see e.g. Sadzik (2006). Beware that even though we sometimes refer to our actions as *epistemic* actions, they also allow the possibility of expressing factual (ontic) change via the postcondition mapping.

Definition 3.3 (Product update of a state with an action)**.** Given are a state (M, W_d) and an action $(\mathcal{E}, E_d)$, where $\mathsf{M} = (W, R, V)$ and $\mathcal{E} = (E, Q, pre, post)$. The *product update* of the state (M, W_d) with the action $(\mathcal{E}, E_d)$ is defined as the state $(\mathsf{M}, W_d) \otimes (\mathcal{E}, E_d) = ((W', R', V'), W'_d)$, where

- $W' = \{(w, e) \in W \times E \mid \mathsf{M}, w \models pre(e)\}$
- $R'_i = \{((w, e), (v, f)) \in W' \times W' \mid wR_iv \text{ and } eQ_if\}$
- $V'(p) = (\{(w, e) \in W' \mid \mathsf{M}, w \models p\} - \{(w, e) \in W' \mid post(e) \models \neg p\}) \cup \{(w, e) \in W' \mid post(e) \models p\}$
- $W'_d = \{(w, e) \in W' \mid w \in W_d \text{ and } e \in E_d\}$

Definition 3.4 (Applicability of an action in a state)**.** Given are a state (M, W_d) and an action $(\mathcal{E}, E_d)$. The action $(\mathcal{E}, E_d)$ is said to be *applicable* in the local state (M, W_d) if the following holds: For each world $w \in W_d$ there is a least one event $e \in E_d$ such that $\mathsf{M}, w \models pre(e)$.

The intuition behind this definition is the following. First, note that if both the state and the action are global, that is of the form $W_d = \{w\}$ and $E_d = \{e\}$, this reduces to the condition that the precondition of the designated event e holds in the designated world w ($\mathsf{M}, w \models pre(e)$). This corresponds to the condition of "possibility" introduced with a similar purpose in Löwe et al. (2010). The point is, that the designated world w denotes the *current world* and the designated event e denotes the event that *actually* takes place, so the condition simply ensures that the precondition of the event that takes place is satisfied in the current world. If not, the pointed set W'_d of the product update $(\mathsf{M}, \{w\}) \otimes (\mathcal{E}, \{e\})$ would be empty.

Now consider the case of local states and actions. When we update a local state (M, W_d) of agent i with a local action $(\mathcal{E}, E_d)$ of the same agent, the local action is assumed to present agent i's view on what the action will bring about (it could be an action executed by i himself, but could also be an action executed by some other agent, or a joint action of several agents). The condition of Definition 3.4 then has the following meaning: For each of the worlds that agent i considers possible, the action specifies at least one applicable event that i considers possible.

Lemma 1. *If a local action $(\mathcal{E}, E_d)$ of an agent i is applicable in a local state (M, W_d) of the same agent, then the product update $(\mathsf{M}, W_d) \otimes (\mathcal{E}, E_d)$ is again a local state of i.*

We leave the proof as an (easy) exercise for the reader.

Example 2. The following example is inspired by the *Sally-Ann test* used in cognitive psychology to test whether children possess a so-called *theory of mind*

Wimmer and Perner (1983). There are three agents, Sally (denoted by i), Ann (denoted by j) and an observer, the child (denoted by k). Sally has a basket and Ann has a box. There is a marble, which can either be in the basket or in the box. We use b to denote the proposition "the marble is in the basket". In the initial situation, the marble is in the basket, and this is common knowledge. Thus the following local state s_0 describes all three agents' initial view of the world:

$$s_0 = \quad \circledcirc\, w_1 : b$$

Now Sally (i) leaves the room, and in the meantime Ann (j) moves the marble to the box. The observer sees this, but Sally doesn't. However, it is common knowledge that when Sally leaves the room, Ann has the possibility of moving the marble to the box. The observer can represent the action taking place by the following local action a_1:

$$a_1 = \quad \bullet \;\; e_1 : \langle b, \top \rangle \;\; \overset{i}{\text{———}} \;\; e_2 : \langle b, \neg b \rangle \;\; \circledcirc$$

Labeling events by the pair $\langle \phi_1, \phi_2 \rangle$ means that the event has precondition ϕ_1 and postcondition ϕ_2. In the action a_1, event e_1 represents the possibility that Ann doesn't move the marble, and event e_2 represents the possibility that she does. Since Sally (i) has left the room, she cannot distinguish these two events, which is represented by the two events being connected by an i-relation. There is, however, no j- or k-relation connecting the two, since both Ann and the observer can distinguish between the marble being moved and not being moved. The designated event is e_2, since the observer sees the marble being moved. Taking the product update of s_0 with a_1, we then obtain the observers updated view of the world after the action has taken place:

$$s_0 \otimes a_1 = \quad \bullet \;\; (w_1, e_1) : b \;\; \overset{i}{\text{———}} \;\; (w_1, e_2) : \neg b \;\; \circledcirc$$

We can see that the observer now knows that Sally (i) doesn't know where the marble is (doesn't know whether b or $\neg b$ holds). It has been shown in Wimmer and Perner (1983) that children under the age of 4, and autistic children in general, will—when playing the role of the observer—conclude that Sally knows that the marble is now in the box ($\neg b$).

With epistemic actions and states defined, we now show how these are employed in a planning context.

4 Epistemic planning domains and problems

Following Ghallab et al. (2004), any classical planning domain can be represented as a *restricted state-transition system* $\Sigma = (S, A, \gamma)$, where

- S is a finite or recursively enumerable set of *states*.
- A is a finite set of *actions*.
- $\gamma : S \times A \hookrightarrow S$ is a computable *state-transition function*. The state-transition function is partial, that is, for any $(s, a) \in S \times A$, either $\gamma(s, a)$ is undefined or $\gamma(s, a) \in S$.

A *classical planning problem* is then represented as a triple (Σ, s_0, S_g), where

- Σ is a *restricted state-transition system*.
- s_0 is the *initial state*, a member of S.
- S_g is the set of *goal states*, a subset of S.

A *solution* to a classical planning problem (Σ, s_0, S_g) is a finite sequence of actions (a *plan*) $a_1, a_2, \ldots, a_n$ such that

$$\gamma(\gamma(\ldots\gamma(\gamma(s_0, a_1), a_2), \ldots, a_{n-1}), a_n) \in S_g.$$

Note that finding solutions to classical planning problems is always at least semi-decidable: given a planning problem, we can compute its *state space* (the space of states reachable by a sequence of actions applied to the initial state) in a breadth-first manner, and if one of the goal states is reachable, we will eventually find it. Next is the definition of epistemic planning domains, which are special cases of classical planning domains.

Definition 4.1 (Epistemic planning domains)**.** Given are a finite set P of atomic propositions and a finite set $\mathcal{A}$ of agents. An *epistemic planning domain* on $(P, \mathcal{A})$ is a restricted state-transition system $\Sigma = (S, A, \gamma)$, where

- S is a finite or recursively enumerable set of epistemic states of $\mathcal{L}_{KC}(P, Ag)$.
- A is a finite set of actions of $\mathcal{L}_{KC}(P, Ag)$.

- γ is defined by:

$$\gamma(s,a) = \begin{cases} s \otimes a & \text{if } a \text{ is applicable in } s \\ \text{undefined} & \text{otherwise} \end{cases}$$

If all states and actions are from $\mathcal{L}_K(P, Ag)$ it is called an epistemic planning domain *without common knowledge*. If $|\mathcal{A}| = 1$ it is called a *single-agent* epistemic planning domain.

Definition 4.2 (Epistemic planning problems)**.** An *epistemic planning problem* is a triple (Σ, s_0, ϕ_g), where

- $\Sigma = (S, A, \gamma)$ is an epistemic planning domain on $(P, \mathcal{A})$.
- s_0, the *initial state*, is a member of S.
- ϕ_g is a formula in $\mathcal{L}_{KC}(P, Ag)$ called a *goal formula*. The set of *goal states* is $S_g = \{s \in S \mid s \models \phi_g\}$.

If all states, actions, and formulas are from $\mathcal{L}_K(P, Ag)$ it is called an epistemic planning problem *without common knowledge*. If $|\mathcal{A}| = 1$ it is called a *single-agent* epistemic planning problem.

Epistemic planning problems are special cases of classical planning problems. A *solution* to an epistemic planning is thus, according to the definition above, a sequence of actions $a_1, a_2, \ldots, a_n$ s.t. $\gamma(\gamma(\ldots\gamma(\gamma(s_0, a_1), a_2), \ldots, a_{n-1}), a_n) \in S_g$, that is, s.t. $s_0 \otimes a_1 \otimes a_2 \otimes \cdots \otimes a_n \models \phi_g$. As noted, finding solutions is at least semi-decidable. A further look at the complexity of epistemic planning is found in later in the paper. Before examining (in the next section) some of the different types of actions that can be defined in epistemic planning problems, we briefly touch upon the relation between epistemic planning and Dynamic Epistemic Logic (DEL).

Note that in our framework for epistemic planning, our only take away from DEL is event models and product updates. We do not make use of the full DEL language, that is, epistemic logic extended with action modalities. Action modalities are used in DEL to express the logical consequences of performing actions encoded as event models. This means that we have a logical language in which it is possible to represent and reason about actions and their dynamics. In classical planning, on the other hand, the underlying logical language only

describes static states of affairs. The dynamics is instead captured in a meta-language. This meta-language describes actions in terms of how they modify state descriptions given as formulas of the object language (cf. e.g. STRIPS). In other words, in classical planning the object language describing states is completely separate from the meta-language describing actions. We have here taken a similar approach, where the object language for describing states is simply standard epistemic logic, and the meta-language for describing actions is event models.

It would of course also be possible in our framework to include action modalities in the object language, that is, make it the full language of DEL. This would allow us to include formulas with action modalities in pre- and post-conditions of actions as well as in goal formulas. It would thus allow us to, for instance, express goals such as "achieve a state in which it is (im)possible for agent j to perform an action that will result in ϕ." If we were to allow goals to include statements about actions and their consequences, we would also like to be able to state goals such as: "achieve a state in which it is (im)possible for agent j to perform *any sequence of actions* that will result in ϕ." This is not possible in standard DEL, but requires us to introduce iteration of modalities. We leave this for future work.

5 Action types

An action $(\mathcal{E}, E_d)$ is called *purely epistemic* if for all events e in $D(\mathcal{E})$, the post-condition of e is implied by the precondition, that is, $\models pre(e) \rightarrow post(e)$. A purely epistemic action is one that does not make any factual (ontic) changes. An important example of such actions are *public announcements*, which are the purely epistemic atomic actions. Actions that are not purely epistemic are called *ontic*. Along an orthogonal axis, we can distinguish between observable and non-observable actions. An action $(\mathcal{E}, E_d)$ is called *fully observable* or *public*, if all the accessibility relations of $\mathcal{E}$ are identities (that is, no two distinct events are connected). If an action is not fully observable, it is called *partially observable*.

Let there be given a partially observable action $((E, Q, pre, post), E_d)$ and a group of agents $G \subseteq \mathcal{A}$. If for each $i \in G$, the accessibility relation Q_i is the identity, then the action is said to be *group observable* by G. If, in addition, for each $j \in \mathcal{A} - G$ the accessibility relation Q_j is the universal relation, then it is said to be *group observable by G alone*. An action is said to be *privately observable* by an

agent i, if the action is group observable by $\{i\}$ alone. Note, that agents $j \in \mathcal{A} - G$ will know that *something* has happened, though not precisely what.

An action $((E, Q, pre, post), E_d)$ is called *globally deterministic* if all preconditions are mutually inconsistent, that is, $\models pre(e) \wedge pre(f) \rightarrow \bot$ for all distinct $e, f \in E$, in other words, only one event is possible for each world. It is called a *sensing action* if:

- it is purely epistemic
- it is globally deterministic
- its preconditions cover the logical space, that is, $\models \bigvee_{e \in E} pre(e) \leftrightarrow \top$.

Sensing actions are called *answers* in Gerbrandy (2007), but the word "sensing" is better in line with the taxonomy of the automated planning literature.

Example 3. Consider again the Sally-Ann example (Example 2), but now from the perspective of Sally. From her perspective, the event that takes place while she is out of the room is represented by the following local action:

$$a'_1 = \quad e_1 : \langle b, \top \rangle \quad \overset{i}{\text{---}} \quad e_2 : \langle b, \neg b \rangle$$

This action is *group observable by* $\{j, k\}$ *alone* (Ann and the observer). The product update of s_0 with a'_1 gives Sally's view on the world after the event has taken place:

$$s_0 \otimes a'_1 = \quad (w_1, e_1) : b \quad \overset{i}{\text{---}} \quad (w_1, e_2) : \neg b$$

Sally now no longer knows whether the marble is in the basket or not, that is, $s_0 \otimes a'_1 \models \neg K_i b \wedge \neg K_i \neg b$. Sally might after this consider the action of entering the room again and look into the basket, where the marble used to be. This is a *sensing action*, where Sally will get to know whether the marble is in the basket or not. The sensing action looks as follows, again from the viewpoint of Sally:

$$a'_2 = \quad e_1 : \langle b, \top \rangle \qquad e_2 : \langle \neg b, \top \rangle$$

Note that both events are still designated. This is because Sally does not know the outcome of the sensing action at the time she plans the action (see a more

thorough discussion of this in the following section). Now, updating Sally's local state $s_0 \otimes a'_1$ with this sensing action, we get:

$$s_0 \otimes a'_1 \otimes a'_2 = \quad \underset{\circledbullet}{(w_1, e_1, e_1) : b} \qquad \underset{\circledbullet}{(w_1, e_2, e_2) : \neg b}$$

Now there is no longer an i-edge between the two worlds, because it represents the situation after Sally has been sensing which of the two holds. We now have that Sally knows whether the marble is in the basket, that is, $s_0 \otimes a'_1 \otimes a'_2 \models K_i b \vee K_i \neg b$.

In the next section, we will look at how epistemic planning domains generalise some well-known types of planning domains studied in automated planning.

6 Propositional planning and partial observability in single-agent domains

Following Ghallab et al. (2004), a *propositional planning domain* (or *set-theoretic planning domain*) on a finite set P of atomic propositions is a restricted state-transition system $\Sigma = (S, A, \gamma)$ satisfying:

- $S = 2^P$.
- A is a set of pairs $a = (\mathit{precond}(a), \mathit{effects}(a))$, where both $\mathit{precond}(a)$ and $\mathit{effects}(a)$ are finite sets of literals over P. An action a is said to be *applicable* in a state if $\mathit{precond}^+(a) \subseteq s$ and $\mathit{precond}^-(a) \cap s = \emptyset$.[1]
- γ is defined by:

$$\gamma(s, a) = \begin{cases} (s - \mathit{effects}^-(a)) \cup \mathit{effects}^+(a) & \text{if } a \text{ is applicable in } s \\ \text{undefined} & \text{otherwise} \end{cases}$$

Note that propositional planning is decidable, as the set of states is finite. Every propositional planning domain $\Sigma = (S, A, \gamma)$ is equivalent to an epistemic planning domain $\Sigma' = (S', A', \gamma')$ defined as follows:

[1] For any set of literals L, L^+ denotes the set of atoms in L and L^- denotes the set of atoms whose negations are in L.

- S' is the set of atomic states of $\mathcal{L}_{KC}(P, Ag)$ where $|\mathcal{A}| = 1$.

- A' is the set that for each $a \in A$ contains an atomic action $(E_a, Q_a, pre_a, post_a)$ given by $E_a = \{e\}$, $Q_a = \{(e, e)\}$, $pre = \bigwedge_{p \in precond^+(a)} p \ \wedge \bigwedge_{p \in precond^-(a)} \neg p$ and $post = \bigwedge_{p \in effects^+(a)} p \ \wedge \bigwedge_{p \in effects^-(a)} \neg p$.

- γ' is defined as above for general epistemic planning domains.

It is easy to check that Σ and Σ' are indeed equivalent. This shows that propositional planning domains are a special case of epistemic planning domains, and that the propositional planning domains can be precisely characterised as those epistemic planning domains where all states and actions are atomic, and where all actions have purely propositional preconditions. This should come as no surprise, but is still worth noting, as it clarifies the exact link between classical propositional planning and epistemic planning.

Epistemic planning domains also allow for a nice treatment of *partial observability*. Assume we are still in a single-agent domain, that is $|\mathcal{A}| = 1$. Let i denote the element of $\mathcal{A}$. Assume we have a local action $((E, Q, pre, post), E_d)$ of agent i. Let $e, e' \in E_d$. We say that e and e' are *runtime indistinguishable* if eQ_ie', otherwise they are called *runtime distinguishable* or *plan-time indistinguishable*. The point is this. Assume, for example, that the agent is facing a closed box which might either be full (denoted by f) or empty (denoted by $\neg f$), but he doesn't know which. Let c denote the proposition "the box is closed". Then his local state is this:

$$s_0 = \begin{array}{c} w_1 : f \wedge c \\ | \\ i \\ | \\ w_2 : \neg f \wedge c \end{array}$$

Now assume he considers the action of opening the box to see its content. At plan-time (when he is still computing the plan), he only knows that the effect of opening the box will be that he *either* learns f *or* learns $\neg f$, but not which. So the two outcomes are plan-time indistinguishable to him. However, at run-time when actually carrying out the action, he *will* know which of the two is the case. We can model this by the following local action:

$$openBox = \begin{array}{cc} e_1 : \langle f, \neg c \rangle & e_2 : \langle \neg f, \neg c \rangle \end{array}$$

Updating the state above with this action we then get:

$$s_0 \otimes \mathit{openBox} = \begin{array}{c} (w_1, e_1) : f \wedge \neg c \\ \circledcirc \end{array} \qquad \begin{array}{c} (w_2, e_2) : \neg f \wedge \neg c \\ \circledcirc \end{array}$$

The state after the execution only differs from the state before by the substitution of $\neg c$ for c (opening the box) and the removal of the edge between the f-world and the $\neg f$-world. This means that after the action the agent will be able to distinguish between f and $\neg f$. However, as the agent at plan time still doesn't know which it will be, we need to keep both worlds in the set of distinguished worlds. This explains the need of the set of distinguished worlds, W_d, and the need of a special definition of bisimulation between states.

Constructing actions that combine runtime indistinguishable events with plan-time indistinguishable events allows us to model partial observability: if two possible outcomes (events) will be indistinguishable even when the action is performed at runtime (no observation), then they should be in the same Q_i equivalence class; if the two possible outcomes will be distinguishable when the action is performed (observable), then they should be in distinct Q_i equivalence classes. And, obviously, if the exact outcome is known already at plan-time, the action will contain only a single event representing this outcome. Note that this approach to partial observability is consistent with the definition of *fully observable actions* introduced earlier. According to this definition, a single-agent action is fully observable if and only if its accessibility relation is the identity, that is, all pairs of events are runtime distinguishable. The articles Bacchus and Petrick (1998), Petrick and Bacchus (2002) argue in favour of a similar approach to partial observability.

As it can easily be seen, in the single-agent case, each local state can—modulo bisimulation—be uniquely described by a set of atomic actions and a description of which of the actions are runtime distinguishable. Thus it seems fair to say that epistemic planning with one agent captures *exactly* what is involved in propositional planning in (nondeterministic) domains with partial observability.

Example 4. Continuing the example with the agent and the box, consider the following action:

$$\mathit{emptyBox} = \begin{array}{c} e_1 : \langle f \wedge \neg c, \neg f \rangle \\ \circledcirc \end{array} \qquad \begin{array}{c} e_2 : \langle \neg f, \top \rangle \\ \circledcirc \end{array}$$

This is an action for emptying the box. Note that it distinguishes between two cases: one covering the case where the box is full and open, and another covering the case where it's already empty. Note that the action is only applicable when the agent knows that either the box is open or already empty. We now get:

$$s_0 \otimes openBox \otimes emptyBox = \begin{array}{c} (w_1, e_1, e_1) : \neg f \wedge \neg c \\ \circledcirc \\ \end{array} \quad \begin{array}{c} \\ \\ (w_2, e_2, e_2) : \neg f \wedge \neg c \\ \circledcirc \end{array}$$

Thus a solution to the planning problem of satisfying the goal formula $\neg f$ given the initial state s_0 would be *openBox*, *EmptyBox*. Note that the branching that usually takes place when planning in partially observable domains like this is being internalised in the state descriptions. Note also that if we take the bisimulation contraction of the state $s_0 \otimes openBox \otimes emptyBox$, we get an atomic state. Thus if we want to plan further, e.g. close the box again, we can now work with atomic states.

We will now prove that single-agent epistemic planning is decidable, that is, given any epistemic planning problem we can decide whether a plan exists or not. In the proof we actually show something slightly stronger, since we also show how to construct a plan if one exists.

Theorem 1. *Single-agent epistemic planning is decidable.*

Proof. Given any single-agent epistemic planning problem, we can perform a breadth-first exploration of the state space. However, after computing each new state, we make sure to replace it by its bisimulation contraction, which can be computed in linear time Dovier et al. (2001). Now it suffices to prove that when $|\mathcal{A}| = 1$, there are only finitely many distinct bisimulation minimal states of $\mathcal{L}_{KC}(P, Ag)$ (recall that P is always assumed to be finite). Consider first connected states of $\mathcal{L}_{KC}(P, Ag)$, that is, states with only one equivalence class. Since the accessibility relation is an equivalence relation, there can be no two worlds satisfying the same atomic propositions in a bisimulation minimal state (the two worlds would be bisimilar). Thus all bisimulation minimal connected states are substates of the following state (up to isomorphism):[2]

$$((2^P, 2^P \times 2^P, V), 2^P), \text{where } V(p) = \{w \mid p \in w\}.$$

[2] A state $((W', R', V'), W'_d)$ is called a *substate* of a state $((W, R, V), W_d)$ if (W', R', V') is a submodel of (W, R, V) and $W'_d = W_d \cap W'$.

There can obviously only be finitely many such substates (up to isomorphism). Now consider the case of non-connected states. Note that we can not immediately reduce these to connected states due to the way we defined bisimulations on states. In any case, each equivalence class in the state must again be a substate of the state defined above. Furthermore, there can be no two bisimilar equivalence classes, by bisimulation minimality. Thus, there can also only be finitely many bisimulation minimal non-connected states (up to isomorphism). This is the required conclusion. □

In this section we have only been considering the single-agent case, but obviously the multi-agent case is the most interesting, and, as we will see next, also far more challenging.

7 Multi-agent epistemic planning

We will start this section by giving an example of a multi-agent epistemic planning domain inspired by the well-known *Byzantine Agreement* problem (or *coordinated attack* problem) Fagin et al. (1995).

Example 5. There are three logicians, a philosopher (i), a computer scientist (j), and a mathematician (k). They work at the same university, so usually they go together in the same car, allowing them to discuss logic on the way. One day at work it happens that agent i suddenly recalls that he forgot to turn off the lights of the car, so that the battery is now flat. Let l denote the proposition "the lights are on" and let b denote the proposition "the battery is flat". A possible local state describing agent i's internal view of the world immediately after having realised that the light were left on could then be:

$$s_0 = \begin{array}{l} w_1 : l \wedge b \\ j,k \\ w_2 : \neg l \wedge \neg b \\ j,k \\ w_3 : \neg l \wedge b \end{array}$$

Recall that, by convention, we only show the reflexive transitive reduction of the state, so there is an implicit j,k-edge from w_1 to w_3. In this local state, it is assumed to be common knowledge that j and k are still unaware of whether the lights are on or not. The fact that there is no world labelled $l \wedge \neg b$ means

that it is also common knowledge that *if* the lights are on, *then* the battery is flat, that is, $l \to b$.

Assume further, to keep the example manageable (and because leaving the lights on happens more often than the three logicians would wish to think about), that it is common knowledge that from now on agent i can repeatedly only choose between the following three actions:

1. tell agent j that l:
2. go to the car and turn the lights off if they are still on;
3. tell agent k that l.

Actions 1 and 3 are both announcement of l, so they can both be expressed by the event $\bullet\ \langle l, \top \rangle$. Action 2 is an ontic action that can be expressed by the event $\bullet\ \langle \top, \neg l \rangle$. Agent k cannot distinguish 1 from 2 and agent j cannot distinguish 2 from 3. So we obtain an event model $\mathcal{E}$ looking like this:

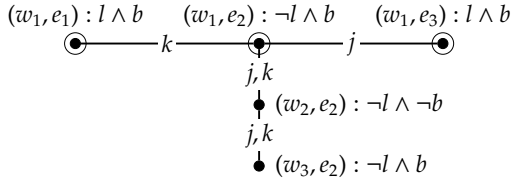
$$\mathcal{E} = \quad e_1 : \langle l, \top \rangle \bullet \overset{k}{\text{———}} \bullet e_2 : \langle \top, \neg l \rangle \bullet \overset{j}{\text{———}} \bullet e_3 : \langle l, \top \rangle$$

The three local actions available to agent i are then $(\mathcal{E}, \{e_1\})$, $(\mathcal{E}, \{e_2\})$ and $(\mathcal{E}, \{e_3\})$ (note that agent i has full observability). The three possible actions are modelled by the same event model, only differing in the designated set. This is often the case in multi-agent epistemic planning domains, and it means that the branching that usually takes place in the search for a plan is here internalised in a single epistemic action, and the branching factor of the plan search becomes 1 (if we make sure to label worlds with their update history to be able to keep track of the designated sets). Of course, what you pay for this is that the epistemic models might grow exponentially in size as you move down through the state space, but this is essentially no different from classical planning, where the number of reachable states can grow exponentially with the depth.

Now consider the product update of the initial state s_0 with the local action $a = (\mathcal{E}, \{e_1, e_2, e_3\})$, where we have chosen all three events as designated to postpone the decision of which to pick (corresponds to nondeterministic, but

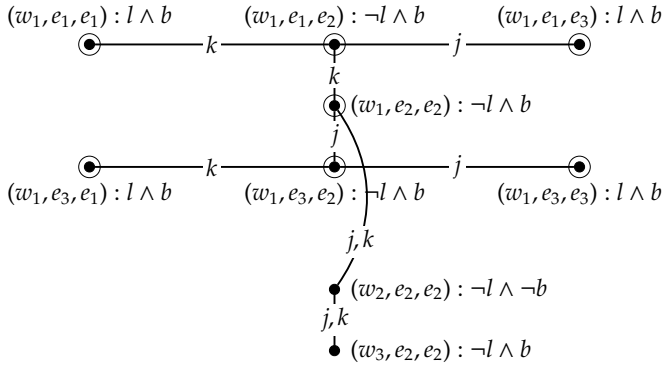
observable, choice, cf. Section 5):

$s_0 \otimes a =$



It can be seen that if agent i e.g. chooses to do e_1 (tell l to j), then afterward $K_j b$ and $\neg K_k b$ (since these formulas hold in the world (w_1, e_1)). Now consider a second update with the action a:

$s \otimes a \otimes a =$



If agent i first chooses e_1 and then e_3, the designated world (actual world) will become (w_1, e_1, e_3). We here have $K_k K_j b$ but not $K_j K_k K_j b$. It might at first seem intuitively puzzling that k knows that j knows the battery to be flat after i having performed only the action sequence e_1, e_3 (tell l to j, tell l to k). To explain the intuition, first note that by choice of event model, whenever agent i performs an action, all three agents will "know" that one of e_1, e_2 or e_3 has happened, but not necessarily which. Thus after i having performed first e_1 and then e_3, we can intuitively think of agent k as being able to perform the following line of reasoning: "Agent i's second action was to tell me l. Thus the lights must still be on. Therefore agent i's first action can not have been to turn off the lights. Since his first action wasn't to let me know about the lights either, his first action must have been to let agent j know that the lights are on. From this, j must have been able to conclude that the battery is flat." This reasoning leads k to conclude that j knows b (no sequence of actions can change the truth value of b). This provides the informal intuition behind why $K_k K_j b$ holds at (w_1, e_1, e_3).

It can easily be shown that in general in the world

$$(w_1, \overbrace{e_1, e_3, e_1, e_3, \ldots, e_1, e_3}^{2(n+1)})$$

of the state $s_0 \otimes a^{2(n+1)}$ we have $(K_k K_j)^{n+1} b$ but not $K_j (K_k K_j)^{n+1} b$. It's a bit like an inverse Muddy Children puzzle: instead of iteratively decreasing the depth of the agents' uncertainty, it is iteratively increased. This is similar to the situation obtained in the Byzantine Agreement problem. From this we can infer that there is no upper bound on the size of the models in the chain $s_0 \otimes a, s_0 \otimes a^2, s_0 \otimes a^3, \ldots$, not even if we take the bisimulation contractions of the states. Similar to Byzantine Agreement, it is also easy to infer that no matter which sequence of choices agent i makes, it will never become common knowledge that the battery is flat (there will always be exactly one world in which $\neg l \wedge \neg b$ holds, and this world is accessible from all other worlds by some path).

The example we just gave made use of ontic actions. However, in Sadzik (2006) it is shown that even allowing only *purely epistemic* actions with propositional preconditions, we can still get iterated updates of arbitrary size (using a variant of the coordinated attack problem). From the fact that in multi-agent epistemic planning problems there is in general no upper bound on the size of the reachable epistemic states, one might fear that planning is not even decidable in the general case. Indeed, this is exactly the case, as we will now show.

Theorem 2. *Multi-agent epistemic planning is undecidable (even without common knowledge).*

Proof. Undecidability here means that there is no decision procedure that for arbitrary multi-agent epistemic planning problems can determine whether a solution exists or not. We give the proof by showing that for any Turing machine M we can construct an epistemic planning problem P_M that has a solution if and only if M halts. As the halting problem is undecidable, so is epistemic planning. The underlying idea is this. Given any Turing machine M, we can encode its configurations (state, tape content and head position) as epistemic models—models containing exactly one world per non-blank tape cell. Furthermore, we can encode the possible transitions of M as epistemic actions. In this way, we achieve that any run of M can be simulated by a sequence of epistemic actions applied to the epistemic state representing the initial configuration of M. Suppose M has only a single halting state which we

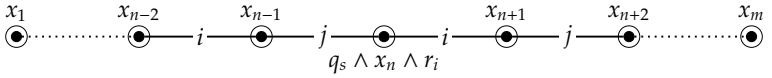
represent in the epistemic language by a special propositional symbol q_f. We can then conclude that M halts if and only if there is a sequence of epistemic actions leading from the (representation of the) initial epistemic state to an epistemic state in which q_f holds in one of the worlds. In this way, the Turing machine halts if and only if there is a solution to the planning problem in which the goal is that q_f should hold in one of the worlds. This gives us the required planning problem P_M that has a solution if and only if M halts.

We now proceed with the details. Let there be given a deterministic Turing machine M with two-way infinite tape, and states $q_0, q_1, \ldots, q_f$, where q_0 is the initial state and q_f is the (only) halting state. The set of tape symbols is some finite set Γ including a blank symbol, b. We will now show how to construct an epistemic planning problem in which the Turing machine's configurations (state, tape content and head position) are encoded as epistemic states, and the transitions of the Turing machine are encoded as epistemic actions. First we will make use of common knowledge, but later we will show how the role of common knowledge can be replaced by the introduction of an additional agent. We build the planning problem on the language $\mathcal{L}_{KC}(\{q_0, \ldots, q_f\} \cup \Gamma \cup \{r_i, r_j\}, \{i, j\})$.

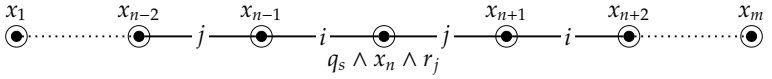
Assume we are given a configuration of M with instantaneous description (ID)

$$x_1 \cdots x_{n-2} x_{n-1} q_s x_n x_{n+1} \cdots x_m,$$

where $q_s \in \{q_0, \ldots, q_f\}$ and all $x_i \in \Gamma$ (see Hopcroft et al. (2006) for details on Turing machines and instantaneous descriptions). Then this instantaneous description will be encoded as either of the following local states:



(1)



(2)

We call these local states the states *representing* the ID. There is exactly one world to represent each of the non-blank tape cells of the Turing machine, and this world is labelled by the symbol representing the content of the cell. In addition, the world representing the current tape cell is labelled by two additional atomic propositions: the name of the current state (q_s) and either the proposition r_i or

the proposition r_j. The purpose of the propositions r_i and r_j will be explained in a moment. First note the alternation of i- and j-edges in these models. This is to ensure that the local states represent linear structures where each world has exactly one left and one right neighbour. If instead all edges were i-edges, all pairs of worlds would be each others neighbours, as all accessibility relations are assumed to be equivalence relations. Thus, if all edges were i-edges, we would be representing a set rather than a linear structure. The linear structure is required, since this is the only way we can encode the tape of a Turing machine. This also indicates why the current proof wouldn't work in the single-agent case, at least as long as we insist on using only equivalence classes, that is, insist on representing *knowledge*.

Now back to the propositions r_i and r_j. The purpose of r_i and r_j is to mark which indistinguishability relation (either i or j) will lead to the tape cell to the right of the current one (the "next" tape cell). If r_i holds at the world representing the cell currently scanned, it means that the tape cell to the right is represented by the neighbouring world reached by following the i-edge—and vice versa for r_j. Since in 1, r_i holds at the world representing the current tape cell, it means that the world representing the tape cell to the right is the one labelled x_{n+1}. If we replaced r_i by r_j in 1, it would correspond to changing the direction of the tape, and the right neighbour would instead become x_{n-1}.

The initial configuration of M (empty tape) will be represented by the singleton local state

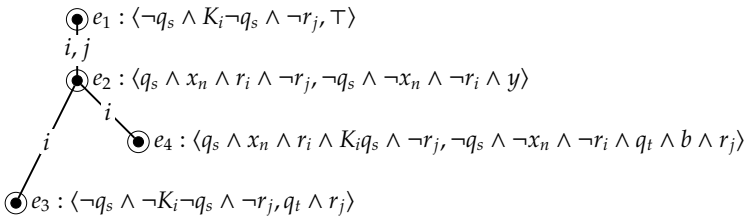
$$s_0 = \textcircled{\bullet} q_0 \wedge b \wedge r_i$$

This will be the initial state of our planning problem P_M. In the planning problem, we put two (symmetric) local actions for each of the transitions of the Turing machine. We will only show the local actions for transitions of the form

$$\delta(q_s, x_n) = (q_t, y, R), \text{where } x_n \neq y, \tag{3}$$

as transitions of the form $\delta(q_s, x_n) = (q_t, y, L)$ and transitions with $x_n = y$ can be handled similarly. There are two local actions corresponding to (3), where one is obtained from the other by interchanging i with j everywhere. We thus only

show one of the two:



(4)

We call these local actions the actions *representing* the transitions of the Turing machine. Suppose M can perform a move from an instantaneous description ID_1 to an instantaneous description ID_2, and let a be the local action representing the transition used in the move. The point is now that if s is a local state representing ID_1, then the product update $s \otimes a$ will be representing ID_2. Before providing the details, we will try to explain the intuition behind the construction of the local actions, and how they can simulate the moves of the Turing machine.

Let a denote an action of the form shown above, representing a transition of type (3). Let s denote a local state representing an instantaneous description of M, that is, s is on the form (1) or (2). Suppose a is applicable in the local state s. Then, by the applicability condition (Definition 3.4), s can only be of the form (1), as none of the events of a have preconditions that satisfy r_j. Now consider what happens when a is applied to s, that is, when we form the product update $s \otimes a$. We denote the world of s in which q_s holds by w_c. The world w_c represents the current tape cell of the Turing machine (before the update). We now consider how the events $e_1, \ldots, e_4$ of a affect the product update.

Event e_1 has its precondition satisfied in all worlds of s except w_c (because of the conjunct $\neg q_s$) and its right neighbour (because of the conjunct $K_i \neg q_s$). Since e_1 has an empty postcondition, this implies that in $s \otimes a$ all worlds of s except w_c and its right neighbour will be kept unchanged (paired with e_1). In other words, the tape excluding the current cell and its right neighbour remain the same after the update, as it should.

Event e_2 has its precondition satisfied in w_c and this world only. Its postcondition deletes q_s, x, and r_i, and instead adds y. Thus, event e_2 makes sure to change the symbol at the current tape cell from x to y, and remove the head from this cell.

Event e_3 has its precondition satisfied in the right neighbour of w_c, if such a right neighbour exists in s. In case it exists, q_t and r_j will be added as conjuncts

to it. Thus e_3 makes sure to place the head at the right neighbour of the previous current cell, and to update the state from q_s to q_t.

In case the right neighbour of w_c doesn't exist, the action a makes sure to construct such a right neighbour. This is done via the event e_4. In case w_c has no right neighbour, the precondition of e_4 will be satisified in w_c. This implies that the product update $s \otimes a$ will contain both a world (w_c, e_2) and a world (w_c, e_4). The first of these is the "updated version" of w_c, whereas (w_c, e_4) is a "new" world. This new world is accessible from (w_c, e_2) by an i-edge. It is the new right neighbour of w_c. The postcondition of e_4 makes sure that $q_t \wedge b \wedge r_j$ will hold in this new right neighbour. Thus e_4 makes sure to construct a new right neighbour cell (if needed), make this the new current cell, put a blank symbol into it, and update the state.

Note that given any state s, either it will contain a world satisfying $\mathsf{pre}(e_3)$ or a world satisfying $\mathsf{pre}(e_4)$, but not both. There will be a world satisfying $\mathsf{pre}(e_3)$ if the head stays within the previously used part of the tape when the transition represented by a is executed, otherwise there will be a world satisfying $\mathsf{pre}(e_4)$.

We can now finalise the proof. The set of local actions of the planning problem P_M is taken to be the set of local actions representing the transitions of M. Now suppose there is a move of the Turing machine from an instantaneous description ID_1 to an instantaneous description ID_2. We then need to prove that if s_1 is a local state representing ID_1, then for all actions a applicable in s_1, the local state $s_1 \otimes a$ represents ID_2. We need to split the proof into cases, distinguishing the cases where the tape head moves into the previously unused part of the tape, and those where it doesn't. We will only cover one of the cases here, as they are largely similar. Let us consider the most tricky case, where the tape gets extended. So assume ID_1 is an instantaneous description of the following form:

$$\mathrm{ID}_1 = x_1 \cdots x_{n-2} x_{n-1} q_s x_n,$$

and the move performed is a result of the following transition:

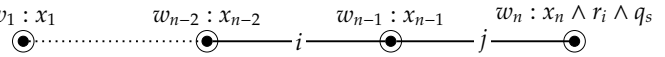
$$\delta(q_s, x_n) = (q_t, y, R), \text{where } x_n \neq y. \tag{5}$$

The move will then result in the following instantaneous description:

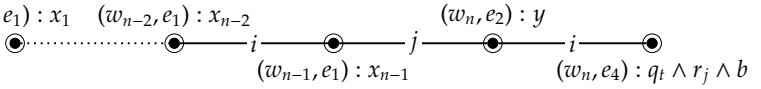
$$\mathrm{ID}_2 = x_1 \cdots x_{n-2} x_{n-1} y q_t b.$$

There are two local states that can represent ID_1, but they are symmetric, so we can without loss of generality assume that the local state s_1 representing ID_1 is

the following:

$s_1 =$ 

Now, the only actions applicable in this state are actions containing at least one event with its preconditions being satisfied by $x_n \wedge r_i \wedge q_s$. This implies that the only action applicable is the one representing the transition (5) (recall that the Turing machine is deterministic). This is the action shown in (4). Taking the product update of s_1 with the action (4), we get the following local state s_2:

$s_2 =$ 

The reader is encouraged to check that this is indeed the correct product update of s_1 with the action (4). Action s_2 is immediately seen to be a representation of the instantaneous description ID_2, as required.

It now follows that if $a_1, a_2, \ldots$ is any sequence of actions in P_M where each a_i is applicable in $s_0 \otimes a_1 \otimes \cdots \otimes a_{i-1}$, then the sequence $s_0, s_0 \otimes a_1, s_0 \otimes a_1 \otimes a_2, \ldots$ will be a representation of the sequence of the moves of the Turing machine M. Now choose the goal formula of P_M to be $\neg C \neg q_f$. This formula expresses that there is world accessible by some path at which q_f holds. It holds exactly in those epistemic states representing halting states of the Turing machine. Thus we now have that the planning problem has a solution if and only if M halts. This is the required result.

The proof just given makes use of common knowledge. We can do away with common knowledge by introducing a third agent, k, instead. The idea is quite simple: whenever there is an i- or j-edge in any of the states or actions introduced above, we also add a k-edge. Since all accessibility relations are equivalence relations, this means that if a world of a state is accessible by *any* path, then it is accessible by a single k-edge. Thus we can replace the goal formula $\neg C \neg q_f$ by the formula $\neg K_k \neg q_f$. The rest of the proof remains unchanged. Whether epistemic planning with only two agents and no common knowledge is decidable or not is an open problem. □

There seem to be no direct equivalents of this result in the existing literature, although there are obviously some connections to the non-stabilisation results of

iterated updates over various types of purely epistemic actions in Sadzik (2006), and to the undecidability of the logic of iterated public announcements in Miller and Moss (2005). The result above is of course not encouraging for epistemic planning in the general case, however, semi-decidability can sometimes be sufficient for a planner, as a planner embedded in an agent architecture (e.g. a BDI agent) would usually in any case rely on being timed out if a plan is not found within a reasonable time. From a positive perspective, this result shows that we have been introducing a very expressive planning framework, more expressive than previous frameworks suggested for planning based on epistemic logic van der Hoek and Wooldridge (2002), Petrick and Bacchus (2002) (since these other frameworks are known to be decidable, and ours is only semi-decidable in the most general case). In any case, an interesting problem of course becomes to find fragments of epistemic planning that *are* decidable. We already saw one such fragment, the single-agent case. Another decidable fragment is the one that only allows globally deterministic actions (cf. Section 5). This covers e.g. public announcements, atomic ontic actions, and sensing actions. That this fragment is decidable follows trivially from the fact that updates with actions having mutually inconsistent preconditions can never increase the model size. In Löwe et al. (2010), it is shown that interesting planning problems can be expressed even within these restrictived fragments. A more thorough investigation of which fragments of multi-agent epistemic planning are decidable is left for future work.

8 Related and future work

Work on using Dynamic Epistemic Logic in planning was recently independently initiated by Löwe, Pacuit and Witzel Löwe et al. (2010). Their work however differs from ours in a number of ways. They only consider purely epistemic actions (no postconditions), but on the other hand they allow arbitrary accessibility relations in models. Both accounts can surely be extended to cover both ontic actions and arbitrary accessibility relations with a bit of extra work. In Löwe et al. (2010), a restricted planning fragment is shown to be decidable by giving an upper bound on model size, but general algorithms and decidability issues are not covered. Our work also differs in allowing the *internal perspective* on planning, where the epistemic models represent the planning agent's internal view of the world. We have shown that this gives a nice and natural way of dealing with partial observability in planning (even relevant in

the single-agent case). The closest relative to our idea of epistemic models from an internal perspective appears to be the recent work by Aucher Aucher (2010), however, his approach is technically slightly different.

Another line of research considers *planning as model checking*. The idea is here to represent the state space of the planning problem as a model in a suitable temporal logic, and then recast the planning problem as a model checking problem (model checking of a formula expressing reachability of the goal state). The article van der Hoek and Wooldridge (2002) considers epistemic planning from this perspective. It however assumes that the state space is already given, and that it is finite. Thus, it doesn't consider the problem of how to express actions in a convenient formalism, and it doesn't allow the expressiveness we have in our formalism. In this approach, the treatment of epistemic and ontic change is similar—either way it is just a next step in a run, and how the valuation between different point changes is not essential to define or describe the transition van Ditmarsch and Kooi (2008). Note also that *tractable* in the article van der Hoek and Wooldridge (2002) refers to the fact that the model checking algorithm is polynomial in the *size of the model*, that is, in the size of the entire state space. Usually the complexity of planning problems is stated as functions of the size of the descriptions of the available actions, and from this perspective even classical propositional planning is PSPACE-complete. Of further interest along this axis of inquiry is Hans van Ditmarsch and Ruan (2011) where formulas in Dynamic Epistemic Logic are model checked by an explicit construction of the corresponding interpreted system.

Currently, the authors are involved in work on a logic of branching plans, clearly defining what it means for a plan to be a solution to a particular problem, by way of reduction axioms to DEL. This allows plans validation in standard DEL terms. Parallel to this work, a beginning algorithm for epistemic planning is being developed, making concrete the process of plan synthesis. We are also looking into using plausibility models (like those of Baltag and Smets (2007)), rather than purely epistemic models for planning. This would allow a planning agent to focus first on finding a plan that works for those outcomes which are found most plausible, rather than having to take all possibilities into account. Particularly in multi-agent scenarios, this may vital if planning is to be feasible.

Further work includes generalising the framework to arbitrary accessibility relations, in particular for representing belief rather than knowledge. We would also like to carry through a more thorough investigation of the decidable fragments of epistemic planning. Finally, we wish to develop languages suitable

for describing local actions in a form which is more manageable and closer to the traditional planning languages like STRIPS. Such languages should not necessarily possess the full expressivity of local actions, but should be tailored for expressing a non-trivial subset relevant for actual planning problems in multi-agent domains (e.g. with different predefined action types for dealing with sensing, announcements, and ontic change, and parameters for specifying the observers of the action).

Acknowledgements The authors would like to thank the following persons for valuable comments and ideas for improvement of the original draft: Torben Braüner, Valentin Goranko, Jens Ulrik Hansen, Martin Holm Jensen, and two anonymous reviewers. The work is partially funded by the Danish Natural Science Research Council under project HYLOCORE.

References

G. Aucher. An internal version of epistemic logic. *Studia Logica*, 94(1):1–22, 2010.

F. Bacchus and R. P. A. Petrick. Modeling an agent's incomplete knowledge during planning and during execution. In *KR*, pages 432–443, 1998.

A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. In G. Bonanno, W. van der Hoek, and M. Wooldridge, editors, *Texts in Logic and Games*, volume 3, pages 11–58. Amsterdam University Press, 2007.

P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*, volume 53 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, Cambridge, UK, 2001.

A. Dovier, C. Piazza, and A. Policriti. A fast bisimulation algorithm. In G. Berry, H. Comon, and A. Finkel, editors, *CAV*, volume 2102 of *Lecture Notes in Computer Science*, pages 79–90. Springer, 2001. ISBN 3-540-42345-1.

R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning About Knowledge*. MIT Press, 1995.

J. Gerbrandy. Communication strategies in games. *Journal of Applied Non-Classical Logics*, 17(2), 2007.

M. Ghallab, D. S. Nau, and P. Traverso. *Automated Planning: Theory and Practice*. Morgan Kaufmann, 2004.

W. v. d. H. Hans van Ditmarsch and J. Ruan. Connecting dynamic epistemic and temporal epistemic logics. *Logic Journal of the IGPL, to appear.*, 2011.

J. E. Hopcroft, R. Motwani, and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation (3rd Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2006. ISBN 0321455363.

B. Löwe, E. Pacuit, and A. Witzel. Planning based on dynamic epistemic logic. technical report, ILLC, University of Amsterdam, May 2010.

J. S. Miller and L. S. Moss. The undecidability of iterated modal relativization. *Studia Logica*, 79(3):373–407, 2005.

R. P. A. Petrick and F. Bacchus. A knowledge-based approach to planning with incomplete information and sensing. In M. Ghallab, J. Hertzberg, and P. Traverso, editors, *Proceedings of the Sixth International Conference on Artificial Intelligence Planning and Scheduling (AIPS-2002)*, pages 212–221, Menlo Park, CA, Apr. 2002. AAAI Press.

T. Sadzik. Exploring the iterated update universe. technical report, ILLC, University of Amsterdam, 2006.

W. van der Hoek and M. Wooldridge. Tractable multiagent planning for epistemic goals. In *AAMAS '02: Proceedings of the first international joint conference on Autonomous agents and multiagent systems*, pages 1167–1174, New York, NY, USA, 2002. ACM. ISBN 1-58113-480-0. doi: http://doi.acm.org/10.1145/545056.545095.

H. van Ditmarsch and B. Kooi. Semantic results for ontic and epistemic change. In G. Bonanno, W. van der Hoek, and M. Wooldridge, editors, *Logic and the Foundation of Game and Decision Theory (LOFT 7)*, Texts in Logic and Games 3, pages 87–117. Amsterdam University Press, 2008.

H. Wimmer and J. Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13:103–128, 1983.

Dynamic Epistemic Analysis for IERS

Jianying Cui and Xudong Luo

The Institute of Logic and Cognition, Sun Yat-Sen University, China
cuijiany@mail.sysu.edu.cn, david.x.d.luo@gmail.com

Abstract

The iterated regret minimization solution exhibits good qualitative behavior, as observed in experiments in many games that have proved problematic for Nash Equilibrium (NE) solutions. So, it is interesting to explore epistemic characterizations unearthing players' rationality for an algorithm of Iterated Eliminations Regret-dominated Strategy (*IERS*) related to the solution. In this paper, firstly, based on dynamic epistemic logic (Public Announcement Logic and Plausible Belief Revision Logic), we develop two epistemic regret-game models, and define a new rationality. Then, we characterize the Iterated Elimination Regret-Dominated procedure as a process of dynamic information exchange by taking the players' rationality, respectively, as a proper announcement assertion and a radical upgrade proposition. Thereby, we provide a new perspective on the outcomes of the *IERS* algorithm.

1 Introduction

Deviations between behavioral rationality and Bayesian rationality in some games will appear if we set the players' rationality as a precondition of game analysis: Bayesian rationality may predict outcomes (Nash Equilibria)

that are inconsistent with empirical observations in these games (cf.McKelvey (1992),Basu (2007)). One of the main issues in the study of rational behavior of players in game theory has been to explore the causes of such deviations and to provide reasonable epistemic foundations for the algorithms related to the deviations. This study has recently spread to other subject fields (cf. Bernheim (1984),David (1984)).

In Renou and Schlag (2011) and Halpern and Pass (2012), a new iterated algorithm has been proposed named Iterated Eliminations Regret-dominated Strategy (*IERS*). It is one way to capture the intuition that a player wants to do well no matter what the other players do. With the algorithm, firstly, one player needs to figure out the maximal regret value of each of his strategies according to some rules, given that the players are uncertain about their opponents' strategies(i.e., players regard any strategy of their opponents is possible). Then strategies are chosen corresponding to a minimum regret value after comparing these maximum regret values, and a new subgame is formed with the strategies chosen at the previous step, repeating the process in the new subgame until the subgame no longer changes. Halpern etc. in Halpern and Pass (2012) take the strategy profiles of the final subgame obtained as new Game Solutions(GS), Iterated Regret Minimization (IRM). They proved that the new solutions exhibit the same behavior as that observed in experiments in real life for many famous games that have proved problematic for NE, including the Traveler's Dilemma, the Centipede Game, and Nash Bargaining. The game solution and its algorithm, *IERS*, are particularly appealing when considering inexperienced but intelligent players that play a one-shot game for the first time. For example, in the Traveler's Dilemma, given a penalty is 2, minimax regret equilibrium is precisely $(97, 97)$, and agrees well with the experimental results in Becker et al. (2005).

Accordingly, it becomes interesting to explore the epistemic characterization unearthing players' rationality for the algorithm *IERS*. Halpern etc. in Halpern and Pass (2012) provided an epistemic characterization for *IERS* based on an epistemic logic. However, in their characterization, an epistemic paradox (cf.Halpern and Pass (2012)) arise so that they had to assign successively lower probabilities to higher orders of rationality, and to weaken a basic premise in game theory(i.e., "Rationality is common knowledge among players") by insisting that the higher levels of belief regarding other players' rationality do not involve common knowledge or common belief. However, rationality of common knowledge as a basic premise is recorded in almost all of game textbooks, and supported by many game experts and researchers (cf. Au-

mann (1997)Rubinstein (1994)). Therefore, inspired by van Benthem (2011),van Benthem (2007) and Bonanno (2008), we construct two regret epistemic game models for different dynamic epistemic analyses of the *IERS* algorithm. Based on the epistemic models, we describe an iterated elimination regret-dominated procedure as a process of dynamic information exchange by defining players' rationality, respectively, as a proper announcement assertion and a radical upgrade proposition. Then we show that the two different interactive epistemic results among players are both line with the outcomes of *IERS*. We thus provide a new characterization of the algorithm *IERS* in a simpler and more intuitive way. The characterization avoids the paradox in the *IERS* algorithm and retains a classic rule in game theory, namely, that rationality should be common knowledge among players. Moreover, we can construct a uniform frame to analyze and explore rationality in iterated algorithms from players' regret perspective, such as restating the concepts of Weak Rationality (*WR*) and *SR* (Strong Rationality) in van Benthem (2007) based on our regret-epistemic game model. We thus offer a new perspective to explore logic characterizations for algorithms of Iterated Elimination Strictly Dominated strategies (*IESD*) and algorithms of Rationalizability corresponding to Bernheims version.

The rest of this paper is organized as follows. The preliminaries about game theory and dynamic epistemic logic are contained in Section 2. Section 3 - 5 are the heart of the paper: we construct two regret epistemic game models based on the given games, after providing semantic interpretations for those propositions related to a game, especially, giving the semantic definition of the proposition "a player is rational" at a world in these models, we study the game solutions in the context of the rational players with a soft information upgrade and hard information update, and prove a consistency between outcomes of the algorithm *IERS* and the results from players' public announcement and dynamic upgrading with the rationality. In Section 6, we discuss related approach, and conclude in Section 7.

2 Preliminaries

Recently researchers systematically the iterated minimax regret algorithm and its solution (cf.Renou and Schlag (2011), Renou and Schlag (2010), Halpern and Pass (2012)and Stoye (2011) etc.). Originally, the idea of minimax regret was developed in decision theory by Savage(cf. L.J.Savage (1951)). Basically his

approach is to minimize the worst-case regret. The aim of this is to perform as closely as possible to the optimal course. Since the minimax criterion applied here is to the regret rather than to the payoff itself, it is not as pessimistic as the ordinary minimax approach. In finite pure strategies context, we choose a simple version to keep the general proposal as simple as possible, and make the dynamic epistemic logic analysis for the algorithm itself to be the key feature. [1]

Definition 2.1. Let $G = \langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$ be a strategic form game. A regret game of G is a quintuple $G' = \langle N, \{S_i\}_{i\in N}, \{re_i\}_{i\in N}\rangle$ where $\{re_i\}_{i\in N}$ stands for player i's ex-post regret associated with any profile of pure actions (s_i, s_{-i}) as $re_i(s_i, s_{-i}) = max\{u_i(s'_i, s_{-i}), s'_i \in S_i\} - u_i(s_i, s_{-i})$, and let $re_i(s_i) = max\{re_i(s_i, s_{-i}), \forall s_{-i} \in S_i)\}$ states the regret value of choosing s_i for player i.

Definition 2.2. Given the game $G = \langle N, \{S_i\}_{i\in N}, \{re_i\}_{i\in N}\rangle$, let s_i and s'_i be available strategies for player i, and set $S'_{-i} \subseteq S_{-i}$, then s_i is regret-dominated by s'_i on S'_{-i} if $re_i(s'_i) < re_i(s_i)$. And for set $S' \subseteq S$, strategy $s'_i \in S_i$ is unregrettable with respect to S'_i, if no strategy in S'_i regret-dominates s'_i on S'_{-i}. In additional, a regret-dominated strategy s_i is also called regrettable for a player i.

Definition 2.3. The procedure of Iterated Eliminations Regret-dominated Strategies (*IERS*) is as follows. Given a regret game $G' = \langle N, \{S_i\}_{i\in N}, \{re_i\}_{i\in N}\rangle$, let *IUD* respectively be the se t of iterated regret-undominated strategies of the G' recursively defined as follows:
$IUD = \prod_{i\in N} IUD_i$, where $IUD_i = \bigcap_{m\geqslant 0} IUD_i^m$, with $IUD_i^0 = S_i$ and $RD_i^0 = \{s_i \in IUD_i^0 \mid s_i$ is regrettable with respect to IUD_i^0 in $G'\}$. For $m \geqslant 1, IUD_i^m = IUD_i^{m-1} \setminus RD_i^{m-1}$, where, $RD_i^m = \{s_i \in IUD_i^m \mid s_i\}$ is regrettable with respect to IUD_i^m in a subgame G'^m.[2].

It is assumed that at each stage all dominated strategies are simultaneously deleted in Definition 2.3. In contrast to most equilibrium concepts, *IERS* yields a rectangular set of strategy profiles, i.e., a Cartesian product of sets. This *IERS* procedure is illustrated in Figure 1.

$$IUD_1^0 = \{X, Y, Z\}, RD_1^0 = \{X\}, IUD_2^0 = \{a, b, c\}, RD_2^0 = \{b\};$$

$$IUD_1^1 = \{Y, Z\}, RD_1^1 = \varnothing, IUD_2^1 = \{a, c\}, RD_2^1 = \{c\};$$

[1] Most of our conclusions in this paper can be extended to mixed strategy context. However, it is beyond the scope of this paper. We will explore the problem in our future works

[2] G'^m is a subgame of G', in which $S_i = IUD_i^m$ and $G'^0 = G'$

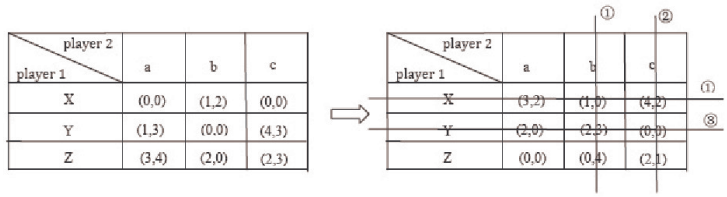


Figure 1: *IERS* procedure

$IUD_1^2 = \{Y, Z\}, RD_1^2 = \{Y\}, IUD_2^2 = \{a\} = IUD_2, RD_2^2 = \varnothing;$

$IUD_1^3 = \{Z\} = IUD_1.$
Thus, $IUD = \{(Z, a)\}$.

IUD is not consistent with NE in many games, but as we have known that traditional game-theoretic solution concepts (most notably NE) predict outcomes that are inconsistent with empirical observations, that is the main reason why researchers introduce the algorithms of minimax regret into the game theory.

3 Public Announcement Logic and Plausible Belief Revision Logic

Exploring update or upgrade scenarios for scenarios of virtual communication in games, in a dynamic epistemic logic for changing game models, has developed over the past decade (cf. Baltag et al. (1999), Ditmarsch et al. (2007), van Benthem (1996), van Benthem (2011) etc.). As a basis for most dynamic epistemic logics, Public Announcement Logic (PAL) can deal with the change of information arising from the action of public announcement by adding a dynamic modality $[\varphi]$ to standard epistemic logics[3], where $[\varphi]\psi$ means *after a truthful public announcement of φ, formula ψ holds*. Its truth condition is that:

$M, w \vDash [\varphi]\psi$ iff $M, w \vDash \varphi$ implies $M|_{\varphi}, w \vDash \psi$.[4]

[3]We assume the reader is familiar with Epistemic Logics.(for example, the concept of common knowledge and the prosperities of various epistemic logic systems(S5,KD45 etc.). The reader can consult the recent books and papers (cf. van Benthem (2011)Halpern (2003)) for details.

[4]$M|_{\varphi}$ is a submodel of M in which φ is true.

With this language, we can say things like $[!\varphi]K_i\psi$: *after a truthful public announcement of* φ*, agent i knows* ψ, or $[\varphi]C_N\varphi$: *after its announcement,* φ *has become common knowledge in the group N of agents* and so on.

In van Benthem (2007), van Benthem stated the issue of "an announcement limit" has close connections with the equilibria solved by algorithms of those iterated elimination dominated strategies, he showed that for any model M we can keep announcing φ, retaining just those worlds where φ holds, This yields a sequence of nested decreasing sets, which must stop in finite models, i.e., $\sharp(\varphi, M)$:

Definition 3.1. For any model M and formula φ, the announcement limit $\sharp(\varphi, M)$ is the first submodel in the repeated announcement sequence where announcing φ has no further effect.[5].

Public announcements or observations $[\varphi]$ of true propositions φ yield "hard information" that changes the current model irrevocably, discarding worlds that fail to satisfy $[\varphi]$. This is a "hard" information attitude that changes irrevocably what we know. Alternatively, we can consider an agent aware of being subject to continuous belief changes, and taking incoming signals in a softer manner, without throwing away options forever. We call this update 'soft' information attitude. To describe this, we can use worlds with plausibility orderings ($\preceq_i$) supporting dynamic updates.

Here, we start with the simplest model of beliefs: a set of states where each world is associated with the single plausibility orderings, ($w \preceq_i v$) that says "player i considers world v at least as plausible as w, we set $Min_{\preceq_i}(X) = \{v \in W \mid w \preceq_i v \text{ for all } w \in X\}$ and static models for this setting are easily defined[6]:

Definition 3.2. A simple plausibility models are structures $M = (W, \{Ra_i^{re}\}_{i\in N}, \{\preceq_i\}_{i\in N}, V)$, where $(W, \{Ra_i^{re}\}_{i\in N}, V)$ is an epistemic model, and for each $i \in N$, $\preceq_i$ is a well-founded reflexive and transitive relation on W satisfying, for all $w, v \in W$:

plausibility implies possibility: if $w \preceq_i v$ then $v \in Ra_i^{re}(w)$.

locally-connected: if $v \in Ra_i^{re}(w)$, then either $w \preceq_i v$ or $v \preceq_i w$

We can define a basic soft informational attitude:

[5]This definition is in van Benthem (2007)

[6]cf.van Benthem (2004)

Belief: $M, w \models B_i\varphi$ iff for all $v \in Min_{\preceq_i}([w]_i), M, v \models \varphi$, where $[w]_i$ is the equivalence class of w under Ra_i^{re}.

Models like this representing have been extensively used by logicians (cf. Baltag and Smets (2009),van Benthem (2004; 2011)), game theorists (cf. Board (2004)), and computer scientist (cf. P.Lanarre and Y.Shoham (1994)C.Boutilier (1992)), to represent rational agents' (all-out) beliefs.

As stated above, public announcement assumes that agents treat the source of the incoming information as infallible. But in many scenarios, agents trust the source of the information up to a point. We require some 'soft' announcements of a formula φ–we needn't eliminate worlds, but rather modify the plausibility ordering that represents an agent's current hard and soft information states. The goal is to rearrange all states in such a way that φ is believed, and perhaps other desiderata are met. Logicians have studied the policies of soft upgrade van Benthem (2004), but here, we only focus on a radical policy for belief upgrade.

Let $\|\varphi\|_i^w = \{x \mid M, x \models \varphi\} \cap [w]_i$ denote the set of φ worlds:

Definition 3.3. [7] Given an epistemic-doxastic model $M = (W, \{Ra_i^{re}\}_{i\in N}, \{\preceq_i \}_{i\in N}, V)$, and a formula φ, the radical upgrade of M with φ is the model $M^{\Uparrow\varphi} = (W^{\Uparrow\varphi}, \{Ra_i^{re}\}_{i\in N}^{\Uparrow\varphi}, \{\preceq_i\}^{\Uparrow\varphi} \preceq_{i\in N}, V^{\Uparrow\varphi})$ with $W^{\Uparrow\varphi} = W$,
for each $i, \{Ra_i^{re}\}^{\Uparrow\varphi} = \{Ra_i^{re}\}, V^{\Uparrow\varphi} = V$ and finally, for all $i \in N, w \in W^{\Uparrow\varphi}$:

for all $x \in \|\varphi\|_i^w$ and $y \in \|\neg\varphi\|_i^w$, set $x \prec_i^{\Uparrow\varphi} y$,

for all $x, y \in \|\varphi\|_i^w$, set$x \prec_i^{\Uparrow\varphi} y$ iff $x \preceq_i y$, and

for all $x, y \in \|\neg\varphi\|_i^w$, set$x \prec_i^{\Uparrow\varphi} y$ iff $x \preceq_i y$.

Accordingly, we can use modalities $[\Uparrow \varphi]\psi$ meaning "after i's radical upgrade of φ, ψ is true. Formally, $M, w \models [\Uparrow \varphi]\psi$ iff $M^{\Uparrow\varphi}, w \models \psi$.

The "hard and soft upgrade" is respectively illustrated in the following example from Pacuit and Roy (2011) :

[7] The definition is quoted from van Benthem et al. (2011)

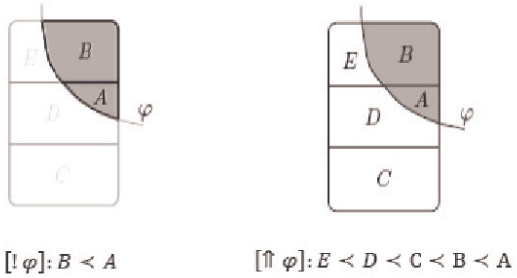


Figure 2: A "hard" and a soft "upgrade".

4 An Epistemic Game Regret Model

In order to give two dynamic epistemic analysis of the game solutions as model changes, this section we present an epistemic game regret model based on the structure of an original game model.

First, a logic is called a 'regret-game logic' ($G' - logic$) if the set of atomic propositions are in the following forms:

> Pure strategy symbols $s_i, t_i, ...$: the intended interpretation of s_i is *player i chooses strategy* s_i;
>
> Symbols Ra_i^{re}, meaning *player i is rational*, symbols Br_i^* interpreted as *the best response of player i* and a symbol GS meaning *it is a Game Solution with max-minimizing regret algorithm*;
>
> Atomic propositions of the form $s_i \succ^1 s_i'$ means the strategy s_i is at most as regrettable as the strategies s_i' for player i, or s_i regret-dominant s_i'.

Next, we define a frame for $G' - logic$ as follows:

Definition 4.1. Given a game with regret G', $\mathfrak{F}_G' = \langle W, \{\sim_i\}_{i\in N}, \{f_i\}_{i\in N}\rangle$ is a frame of $G' - logic$, where

> $W(\neq \emptyset)$: consists of all players' pure strategy profiles.
>
> $\sim_i$ is an accessibility relation for player i, which is defined as the equivalence relation of agreement of profiles in the i'th coordinate.

$f_i : W \to S_i$ is a pure strategic function, which satisfies the following property: $w \sim_i v$ iff $f_i(w) = f_i(v)$.

Simply, a frame for $G' - logic$ adds to a Kripke S5 frame (cf. Blackburn et al. (2007.)) a function that associates with every state w a strategy profile $f(w) = (f_1(w), ..., f_m(w)) \in S$. Here, the restriction $w \sim_i v$ iff $f_i(w) = f_i(v)$. It means that player i knows her own choice: if she chooses strategy s_i, then she knows that she chooses s_i. This accords with our intuition. Here, for convenience, we denote $R_i(w) = \{v \mid w \sim_i v, v \in W\}$, and $\|s_i\| = \{w \in W \mid f_i(w) = s_i\}$.

Definition 4.2. An epistemic model $M_{G'}$ over $G' - logic$ is obtained by incorporating the following valuation on a $\mathfrak{F}'_G$:

$$M_{G'}, w \vDash s_i \quad \Leftrightarrow \quad w \in \|s_i\|$$

$$M_{G'}, w \vDash (s_i \succcurlyeq^1 s'_i) \quad \Leftrightarrow \quad \exists v \in \|s'_i\|), re_i(s_i, f_{-i}(w)) \leq re_i(s'_i, f_{-i}(v))$$

$$M_{G'}, w \vDash (s_i \succ^1 s'_i) \quad \Leftrightarrow \quad \forall v \in \|s'_i\|), re_i(s_i, f_{-i}(w)) < re_i(s'_i, f_{-i}(v))$$

$$M_{G'}, w \vDash Ra_i^{re} \quad \Leftrightarrow \quad M_{G'}, w \vDash s_i \wedge (\bigwedge_{a \neq s_i}) K_i(s_i \succcurlyeq^1 a).$$

According to the above definition, our rationality has a straightforward game-theoretic meaning. That is a rational player always choose the strategies, which she knows are at least as good as her others. In details, player i is rational at a state if she can know what she chooses at the current state is not regret-dominated. That is, the rational players always try to choose an act that minimizes her regret, when she is not sure what her opponents will do. It is easy to verify that Ra_i^{re} fails exactly at the rows or the columns with which the regret-dominated strategies correspond for player i in a general epistemic regret-game model $M^*_{G'}$. For instance, in figure 1 (the left model), Ra_2^{re} fails at the states $(X, b), (Y, b)$ and (Z, b), and Ra_1^{re} fails at the states $(X, a), (X, b)$ and (X, c) of the original model $M_{G'}$.

As we mentioned previously, the dynamic analysis of iterated elimination algorithms always has to do with changing of a epistemic model. So, In the following, we call the above epistemic regret-game model $M_{G'}$ a full epistemic regret-game model, and take any submodel of a full epistemic regret-game model $M_{G'}$ as a general epistemic game model $M^*_{G'}$.

If we characterize the minimax algorithm in a static epistemic logic without any dynamic modal operator, a paradox will arise (cf. Halpern and Pass (2012)), so

that they have to use some complex methods or techniques to provide a reasonable epistemic foundation of the algorithm, such as assigning successively lower probability to higher orders of rationality, and abandoning or relaxing the most foundational rule in game theory, i.e., the common knowledge in rationality is necessary to form a game solution. However, if we analyze solution algorithms as processes of learning which change game models, we cannot only avoid these drawbacks, and since there are always dynamic intuitions concerning activities of deliberation and communication in a game, we also understand equilibriums or empirical observations in some game better. Therefore, it is more appropriate to deal with assumptions about rationality in dynamic epistemic logic.

5 Belief updating in the light of hard information

In this subsection, we will describe the procedure of *IERS* as the process of repeated announcement for the rationality assertion during players deliberate in a one-shot game, and we show "the announcement limit of the rationality assertion" always is consistent with the outcomes by solving a game with the algorithm *IERS*.

First, it is known that the assertions that players publicly announce must be the statement which they know, are true in PAL. The following theorems guarantee that the rationality notion in Definition 4.2 can be as an assertion of a public announcement.

Theorem 1. *Every finite general epistemic regret-game model has worlds with* Ra^{re} *true, where* $Ra^{re} = \bigcap_{i\in n} Ra_i^{re}$.

Proof. Since that atomic proposition Ra_i^{re} fails exactly at the rows or columns with which regret-dominated strategies correspond for player i in a general game model. Consider any general game model $M^*_{G,}$. If there are no regret-dominated strategies for all players in $M^*_{G'}$, then Ra^{re} is true at all the worlds in it. Thus, the iterated announcement of Ra^{re} cannot change the game model and get stuck in cycles in this situation. If there is a regret-dominated action for some player in the game, because of the relativity of the definition of regret-dominated strategy, he must have a strategy which is better than this strategy, i.e., if player i has a regret-dominated strategy a, then he must have a strategy, say a strategy b, which is better than strategy a. Thus, Ra_i^{re} holds at all the

worlds which belong to the row or the column corresponding to the strategy b. On the other hand, for player j, if he has no weakly dominated action, then also Ra_j^{re} holds at all the worlds. Furthermore, Ra_j^{re} holds at the worlds which belong to the row or the column corresponding to strategy b. So, Ra^{re} holds in the general game model. But if player j has also a regret-dominated action, accordingly he must have a dominant action, say action Y, and Ra_j^{re} is true at at the worlds which belong to the row or the column corresponding to the strategy Y. Therefore, Ra^{re} is satisfied at the world (Y, b).

To sum up the above arguments, every finite general game model has worlds with Ra^{re} true. □

It follows from Theorem 1 that the rationality is self-fulfilling on finite general epistemic regret-game models. Additionally, we can easily conclude the following result from the semantic interpretations Ra_i^{re} and the properties of model $M_{G'}$.

Theorem 2. *The rationality is epistemically introspective. i.e.,the formula* $Ra_i^{re} \rightarrow K_i Ra_i^{re}$ *is valid on a general epistemic regret-game model.*

Proof. Consider a general epistemic regret-game model $M^*_{G'}$ and an arbitrary w in $M^*_{G'}$ such that $M^*_{G'}, w \vDash Ra_i^{re}$ but $M^*_{G'}, w \nvDash K_i Ra_i^{re}$. Because $M^*_{G'}, w \nvDash K_i Ra_i^{re}$, $\exists v \in R_i(w)$ and $M^*_{G'}, v \nvDash Ra_i^{re}$. According to Definition 10, we have $f_i(v)$ is a regret-dominated action for i by some her actions. And, by the property of the function f_i: $f_i(w) = f_i(v)$ iff $v \in Ra_i^{re}(w)$, we can conclude that $f_i(w)$ is also a regret-dominated action for i, further, $M^*_{G'}, w \nvDash Ra_i^{re}$, contrast to the precondition, i.e., $M^*_{G'}, w \vDash Ra_i^{re}$. So, the formula $Ra_i^{re} \rightarrow K_i Ra_i^{re}$ is valid on a general game model. □

As a result, these theorems guarantee that we can successively remove the worlds at which Ra^{re} does not hold in model $M^*_{G'}$ after repeated announcing the rationality at some actual world. Although this scenario behind an iterative solution algorithm is virtual, it is significant: we can expect players to announce that they are rational, since they would know it. As van Benthem have shown in van Benthem (2007), because the current game mode could be changed in the light of the information from another player may change the current game model, it is significance to iterate the process, and repeat the assertion Ra^{re} if true.

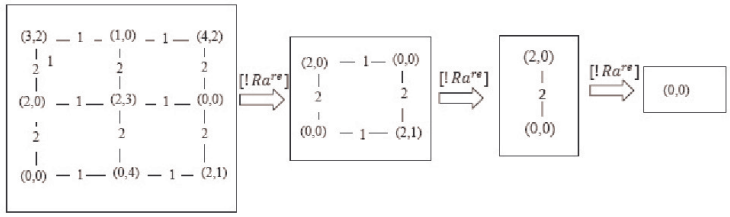


Figure 3: The public announcement of Ra^{re}

In figure 3, the left-most model is the model from fig. figure 1. The other models are obtained by public announcements of Ra^{re} successively for three times. So, in the last submodel, we have:
$M_{G'}, (Z, a) \vDash [!Ra^{re}][!Ra^{re}][!Ra^{re}]C_N$.
It indicates that if the players iteratively announce that they are rational, the process of regret-dominated strategies elimination leads them to a solution that is commonly known to be *GS*.

Theorem 3. *Given a full epistemic game model based on a finite strategic-form G' with regret and an arbitrary world w, w is in a general epistemic game model $M^*_{G'}$ which is stable by repeated announcements of Ra^{re} in the $M_{G'}$ for all players if and only if $f(w) \in IUD$. That is to say, $w \in \sharp(Ra^{re}, M_{G'}) \Leftrightarrow f(w) \in IUD$.*

Proof. (a) From left to right: if $w \in \sharp(Ra, M_{G'})$, that is to say, $w \in M^*_{G'}$, then $M^*_{G'}, w \vDash Ra$, i.e., $M^*_{G'}, w \vDash \wedge_{i \in N} Ra^{re}_i$.

First we show: $\forall i \in N, f_i(w) \notin RD^0_i$. Suppose not. Then $\exists i \in N$ such that $f_i(w) \in RD^0_i$, that is, $f_i(w)$ of player i is regret-dominated in G' by some other strategy $s'_i \in S_i = IUD^0_i$, it means: $re_i(f_i(w)) > re_i(s'_i) \Leftrightarrow max\{re_i(f_i(w), s_{-i}), \forall s_{-i} \in S_{-i}\} > max\{re_i(s'_i, s_{-i}), \forall s_{-i} \in S_{-i}\}$. Thus, let some $s'_{-i} \in S_{-i}$ satisfied $re_i(f_i(w), s'_{-i}) = max\{re_i(f_i(w), s_{-i}), \forall s_{-i} \in S_{-i}\}$, and $s''_{-i} \in S_{-i}$ satisfied $re_i(s'_i, s''_{-i} = max\{re_i(s'_i, s_{-i}), \forall s_{-i} \in S_{-i}\}$. So, we have:
$re_i(f_i(w), s'_{-i}) > re_i(s'_i, s''_{-i})$. Accordingly, there exist a $v' \in R_i(w) \cap \|s'_{-i}\|$ and a $v'' \in R_i(w) \cap \|s''_{-i}\|$, satisfied $re_i(f_i(w), f_{-i}(v')) > re_i(f_i(w), f_{-i}(v''))$, considering $re_i(f_i(w), f_{-i}(v'')) \geq re_i(f_i(w), f_{-i}(v))$, where $\forall v \in \|s_i\|$, thereby, $re_i(f_i(w), f_{-i}(v')) > re_i(f_i(w), f_{-i}(v))$, where $\forall v \in \|s_i\|$. In terms of Definition 4.2, we can conclude that $M^*_{G'}, w \nvDash Ra^{re}_i$, which contradicting the hypothesis that $M^*_{G'}, w \vDash \wedge_{i \in N} Ra^{re}_i$. Since, for every $w \in W$, $f_i(w) \in IUD^0_i = S_i$, it follows that $f_i(w) \in IUD^0_i \setminus RD^0_i = IUD^1_i$.

Next we prove the inductive step. Fix an integer $m \geq 1$ and suppose that, for every player $j \in N$, $f_j(w) \in IUD_j^m$, we want to show that, for every player j, $f_j(w) \notin RD_j^m$. Suppose not. Then there exists a player i, satisfied that $f_i(w) \in RD_i^m$, that is, $f_i(w)$ is a regret-dominated in G'^m by some other strategy $s_i' \in IUD_i^m$. Then, $max\{re_i(f_i(w), s_{-i}), \forall s_{-i} \in IUD_{-i}^m\} > max\{re_i(s_i', s_{-i}), \forall s_{-i} \in IUD_{-i}^m\}$. Since, by hypothesis, for $\forall j \in N, f_j(w) \in IUD_j^m$, it follows-since the prosperity of $f_i(w)$, that is, $v \in R_i(w) \Leftrightarrow f_i(w) = f_i(v)$ that for $\forall v \in R_i(w), f_i(v) \in IUD_i^m$, further, we have: $max\{re_i(f_i(w), f_{-i}(v), v \in R_i(w)\} > max\{re_i(f_i(w'), f_{-i}(v)), v \in R_i(w')\}$, where $w' \in R_i(w) \cap \|s_{-i}'\|$. Thus, similar to the reason above, we can conclude $M^*_{G'^m}, w \nvDash Ra_i^{re}$, again contradicting the fact that $M^*_{G'}, w \vDash Ra_i^{re}$ since $M^*_{G'}$ is a submodel of $M^*_{G'^m}$ and the prosperities of $M^*_{G'}$. So, for every player i, $f_i(w) \in IUD_i^m \setminus RD_i^m = IUD_i^{m+1}$. By induction, $\forall i \in N, f_i(w) \in IUD_i$.

(b) From right to left: Let $f(w) \in IUD = \bigcap_{m \geqslant 0} IUD^m$, by Definition 2.3, $\forall i \in N$ $f_i(w)$ is never regret dominated IUD_i^m. So, it means that after m rounds of public announcement Ra^{re}, $M^*_{G'^m}, w \vDash Ra^{re}$, where $M^*_{G'^m}$ is a general epistemic model related to submodel G'^m. Therefore, in terms of the arbitrary of m and Definition 4.2, it is obvious that $w \in \sharp(Ra^{re}, M_{G'})$ □

6 Belief updating in the light of soft information

Alternatively, we can represent the procedure of *IERS* algorithm as the process of repeated $soft$ announcement of the rationality among players.[8] When this rationality assertion is believed to be true by every player: it is common belief that everybody believes that each of players is a the rational agent, the solution to the one-shot games is just outcome of the *IERS* algorithm.

To do this, firstly, we need to introduce some new concepts into $G' - logic$.

Definition 6.1. [9] A doxastic proposition P is true at a pointed model M[10] if it is true at an actual state w_0 in the model M, a radical upgrade $\Uparrow P$ (where P is a doxastic proposition.)is truthful in a model M if P is true at M. Moveover, a radical upgrade stream $\Uparrow \vec{P} = (\Uparrow P_n)_{n \in N}$ is an infinite sequence of upgrades $\Uparrow P_n$,

[8] The soft announcement refers to a radical update introduced in van Benthem (2004).

[9] Here, Definition 6.1 and 6.2 both are from Baltag and Smets (2009)

[10] A pointed model refers to a epistemic-doxastic model with a designated state w_0, which is called an actual state.

and the upgrade stream $\Uparrow \vec{P}$ is truthful if every $\Uparrow P_n$ is truthful with respect to every model M.

Definition 6.2. A repeated radical upgrade stabilizes $M_{G'}$ if it reaches a fixed point of $\Uparrow \vec{P}$, that is, repeated radical upgrade φ has no further effect.

Next, we redefine a frame $\mathfrak{F}'_{G'}$ of $G' - logic$ and an epistemic-doxastic regret-game model $M'_{G'}$ based on a given game G' with regret. In fact, we can provide the frame $\mathfrak{F}'_{G'}$ just adding a plausibility relation (we mentioned in the section 3) for every player i to the frame $\mathfrak{F}_{G'}$ Meanwhile, let $R'_i(w) = Min_{\preceq_i}([w]_i)$, and $\|s_i\|' = R'_i(w) \cap \|s_i\|$. We introduce the semantic interpretation for those game propositions in $M'_{G'}$ as follows,

Definition 6.3. An epistemic model $M'_{G'}$ over $G' - logic$ is obtained by incorporating the following valuation on a $\mathfrak{F}'_{G}$:

$M'_{G'}, w \vDash s_i \quad \Leftrightarrow \quad f_i(w) = s_i;$

$M'_{G'}, w \vDash (s_i \gtrsim s'_i) \quad \Leftrightarrow \quad \exists v \in \|s'_i\|'), re_i(s_i, f_{-i}(w)) \leq re_i(s'_i, f_{-i}(v))$

$M_{G'}, w \vDash (s_i \succ' s'_i) \quad \Leftrightarrow \quad \forall v \in \|s'_i\|'), re_i(s_i, f_{-i}(w)) < re_i(s'_i, f_{-i}(v))$

$M'_{G'}, w \vDash Ra_i^{re'} \quad \Leftrightarrow \quad M_{G'}, w \vDash s_i \wedge (\bigwedge_{a \neq s_i} B_i(s_i \gtrsim a)).$

Accordingly, by the definitions of radical upgrade in Baltag and Smets (2009), it is easy to justify that the radical upgrade stream $\Uparrow \vec{Ra^{re'}}$ is truthful, since it is reasonable that we take one of the worlds where GS is true as a actual world, and $Ra^{re'}$ always holds at the world. And in the light of a proposition (which is proved in Baltag and Smets (2009)):
Repeated truthful radical upgrade $\Uparrow \vec{P}$ in epistemic-doxastic logic stabilizes every model (with respect to which it is correct), we can have

Corollary 1. *Repeated truthful radical upgrade $\Uparrow \vec{Ra^{re'}}$ in epistemic-doxastic logic stabilizes every model (with respect to which it is correct).*

And similar to the Definition 3.1, we can define:,

Definition 6.4. For any epistemic model $M'_{G'}$ and formula φ, the radical upgrade stabilization $\#(\Uparrow \varphi, M'_{G'})$ is the first model in a repeated upgrade stream where upgrade φ has no further effect, and $W^{\#(\Uparrow \varphi, M'_{G'})}$ is the set of possible worlds, which agents consider the most likely after repeated upgrade φ, i.e., $W^{\#(\Uparrow \varphi, M'_{G'})} = \{w \in Min_{\preceq_{i \in N}}(W) \mid w \vDash \varphi\}$, and call it as a kernel of the $\#(\Uparrow \varphi, M'_{G'})$.

It is illustrated in the figure 4, how a radical upgrade $\Uparrow Ra^{re'}$ upgrades the regret-game illustrated in figure 1. Finally, we also show another characterization

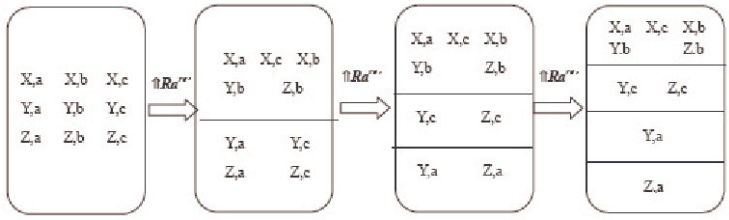


Figure 4: The radical update of $Ra^{re'}$

theorem for *IERS*.

Theorem 4. *Given a full epistemic-doxastic game model $M'_{G'}$, based on a regret-game G' and an arbitrary world w, $w \in W^{\#(\Uparrow Ra^{re'}, M'_{G'})}$ if and only if $f(w) \in IUD$.*

Proof. (a) From left to right: the proof is similar to the induction proof of Theorem 3 and left to the reader.

(b) From right to left: suppose $f(w) \in IUD = \bigcap_{m \geqslant 0} IUD^m$, but $w \notin W^{\#(\Uparrow Ra^{re'}, M'_{G'})}$, then, either $\#(\Uparrow Ra^{re'}, M'_{G'}), w \nvDash Ra^{re'}$ or $w \notin Min_{\leq_i \forall i \in N}(W)$.
On the one hand, if $Ra^{re'}$ does not hold at the world w, i.e., $\#(\Uparrow Ra^{re'}, M'_{G'}), w \nvDash Ra^{re'}$, then $f_i(w)$ is a regret-dominated strategy by some strategy for player i, because of the semantic definition of $Ra_i^{re'}$. Thus, $f_i(w) \notin IUD_i$, further, $f(w) \notin IUD_i$, contraction with the precondition;

On the other hand, if $w \notin Min_{\leq_i}(W)$, then there must be a model $M''_{G'}$, before repeated upgrade $\Uparrow \vec{Ra}^{re'}$ stabilize, so that $M''_{G'}, w \nvDash Ra^{re'}$. So, $\exists i \in N, M''_{G'}, w \nvDash Ra_i^{re'}$. Further, we can derive that $f_i(w)$ must be a regret-dominated for i, thus, $f_i(w) \notin IUD_i$, i.e., $f(w) \notin IUD$, which is also contradiction with the precondition. □

7 Related Work

In Halpern and Pass (2012), Halpern and Pass put forward a new game solution (they called an iterated regret minimization, similar to the regret equilibrium),

and stated the rationality and significance of the game solution by many examples from the game theory. Meanwhile, they also provided the epistemic characterization for the algorithm *IERS* (or say iterated regret minimization solution) using Kripke structure similar to the way we did, they defined the atomic proposition, " player i is rational at a world w in a epistemic game model, denoted by "RAT_i", as her current strategy is a best response to strategy sequences $\langle s(B_i^0(w)), s(B_i^1(w)), ...\rangle$ (where $B_i^0(w)$ consists of the worlds that player i considers most likely at w, and the worlds in $B_i^1(w)$ are less likely, and so on), i.e., $M, w \vDash RAT_i$ if $s_i(w)$ is a best response to the strategy sequence $\langle s(B_i^0(w)), s(B_i^1(w)), ...\rangle$. Moreover they proved, this game solutions resulted from the algorithm *IERS* involves higher and higher levels of belief regarding other players' rationality. At this point, we have the same viewpoint as theirs. We view an iterated elimination dominated procedure as a process of dynamic information exchange in the dynamic epistemic logic (PAL or Plausible Belief Revision Logic), it is natural that these higher levels of belief regarding other players' rationality become an implicit requirement for players' belief. The implication is derived from the essential prosperities of these dynamic logic, for example, after public announcing a formula φ in PAL, player i can delete the worlds in her mind, which are not satisfied the formula φ. In other words, she never reconsider the worlds as epistemic possible worlds for her. Similar scenario will happen in the belief revision with the radical upgrade $\Uparrow Ra^{re'}$, because agent i thinks the Ra^{re}-worlds become better than all the $\neg Ra^{re'}$, and keep the ordering at the later upgrade. Thus, she just considers her strategy based on those worlds satisfied Ra^{re} (or the worlds in her the best plausible area). It implies that player i's final choice must be based on the higher and higher the knowledge (or the belief) of other players' rationality. Nevertheless, in order to avoid a paradox similar to the paradox in the Iterated Admissibility (cf. Halpern and Pass (2012) and Brandenburger et al. (2008)), they also insisted that the higher and higher levels of belief regarding other players' rationality does not involves common knowledge or common belief, rather, higher levels of beliefs are accorded lower levels of likelihood (i.e., they assigned successively lower probability to higher orders of rationality). Considering this paradox does not arise in our way for the essential prosperities of our dynamic logic again, and the rationality defined by us is self-fulfilling. So, we keep well the classic rule in game theory. That is, it is necessary for analyze a game that rationality is common knowledge among players. Therefore, our dynamic analysis for *IERS* is more appealing, and it may be more suitable to be extended to Dynamic

Model Checking in computer science.[11]

Additionally, there is also a large amount of literatures on the algorithms of iterated elimination either in the field of logic, computer science and game theory. Bonanno (2008),Halpern and Pass (2012),van Benthem et al. (2011),Halpern and Pass (2009) etc. In particular, van Benthem (2007) describe and characterize different algorithms in game theory by redefining rationalities based on epistemic logic. Our intellectual debt towards van Benthem (2007) is clear. Compared with their work, we extend their findings in some sense. In fact, we can also restate their results based on our epistemic regret-game frame, provide a new kind of epistemic characterization for the algorithms, which have been studied by them. For example, van Benthem (2007) defined two types of rationality, the weak rationality and the strong rationality, which are denoted by WR_i and SR_i. Here, we redefine these rationality assertion on the epistemic regret-game frame as follows:

$$
\begin{array}{lcl}
M_{G'}, w \vDash (s_i \succcurlyeq^2 s_i') & \Leftrightarrow & (re_i(s_i, f_{-i}(w)) \leq re_i(s_i', f_{-i}(w))) \\
M_{G'}, w \vDash (s_i \succ^2 s_i') & \Leftrightarrow & (re_i(s_i, f_{-i}(w)) < re_i(s_i', f_{-i}(w))) \\
M_{G'}, w \vDash WR_i' & \Leftrightarrow & (M_{G'}, w) \models s_i \wedge (\bigwedge_{a \neq s_i} \langle K_i \rangle (f_i(w) \succcurlyeq^2 a)) \\
M_{G'}, w \vDash SR_i' & \Leftrightarrow & (M_{G'}, w) \models s_i \wedge \langle K_i \rangle (\bigwedge_{a \neq s_i} (f_i(w) \succcurlyeq^2 a)))
\end{array}
$$

Thus, a weak rational player i thinks it is possible that the regret raised by the current strategy is not greater than her other strategies. For example, she can know that there is no alternative action that she knows to reduce her regret, and strong rational player i thinks it is possible that the current strategy does not make her regret more. In other words, a player with strong rationality always a bit optimistic

Based the uniform structure, one can analyze and explore rationality implied iterated algorithms from players' regret perspective, also she can compare the strength of these rationality and the size of stable models.

Theorem 5. *The rationality Ra^{re} is stronger than the rationality WR', i.e., $Ra^{re} \rightarrow WR'$, and but not vise versa.*

Proof. The proof is common and easy, so we left it to the reader. The reader can find the similar proof in Cui and Tang (2010) and van Benthem (2007). □

[11]Some of our dynamic epistemic analysis for iterated elimination algorithms in the game theory have been extend in the field of Dynamic Model Checking, cf.Cui and Tang (2010)

Corollary 2. $\sharp(Ra^{re}, M_{G'}) \subseteq \sharp(WR', M_{G'})$

However, there is no relation between Ra^{re} and SR'. For instance, in the game G3 is from van Benthem (2011), and model G'3 is the regret game of G3. Ra_2^{re} holds at the worlds :$(A,a), (B,a), (C,a), (A,c)$,
$(B,c), (C,c)$, but SR'_2 is true at the worlds: $(A,a), (B,a), (C,a), (A,b), (B,b), (C,b)$.

player 1 \ player 2	a	b	c
A	(2,3)	(1,0)	(1,1)
B	(0,0)	(4,2)	(1,1)
C	(3,1)	(1,2)	(2,1)

G3

player 1 \ player 2	a	b	c
A	(1,0)	(3,3)	(1,2)
B	(3,2)	(0,0)	(1,1)
C	(0,1)	(3,0)	(0,1)

G'3

Figure 5: comparing Ra^{re} to SR'

8 Conclusion and Further Direction

The paper showed iterated hard and soft update is a powerful method and also can be applied into non-standard game solution algorithms such as minimizing regret. This brings more kinds of recent work in the foundations However, now it is time to go beyond proving such single results. Here are a few directions that we intend to pursue:

Introducing logics for qualitative reasoning: We will introduce modal logics over matrix games in the spirit of van Benthem et al. (2011) referring to agents available strategies, knowledge and preferences with propositional constants for positions of rationality and/or regret, and study the qualitative calculus of reasoning about these notions in interactive behavior.

Comparing, combining, and reducing methods: Comparing methods like *IESD* and *IERS*, we see that one may be better than another depending on the structure of a given game. We will investigate what happens when agents have a variety of such methods available. One possibility is that one method may simulate another, by means of translating the given game systematically into one with changed outcome values. Moreover, there are games where both methods make sense intuitively. We will start with sequential combinations of

solution methods, starting from very concrete questions such as whether or not $IESD; IERS = IESD; IERS$. The eventual goal would be an algebra of solution methods.

Linking up with limit behavior in learning theory: We have only considered cases where games get solved through iterated soft updates with regret statements. But many other scenarios can have the same features, including infinite sequences where the approximation behavior itself is the focus of interest. In particular, we are interested in connecting our setting with the learning-theoretic scenarios and extended temporal update logics suggested by the results of Baltag et al. (2011) and Pacuit and Roy (2011).

Acknowledgements We are grateful to Johan van Benthem and Olivier Roy for valuable suggestions. Cui is supported by Supported by National Natural Science Foundation of China (No.61173019), The fiftieth China Postdoctoral Science Foundation (No.2011M501370) and Philosophy and Social Science Youth Projects of Guang Dong Province (No.GD11YZX03).

References

R. J. Aumann. Rationality and bounded rationality. *Games and Economic Behavior*, 53:2–14, 1997.

A. Baltag and S. Smets. Group belief dynamics under iterated revision: fixed points and cycles of joint upgrades. In *Proceedings of the 12th Conference on Theoretical Aspects of Rationality and Knowledge*, TARK '09, pages 41–50, New York, NY, USA, 2009. ACM.

A. Baltag, L. S. Moss, and S. Solecki. The logic of public announcements, common knowledge and private suspicious. Technical Report SEN-R9922, CWI, Amsterdam, 1999.

A. Baltag, N. Gierasimczuk, and S. Smets. Belief revision as a truth-tracking process. In *TARK*, pages 187–190, 2011.

K. Basu. The traveler's dilemma. *Journal of the American Statistical Association*, 46:55–67, 2007.

T. Becker, M. Carter, and J. Naeve. Experts playing the traveler¡¯s dilemma. Discussion paper, 2005. 252.

D. Bernheim. Rationalizable strategic behavior. *Econometrica*, 52(4):1007–1028, 1984.

P. Blackburn, J. van Benthem, and F. Wolter. *Handbook of Modal Logic*. Elsevier Science Inc, 2007.

O. J. Board. Dynamic interactive epistemology. *Games and Economic Behavior*, 49:49–80, 2004.

G. Bonanno. *A syntactic approach to rationality in games with ordinal payoffs*, volume 3. Amsterdam: Amsterdam University Press, 2008.

A. Brandenburger, A. Friedenberg, and H. J. Keisler. Admissibility in games. *Econometrica*, 76(2):307–352, 2008.

C.Boutilier. *Conditional logics for defaoult reasoning and belief revision*. PhD thesis, Univerity of Toronto, 1992.

J. Cui and X. Tang. A method for solving nash equilibria of games based on public announcement logic. *Science China (Information Science)*, 53(7):1358–1368, 2010.

P. David. Rationalizable strategic behavior and the problem of perfection. *Econometrica*, 52(4):1029–1050, 1984.

H. Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*. Springer, Berlin, 2007.

J. Y. Halpern. *Reasoning About Uncertainty*. The MIT Press, Cambridge, Mass., 2003.

J. Y. Halpern and R. Pass. A logical characterization of iterated admissibility. In *Proceedings of the 12th Conference on Theoretical Aspects of Rationality and Knowledge*, TARK '09, pages 146–155, New York, NY, USA, 2009. ACM.

J. Y. Halpern and R. Pass. Iterated regret minimization: A new solution concept. *Games and Economic Behavior*, 74(1):184–207, 2012.

L.J.Savage. The theory of statistical decision. *J.Amer.Statist.Ass*, 46:55–67, 1951.

T. McKelvey, R.and Palfrey. An experimental study of the centipede game. *Econometrica*, 60(4):803–836, 1992.

E. Pacuit and O. Roy. A dynamic analysis of interactive rationality. In *Proceedings of the Third international conference on Logic, rationality, and interaction*, pages 244–257, 2011.

P.Lanarre and Y.Shoham. Knowledge, certainty, belief and conditionalisation. In *Proceedings of the International Conference on Knowledge Representaton and Reasning*, pages 415–424, 1994.

L. Renou and K. H. Schlag. Minimax regret and strategic uncertainty. *Journal of Economic Theory*, 145(1):264–286, 2010.

L. Renou and K. H. Schlag. Implementation in minimax regret equilibrium. *Games and Economic Behavior*, 71(2):527–533, 2011.

A. Rubinstein. *A Course in Game Theory*. The MIT Press, Cambridge, Mass., 1994.

J. Stoye. Axioms for minimax regret choice correspondences. *Journal of Economic Theory*, 146(6):2226–2251, 2011.

J. van Benthem. *Exploring Logical Dynamics*. ESLI Publications, Stanford,Mass., 1996.

J. van Benthem. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics*, 14(2):129–155, 2004.

J. van Benthem. Rational dynamics and epistemic logic in games. *International Game Theory Review*, 9(1):13–45, 2007.

J. van Benthem. *Logical Dynamcis of Information*. Cambridge University Press, 2011.

J. van Benthem, E. Pacuit, and O. Roy. Toward a theory of play: A logical perspective on games and interaction. *Games*, 2(1):52–86, 2011.

Questions about Voting Rules, With Some Answers

Jan van Eijck and Floor Sietsma

CWI & ILLC, CWI
jve@cwi.nl, f.sietsma@cwi.nl

Abstract

We raise questions about voting rules, and provide some of the answers. The method is to define a number of new formal properties of voting rules, and use these for classification and analysis. The aim is to get a better perspective on vices and virtues of individual voting rules.

Keywords: Voting rules, collective choice, multi-agent decision making.

1 Ballots, Profiles, Voting Rules

Voting is the process of selecting an item or a set of items from a finite set A of alternatives, on the basis of the stated preferences of a set of voters. See Brams and Fishburn (2002) for a detailed account. By calling the voters agents, voting can be seen as a form of multi-agent decision making.

A ballot is a linear ordering of A. Let $\mathbf{ord}(A)$ be the set of all ballots on A. We assume that the preferences of a voter are represented by a ballot. A profile is a vector of ballots, one for each voter. We assume voter anonymity, so it does not matter which voter has which ballot. The only thing that matters is the number

of voters holding a certain ballot. Under this assumption voting profiles can be represented as mappings from ballots to non-negative integers. Another way to say this is by saying that our profiles are *quantified*, and the voting rules to be introduced below act like quantifiers: they calculate an outcome based on numbers (of voters holding certain ballots).

We will use $\mathbf{P}, \mathbf{Q}$ to range over profiles, and $\mathbf{b}, \mathbf{b}'$ to range over ballots.

Profiles can be represented as lists of non-negative integers, where the length of the list equals $m!$, with m the number of alternatives. The size of a profile is equal to the length of its ballots. If a profile $\mathbf{P}$ has size m, this means that its alternative set A has $|A| = m$.

If $\mathbf{b}$ is a ballot and $\mathbf{P}$ a profile, we use $\mathbf{P}(\mathbf{b})$ for the number of voters with ballot $\mathbf{b}$ in $\mathbf{P}$.

For example, assume the set of alternatives A equals $\{a, b, c\}$. Then the ballot that has a in first position, b is second position, and c in third position is abc. The following represents the profile $\mathbf{P}$ with $\mathbf{P}(abc) = 2$ (two voters hold ballot abc), $\mathbf{P}(bca) = 6$ (six voters hold ballot bca), and so on:

$$(abc, 2), (bca, 6), (cab, 0), (acb, 4), (cba, 0), (bac, 2).$$

Profiles can be normalized by dividing with the gcd of the list of all nonzero vote numbers. If $\mathbf{P}$ is a profile, we use $\mathbf{P}^\circ$ for the normalized form of the profile. The normalized form of the above example profile is:

$$(abc, 1), (bca, 3), (cab, 0), (acb, 2), (cba, 0), (bac, 1).$$

Definition 1.1. An (anonymous) voting rule V for set of alternatives A is a function from A-profiles to $\mathcal{P}^+(A)$ (the set of non-empty subsets of A). A voting rule V is *resolute* if V maps every profile to a singleton set. If $V(\mathbf{P}) = B$, then the members of B are called the winners of $\mathbf{P}$ under V.

Anonymity means that all voters are treated equally. This is built into our framework because we take profiles to be given by numbers of voters for each ballot.

If $\mathbf{P}$ is a profile for A, and π is a permutation of A, then $\mathbf{P}^\pi$ is the result of replacing x by $\pi(x)$ everywhere in $\mathbf{P}$. If $B \subseteq A$, then $\pi(B) = \{\pi(x) \mid x \in B\}$.

Definition 1.2. A voting rule V is *neutral* if for every profile $\mathbf{P}$ and for every permutation π of the set A of alternatives,

$$V(\mathbf{P}^\pi) = \pi(V(\mathbf{P})).$$

Neutrality means that all alternatives are treated equally.

Definition 1.3. A voting rule V is *normal* if it holds for every profile $\mathbf{P}$ that $V(\mathbf{P}) = V(\mathbf{P}^\circ)$.

Proposition 1. *There are anonymous and neutral voting rules that are not normal.*

Proof. Let V_k be given by $x \in V_k(\mathbf{P})$ if at least k voters have x at the top of their ballots. Then V_k is anonymous and neutral, but V_k is not normal. □

Question 1. *Characterize the normal voting rules.*

A scoring vector for ballots of size m is a list of non-negative integers $(w_0, \ldots, w_{m-1})$ satisfying $w_i \geq w_{i+1}$. The number w_i indicates the weight of position i in the ballot. The plurality rule has scoring vector $(1, 0, \cdots, 0)$. The anti-plurality rule (or: veto rule) has scoring vector $(1, \cdots, 1, 0)$. The Borda rule (see Borda (1781)) has scoring vector $(m-1, m-2, \cdots, 1, 0)$. The trivial voting rule that always returns the set of all alternatives has scoring vector $(0, \ldots, 0)$.

Every scoring vector w determines a voting rule S_w by means of:

$$S_w(\mathbf{P}) = \{x \in A \mid x \text{ has maximal } w\text{-scores in } \mathbf{P}\}.$$

For any scoring vector $w = (w_0, \ldots, w_{m-1})$, let w° be the result of dividing out common factors in $(w_0 - w_{m-1}, \ldots, w_{m-2} - w_{m-1}, 0)$. Call w° the normalization of w.

Proposition 2. *Scoring vector normalization does not affect the set of winners: for all* $\mathbf{P}$ *and all scoring vectors* w *it holds that* $S_w(\mathbf{P}) = S_{w^\circ}(\mathbf{P})$.

Proof. Let $(w_1, \ldots, w_{m-1})$ be a scoring vector. If x is a winner under this vector for profile $\mathbf{P}$, this means that the score N of x for $\mathbf{P}$ is maximal among the scores, i.e., greater than or equal to the score M of any alternative $y \neq x$. Scoring for the vector $(w_1 - w_{m-1}, \ldots, w_{m-2} - w_{m-1}, 0)$ give scores $N - kmw_{m-1}$ and $M - kmw_{m-1}$, so the score of x is still maximal. In the other direction, the scores change by adding a constant, so winners are also preserved.

Next, compare $(w_1, \ldots, w_{m-1})$ and $(w_1K, \ldots, w_{m-1}K)$, with $K > 1$. Scores M and N for x and y under $(w_1, \ldots, w_{m-1})$ change into MK and NK. Since $M > N$ iff $MK > NK$, winners are not affected in either direction. □

Absolute majority is the voting rule that selects an alternative with more than 50 % of the votes as winner, and returns the whole set of alternatives otherwise. This is not the same as plurality, which selects an alternative that has the maximum number of votes as winner, regardless of whether more than half of the voters voted like this or not. Unanimity: if all voters have an alternative a at the top of their ballots then a is the winner, otherwise all alternatives tie for a win. Near-unanimity: if all but at most one of the voters have an alternative a at the top of their ballots then a is the winner, otherwise all alternatives tie for a win.

In the examples below we also use the Condorcet rule. the Copeland rule and the Hare rule. Here are the definitions (see also Taylor (2005)).

A Condorcet winner is an alternative that beats every other alternative in pairwise contests. An alternative x beats another alternative y in a one-to-one contest if more than half of the voters prefer x to y. The Condorcet voting rule (proposed in 1785 by the marquis of Condorcet in Condorcet (1785)) selects the Condorcet winner if it exists, and the set of all alternatives otherwise. The Copeland voting rule Copeland (1951) selects the alternative that maximizes the difference between the number of won and lost pairwise majority contests. The voting rule of single transferable vote, also known as the Hare rule (see Hare (1861); the rule is also described by John Stuart Mill, with an attribution to Thomas Hare, in Mill (1861)), works as follows. If one of the candidates gets an absolute majority, that candidate wins. Otherwise prune the candidate(s) who is/are ranked first by the fewest number of voters from the profile, and repeat.

2 Profile Restriction

Profile restriction is computing a new profile for a subset of the alternative set of the original profile. The relative preferences of the voters in the new profile should remain unchanged.

If $B \subseteq A$, we use $\mathbf{P}^B$ for the result of restricting $\mathbf{P}$ to B. Formally, let $\mathbf{b} \sim_B \mathbf{b'}$ if the ballots $\mathbf{b}$ and $\mathbf{b}'$ become the same after restriction to the set B. Then $\mathbf{P}^B$ is given by

$$\mathbf{P}^B(\mathbf{b}) = \sum \{\mathbf{P}(\mathbf{b}') \mid \mathbf{b}' \in \mathbf{ord}(A), \mathbf{b} \sim_B \mathbf{b}'\}.$$

For example, let **P** be the following profile:

$$(abc, 1), (bca, 2), (cab, 0), (acb, 3), (cba, 0), (bac, 2).$$

Then the restriction of **P** to $\{a, b\}$ is given by

$$(ab, 4), (ba, 4),$$

the restriction of **P** to $\{a, c\}$ is given by

$$(ac, 6), (ca, 2),$$

and the restriction of **P** to $\{b, c\}$ is given by

$$(bc, 5), (cb, 3).$$

Definition 2.1. A voting rule V is invariant for restriction if it holds for every $B \subseteq A$ and every profile **P** that

$$V(\mathbf{P}) \neq A \text{ and } V(\mathbf{P}) \cap B \neq \emptyset \text{ imply } V(\mathbf{P}) \cap B = V(\mathbf{P}^B).$$

Note: Invariance for restriction can be viewed as a strengthening of a property that is known as Chernoff's condition Chernoff (1951), or as Sen's property alpha Sen (1970), or as Arrow's principle of invariance for irrelevant alternatives Arrow (1951, second edition: 1963), applied to voting rules. A voting rule V satisfies this condition if winners in a subset B of the set of all alternatives remain winners if the choice is limited to B. In our terminology: if $V(\mathbf{P}) \cap B \subseteq V(\mathbf{P}^B)$.

Proposition 3. *The Hare rule and the Copeland rule are not invariant for restriction.*

Proof. For the Hare rule, consider the following profile **P** (ballots that are not mentioned get 0 votes):

$$(abc, 3), (bca, 2), (cab, 2).$$

If V is the Hare rule we get $V(\mathbf{P}) = \{a\}$. The restricted profile $\mathbf{P}^{\{a,c\}}$ looks like this:

$$(ac, 3), (ca, 4).$$

This gives $V(\mathbf{P}^{\{a,c\}}) = \{c\}$.

For the Copeland rule, consider the following profile:

$$(bacde, 1), (acdeb, 1), (debac, 1).$$

Under the Copeland rule, this is a win for a. Next, restrict the profile to $\{a, b, c\}$. This gives:

$$(bac, 2), (acb, 1).$$

Now b is the Copeland winner. □

Theorem 1. *The Condorcet rule is invariant for restriction.*

Proof. If there are no Condorcet winners then there is nothing to prove. A winner in the contest between a and b in $\mathbf{P}$ is still a winner in a contest between a and b in $\mathbf{P}^B$ for any B with $\{a, b\} \subseteq B$, and vice versa. □

Theorem 2. *Positional scoring rules with weights* $(w_0, ..., w_{m-1})$ *such that* $w_0 > w_{m-1}$ *are not invariant for restriction.*

Proof. Consider the following profile P consisting of 3 ballots of 7 voters:

$$(abc, 3), (bca, 2), (cab, 2).$$

Suppose the scoring rule gives weights (w_0, w_1, w_2) to the three positions. Then the scores of the candidates are as follows:

$$\begin{aligned} a &: 3w_0 + 2w_1 + 2w_2 \\ b &: 2w_0 + 3w_1 + 2w_2 \\ c &: 2w_0 + 2w_1 + 3w_2 \end{aligned}$$

Note that the difference in score between a and c is exactly $w_0 - w_2$. Since by assumption $w_0 > w_2$, the score of a is larger than that of c. This means that the set of winners is either $V(P) = \{a\}$ or $V(P) = \{a, b\}$. Now let us remove b from the set of candidates. In both cases, the intersection of $V(\mathbf{P})$ with the set of remaining candidates is $\{a\}$. The profile that remains after removing b is the following:

$$(ac, 3), (ca, 4).$$

Now since there is a different number of candidates, the scoring rule may give different weights to the positions. Suppose the weights are (v_0, v_1). Then the scores of the candidates are as follows:

$$\begin{aligned} a &: 3v_0 + 4v_1 \\ c &: 4v_0 + 3v_1 \end{aligned}$$

By assumption $v_0 > v_1$, so c wins the election. Because $V(\mathbf{P}) \cap \{a, c\} = \{a\}$, this shows that V is not invariant for restriction. □

Question 2. *Characterize the voting rules that are invariant for restriction.*

Question 3. *Is invariance for restriction a desirable property for a voting rule to have, or not?*

The last question may seem a bit vague, but in any case, here are some relevant observations. Notice that restriction destroys information. If there are m alternatives and k voters then there are $m!$ possible ballots. The number of integer solutions for

$$x_1 + \cdots + x_n = k$$

under the condition that $x_i \geq 0$ for all $i = 1, \ldots, n$ is $\binom{n+k-1}{k}$ (Jukna 2011, Proposition 1.5). Thus, for m alternatives and k voters there are

$$\binom{m! + k - 1}{k}$$

possible profiles. There are m ways to prune away one alternative. After pruning, there are $(m-1)!$ possible ballots, which leaves

$$\binom{(m-1)! + k - 1}{k}$$

profiles. All in all this gives

$$m\binom{(m-1)! + k - 1}{k}$$

possibilities.

To put these outcomes in perspective, here are some calculations for $m = 4$ and $k = 10$:

$$\binom{4! + 10 - 1}{10} = 92561040.$$

$$4\binom{3! + 10 - 1}{10} = 12012.$$

To see that the information destruction is vast, consider the case where the pruning process leaves only pairs. m alternatives give $m(m-1)$ pairs, so after pair pruning there are only $m(m-1)(k+1)$ possibilities left, since there are $k+1$ ways to split k into non-negative integers k_1, k_2 with $k_1 + k_2 = k$. For 4 alternatives and 10 voters, this reduces the number of possibilities from 92561040 to 132.

3 Profile Addition, Additivity of Voting Rules

Intuitively, we can merge two elections into a single election, by adding the numbers of votes for the various ballots. Call this operation $\oplus$. Note that the two operand profiles have to be of the same size (i.e., over the same set of alternatives). Note also that $(\mathbf{P} \oplus \mathbf{P})^\circ = \mathbf{P}^\circ$.

Definition 3.1. A voting rule V is *additive* if it holds for all m-profiles $\mathbf{P}$ and $\mathbf{Q}$ that $V(\mathbf{P}) \cap V(\mathbf{Q}) \subseteq V(\mathbf{P} \oplus \mathbf{Q})$. Or in words: V is additive if winners of two separate elections concerning the same set of alternatives remain winners if the elections are merged.

The following definition is from Young (1975).

Definition 3.2. A voting rule V is *consistent* if it holds for all m-profiles $\mathbf{P}$ and $\mathbf{Q}$ that $V(\mathbf{P}) \cap V(\mathbf{Q}) \neq \emptyset$ implies $V(\mathbf{P}) \cap V(\mathbf{Q}) = V(\mathbf{P} \oplus \mathbf{Q})$.

Clearly, every consistent rule is additive, but the property of additivity is weaker than the property of consistency: see Proposition 8 below.

The requirement of additivity seems entirely reasonable. Still, there are respectable voting rules that do not satify it.

Proposition 4. *The Condorcet rule is not additive.*

Proof. Consider the following two profiles $\mathbf{P}$ and $\mathbf{Q}$:

$$(abc, 3), (bca, 0), (cab, 0), (acb, 0), (cba, 0), (bac, 2),$$

$$(abc, 1), (bca, 1), (cab, 1), (acb, 3), (cba, 3), (bac, 3).$$

The first of these has Condorcet winner a, the second has no Condorcet winner. So $V(\mathbf{P}) = \{a\}$ and $V(\mathbf{Q}) = \{a, b, c\}$, and therefore $V(\mathbf{P}) \cap V(\mathbf{Q}) = \{a\}$. Their sum is:

$$(abc, 4), (bca, 1), (cab, 1), (acb, 3), (cba, 3), (bac, 5).$$

The Condorcet winner of this sum is b. □

A voting rule satisfies the *Condorcet Criterion* if it always elects the Condorcet winner if there is one. The above proposition should worry anyone who thinks of the Condorcet criterion as a benchmark for voting rule quality.

Proposition 5. *The Hare rule is not additive.*

Proof. Consider the following two profiles (ballots that are not mentioned have no voters):

$$\mathbf{P} = \{(abcd, 5), (bacd, 6), (cabd, 2), (dabc, 10)\}.$$

$$\mathbf{Q} = \{(abcd, 4), (bacd, 4), (cabd, 8), (dabc, 2)\}.$$

If V is the Hare rule, then $V(\mathbf{P}) = V(\mathbf{Q}) = \{a\}$, and $V(\mathbf{P} \oplus \mathbf{Q}) = \{b\}$. □

Question 4. *Is the Copeland rule additive?*

Proposition 6. *The majority, unanimity and near-unanymity rules are additive.*

Proof. Suppose $\mathbf{P}$ and $\mathbf{Q}$ are m-profiles, V is the majority rule, and $a \in V(\mathbf{P}) \cap V(\mathbf{Q})$. Let $\mathbf{P}$ have N voters and $\mathbf{Q}$ have M voters. Then either no $x \in A$ has an absolute majority, or more than $N/2$ ballots in $\mathbf{P}$ have a in first position. Similarly, either no $x \in A$ has an absolute majority in $\mathbf{Q}$, or more than $M/2$ ballots in $\mathbf{Q}$ have a in first position. It follows that either no $x \in A$ has an absolute majority in $\mathbf{P} \oplus \mathbf{Q}$, in which case $a \in V(\mathbf{P} \oplus \mathbf{Q}) = A$, or $(N + M)/2$ ballots in $\mathbf{P} \oplus \mathbf{Q}$ have a in first position, i.e., a is the majority winner in $\mathbf{P} \oplus \mathbf{Q}$.

Same reasoning for the unanimity and near-unanymity rule. □

Proposition 7. *The near-unanymity rule is not consistent.*

Proof. Let V be the near-unanimity rule and let $\mathbf{P}$ be the following profile:

$$(ab, 2), (ba, 1).$$

Then $V(\mathbf{P}) = \{a\}$ and $V(\mathbf{P} \oplus \mathbf{P}) = \{a, b\}$. This shows that V is not consistent. □

Proposition 8. *Additivity does not imply consistency.*

Proof. Immediate from Propositions 6 and 7. □

Theorem 3. *Every positional voting rule is additive.*

Proof. Let V be a positional voting rule, and let $\mathbf{P}$, $\mathbf{Q}$ be a pair of m-profiles, for some m. Suppose $a \in V(\mathbf{P}) \cap V(\mathbf{Q})$. We have to show that $a \in V(\mathbf{P} \oplus \mathbf{Q})$. But this is immediate from the fact that if the score of a is maximal in $\mathbf{P}$ and $\mathbf{Q}$, it is also maximal in $\mathbf{P} \oplus \mathbf{Q}$. □

Question 5. *Can we prove an if and only if for additivity?*

4 Cycles, Reduction

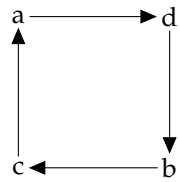
Definition 4.1. A permutation of alternatives π on $A = \{a_0, \ldots, a_{m-1}\}$ is a *full cycle* if π can be given as $a_0 = \pi^0(a_0) \mapsto \pi(a_0) \mapsto \pi^2(a_0) \mapsto \cdots \mapsto \pi^{m-1}(a_0)$, with the $\pi^i(a_0)$ all different.

Any full cycle on A can be considered as a linear ordering on A with a_0 as least element, and vice versa. Thus, there are $(m-1)!$ full cycles on $\{a_0, \ldots, a_{m-1}\}$.

Customary notation for full cycles π on a list of m elements is to give the list:

$$(a_0, \pi(a_0), \pi^2(a_0), \cdots, \pi^{m-1}(a_0)).$$

For example, the full cycle in the following picture can be given as $(adbc)$.



So cycles can also be represented as ballots. Moreover, cycles can be used to classify ballots. Two ballots **b** and **b′** are in the same ballot cycle if there is a full cycle π on A and a number k such that π^k maps **b** to **b′**. If the ballot size is m, then each ballot is part of a cycle of m ballots.

For example, the ballot *abcd* is part of the following cycle:

$$abcd, bcda, cdab, dabc.$$

The following definition is from Saari Saari (1995).

Definition 4.2. A profile is *reduced* if each cycle in the profile contains a ballot with no voters.

Example 1. The profile

$$(abc, 3), (bca, 1), (cab, 0), (acb, 2), (cba, 0), (bac, 2)$$

is reduced.

Explanation: there are two cycles, $\{abc, bca, cab\}$ and $\{acb, cba, bac\}$, and both have a ballot with no voters.

Definition 4.3. A profile is *balanced* if each cycle in the profile is such that each ballot in the cycle has the same number of voters. Use **B** for balanced profiles.

Example 2. The profile

$$(abc, 1), (bca, 1), (cab, 1), (acb, 3), (cba, 3), (bac, 3)$$

is balanced.

Proposition 9. *For every profile* **P** *there exist a reduced* **Q** *and a balanced* **B** *such that* $\mathbf{P} = \mathbf{Q} \oplus \mathbf{B}$.

Definition 4.4. If $\mathbf{P} = \mathbf{Q} \oplus \mathbf{B}$, as in Proposition 9, then call **B** the *surplus* of **P** and **Q** the reduced form of **P**. Use $\mathbf{P}^r$ for the reduced form of **P**.

Proposition 10. *A profile* **P** *is both balanced and reduced iff* **P** *has no voters.*

Definition 4.5. Call the operation of subtracting a balanced profile from **P** *reduction*. Call the operation of adding a balanced profile to **P** *dilution*.

> Here is an obvious **algorithm** for putting a profile **P** in reduced form:
>
> For each cycle π of **P**, let the minimum of the vote numbers in that cycle be k. Subtract k from every vote number in the cycle.

The surplus of a profile indicates by how much the profile can be reduced.

Example 3. The surplus of the profile

$$(abc, 4), (bca, 2), (cab, 1), (acb, 3), (cba, 3), (bac, 6)$$

is the profile

$$(abc, 1), (bca, 1), (cab, 1), (acb, 3), (cba, 3), (bac, 3).$$

Example 4. The reduced form of the profile

$$(abc, 4), (bca, 2), (cab, 1), (acb, 3), (cba, 3), (bac, 6)$$

is the profile

$$(abc, 3), (bca, 1), (cab, 0), (acb, 0), (cba, 0), (bac, 3).$$

Theorem 4. *Any anonymous and neutral voting rule maps a balanced profile to the set of all alternatives.*

Proof. Let $\mathbf{P}$ be a balanced profile for A. Let V be an anonymous and neutral voting rule. We must prove that $V(\mathbf{P}) = A$.

Suppose not, i.e., suppose there is some $b \notin V(\mathbf{P})$. There also is some $a \in V(\mathbf{P})$, for $V(\mathbf{P}) \neq \emptyset$.

Let σ be any permutation of A that satisfies $\sigma(a) = b$.

Observe that each cycle will remain a cycle under the permutation σ. Therefore, because of anonymity and the fact that $\mathbf{P}$ is balanced: $\mathbf{P}^\sigma = \mathbf{P}$. Because of neutrality $V(\mathbf{P}^\sigma) = \sigma(V(\mathbf{P}))$, and therefore $b = \sigma(a) \in V(\mathbf{P}^\sigma) = V(\mathbf{P})$, and contradiction. □

Theorem 5. *If $|A| = m$ then the number of voters in any balanced profile for A is a multiple of m.*

Proof. Each cycle in an m-profile has m elements. There are $(m-1)!$ cycles. Let cycle i have k_i voters. Then all in all we have $m \sum_{i=1}^{(m-1)!} k_i$ voters. □

Definition 4.6. A voting rule *V is safe for dilution* if it holds for all profiles $\mathbf{P}$ and balanced profiles $\mathbf{B}$ that $V(\mathbf{P}) \supseteq V(\mathbf{P} \oplus \mathbf{B})$.

Safety for dilution means that dilution does not introduce new winners.

Definition 4.7. A voting rule *V is safe for reduction* if it holds for all profiles $\mathbf{P}$ and balanced profiles $\mathbf{B}$ that $V(\mathbf{P}) \subseteq V(\mathbf{P} \oplus \mathbf{B})$.

Safety for reduction means that reduction does not introduce new winners.

Theorem 6. *Any anonymous, neutral and additive voting rule is safe for reduction.*

Proof. Assume V is anonymous and neutral. Then $V(\mathbf{B})$ equals the set of all alternatives. By additivity we have:

$$V(\mathbf{P}) = V(\mathbf{P}) \cap V(\mathbf{B}) \subseteq V(\mathbf{P} \oplus \mathbf{B}).$$

□

Proposition 11. *The Condorcet rule is neither safe for reduction nor safe for dilution.*

Proof. Consider the profile:

$$(abc, 1), (bac, 3), (bca, 1), (acb, 5), (cab, 4), (cba, 3).$$

The Condorcet winner for this profile is a. The reduced form of this is:

$$(abc, 0), (bac, 0), (bca, 0), (acb, 2), (cab, 3), (cba, 0).$$

The Condorcet winner for the reduced profile is c. □

Proposition 12. *The absolute majority rule is safe for reduction, but not safe for dilution.*

Proof. The example from Proposition 11 works here as well. In the reduced profile

$$(abc, 0), (bac, 0), (bca, 0), (acb, 2), (cab, 3), (cba, 0)$$

there is an absolute majority for c. Dilute this profile with

$$(abc, 1), (bac, 1), (bca, 1), (acb, 3), (cab, 3), (cba, 3).$$

There is no absolute majority in the diluted profile

$$(abc, 1), (bac, 3), (bca, 1), (acb, 5), (cab, 4), (cba, 3).$$

□

Theorem 7. *Any voting rule V with positional scoring will assign to every alternative in a balanced profile* **B** *the same score.*

Proof. Let **B** be a balanced m-profile. Then there are $(m-1)!$ cycles, and there are k_i voters in each ballot in the i-th cycle. Let V be a positional voting rule with $(x_0, \cdots, x_{m-1})$ as its scoring vector. Let π_i be an arbitrary cycle of **P**, let a be an arbitrary alternative, and let j be an arbitrary position (i.e., $0 \leq j < m$). Then the score for a for this position in the cycle under the voting rule is given by $k_i x_j$, for a occurs in this position exactly once in the cycle. Summing over the cycles, we get that a collects the following score in **B**:

$$\sum_{i=1}^{(m-1)!} k_i x_j.$$

Summing over the positions, we see that a collects the score:

$$\sum_{j=0}^{m-1} \sum_{i=1}^{(m-1)!} k_i x_j.$$

Since a was arbitrary, every alternative collects this same score. □

Theorem 8. *Any voting rule V with positional scoring is safe for reduction and safe for dilution.*

Proof. Let $\mathbf{P}$ be an m-profile, and let $\mathbf{B}$ be a balanced m-profile.

Since $\mathbf{B}$ is balanced, it follows from the previous Theorem that the scores for the alternatives under V for $\mathbf{P}$ can be computed from those for $\mathbf{P} \oplus \mathbf{B}$ by subtracting a constant c from each score, and vice versa, by adding a constant c to each score. These subtractions and additions do not affect the outcome of V. □

Question 6. *Does the converse hold as well? If a voting rule is safe in both directions, does it follow that it is positional?*

If this is too difficult to answer, the following questions may be easier:

Question 7. *If a voting rule is safe in both directions, does it follow that it is additive?*

Question 8. *If a voting rule is safe in both directions, does it follow that it is consistent?*

Notice that for all voting rules V that are not invariant under reduction, the derived voting rule V^r defined by $V^r(\mathbf{P}) = V(\mathbf{P}^r)$ is different from V. Also, for any voting rule V, the derived voting rule V^r is invariant for reduction and dilution by definition.

Question 9. *What are the formal properties of the Condorcet^r rule?*

Question 10. *Are there non-positional voting rules V with the property that V^r is positional?*

5 Strategizing

Strategizing is replacing a ballot $\mathbf{b}$ by a different one, $\mathbf{b}'$, in the hope or expectation to get a better outcome (where better is "closer to $\mathbf{b}$" in some sense).

As is explained in Taylor (2005), there are many ways to interpret 'better'. One way is that X is better than Y if X weakly dominates Y, that is if every $x \in X$ is at least as good as every $y \in Y$ and some $x \in X$ is better than some $y \in Y$. Formally:

Definition 5.1. If $X, Y \subseteq A$ $X \neq \emptyset$, $Y \neq \emptyset$, and $\mathbf{b} \in \mathbf{ord}(A)$, then $X >_{\mathbf{b}} Y$ if $\forall x \in X \forall y \in Y$: $x = y$ or x is above y in $\mathbf{b}$, and $\exists x \in X \exists y \in Y$: x is above y in $\mathbf{b}$.

Let $\mathbf{P} \sim_i \mathbf{P}'$ express that $\mathbf{P}$ and $\mathbf{P}'$ differ only in the ballot of voter i.

Definition 5.2. A voting rule is *strategy-proof* if $\mathbf{P} \sim_i \mathbf{P}'$ implies $V(\mathbf{P}) \geq_i V(\mathbf{P}')$, where $\geq_i$ expresses 'betterness' according to the i-ballot in $\mathbf{P}$.

Note: the following definition does not assume voter anonymity. The definition uses $\mathbf{P}^{-i}$ for the result of removing the ballot of voter i from profile $\mathbf{P}$.

Definition 5.3. A voting rule V is *monotone* if for any profile $\mathbf{P}$ and any alternative $a \in V(P)$, if $\mathbf{b}_i'$ is a new ballot for some voter i that results from moving a up in the ranking of i and not changing the order between the other alternatives, then $a \in V(\mathbf{P}^{-i} \cup \mathbf{b}_i')$.

Definition 5.4. A voting rule is *resolute* if $V(\mathbf{P})$ is a singleton for any profile $\mathbf{P}$.

Theorem 9. *Any resolute voting rule that is monotone and invariant for restriction is strategy-proof.*

Proof. Take some resolute voting rule V, profile $\mathbf{P}$ and voter i. Suppose $V(\mathbf{P}) = \{a\}$. Then for any other candidate c, $V(P^{\{a,c\}}) = \{a\}$ by winner preservation under restriction. Suppose i can strategize by submitting some dishonest ballot $\mathbf{b}_i'$ in order to elect some candidate b such that $b >_i a$.

Let $V(\mathbf{P}^{-i} \cup \mathbf{b}_i') = \{b\}$. It is possible that $a >_i' b$. By monotonicity, if we construct the ballot $\mathbf{b}_i''$ by moving b up until $b >_i'' a$ then $V(P^{-i} \cup \mathbf{b}_i'') = \{b\}$. By winner preservation under restriction, $V(\mathbf{P}^{\{a,b\}}) = a$. But because $b >_i'' a$, $(\mathbf{P}^{-i} \cup \mathbf{b}_i'')^{\{a,b\}} = \mathbf{P}^{\{a,b\}}$ so $V((\mathbf{P}^{-i} \cup \mathbf{b}_i'')^{\{a,b\}}) = V(\mathbf{P}^{\{a,b\}}) = \{a\}$. This contradicts our assumption that $b >_i a$, so strategizing is not possible. □

The concept of weak domination is borrowed from game theory (see, e.g., Osborne (2004)). As Taylor (Taylor 2005, p. 39) remarks:

> In point of fact, an election can be thought of as a game in which a strategy for a player (voter) is a choice of ballot, and the outcome of the game is the set of winners in the election.

To formalize this, let a ballot vector for A be a list of A-ballots $(\mathbf{b}_0, \ldots, \mathbf{b}_{n-1})$. We assume that a ballot vector represents the true ballots of voters $\{0, \ldots, n-1\}$, in the sense that $\mathbf{b}_i$ represents the true preferences of voter i.

Define a payoff function in terms of $\geq_{\mathbf{b}}$ from Definition 5.1, as follows.

Definition 5.5. $\text{payoff}(\mathbf{b}, X) = |\{Y \mid Y \in \mathcal{P}^+(A), X >_{\mathbf{b}} Y\}|$.

Thus, the payoff of a voting outcome X, given a ballot $\mathbf{b}$ serving as a point of reference, is the size of the set of possible voting outcomes that are strictly worse than X.

This payoff function can be used to define the value of a move for a player with true ballot $\mathbf{b}$, as follows:

Definition 5.6. $\text{move}(V, \mathbf{P}, \mathbf{b}, \mathbf{b}') = \text{payoff}(\mathbf{b}, V(\mathbf{P}'))$, where $\mathbf{P}'$ is the result of adding ballot $\mathbf{b}'$ to $\mathbf{P}$.

The game for voting rule V and ballot vector $(\mathbf{b}_0, \ldots, \mathbf{b}_{n-1})$ is now given in terms of the move function, as follows.

Definition 5.7. Assume $\mathbf{P}$ is some profile for $n-1$ voters. Then

$$\text{Game}(V, (\mathbf{b}_0, \ldots, \mathbf{b}_{n-1}), \mathbf{P}, i) = \{(\mathbf{b}, \text{move}(V, \mathbf{P}, \mathbf{b}_i, \mathbf{b}) \mid \mathbf{b} \in \mathbf{ord}(A)\}.$$

Thus, we see that a voting rule together with a ballot vector determines an n-player game $\text{Game}(V, (\mathbf{b}_0, \ldots, \mathbf{b}_{n-1}))$, where each voter has a choice between the members of $\mathbf{ord}(A)$ (the possible ballots), and where the payoff for player i for a profile $\mathbf{P}$ for $n-1$ voters, and for (cast) ballot $\mathbf{b}$ is given by $\text{move}(V, \mathbf{P}, \mathbf{b}_i, \mathbf{b})$.

Clearly, a voting rule V is strategy-proof iff it holds for each ballot vector $(\mathbf{b}_0, \ldots, \mathbf{b}_{n-1})$ that the profile $\mathbf{P}$ corresponding to vector $(\mathbf{b}_0, \ldots, \mathbf{b}_{n-1})$ is a Nash equilibrium for $\text{Game}(V, (\mathbf{b}_0, \ldots, \mathbf{b}_{n-1}))$. But we can take a more general perspective:

Question 11. *Characterize the ballot vectors for which Game$(V, (\mathbf{b}_0, \ldots, \mathbf{b}_{n-1}))$ (for given V) has nontrivial pure Nash equilibria.*

The following proposition shows that there are many trivial Nash equilibria.

Proposition 13. *Let V be a voting rule and $\mathbf{P}$ a profile. If for all i and $\mathbf{P}'$ with $\mathbf{P} \sim_i \mathbf{P}'$ it holds that $V(\mathbf{P}) = V(\mathbf{P}')$, then $\mathbf{P}$ is a Nash equilibrium for Game$(V, (\mathbf{b}_0, \ldots, \mathbf{b}_{n-1}))$, for any ballot vector $(\mathbf{b}_0, \ldots, \mathbf{b}_{n-1})$.*

Proof. No voter has an incentive to deviate from his ballot in $\mathbf{P}$, as it makes no difference for the outcome. □

Players who realize they have lost the game have no incentive to strategize. Similarly for players who realize they have won the game. If all players know they are in one of these two categories, no strategizing will occur. Compare also Chopra et al. (2004) for a first analysis of the crucial role of knowledge in strategic voting.

Question 12. *Analyze the abstention game for a voting rule and a ballot vector, where each player has the choice between casting his true ballot or abstaining from the vote. A voting rule V is abstention-proof if it holds for each ballot vector $(\mathbf{b}_0, \ldots, \mathbf{b}_{n-1})$ that the profile corresponding to that vector is a Nash equilibrium for the abstention game for V and $(\mathbf{b}_0, \ldots, \mathbf{b}_{n-1})$. Characterize the voting rules that are abstention-proof.*

6 Conclusion and Further Research

We have introduced a number of concepts to classify and analyze voting rules: invariance for restriction, additivity, safety for dilution, safety for reduction. We have demonstrated the use of these concepts by proving some new results about voting rules. Further clarification of relations between voting rules will no doubt result from finding answers to the list of questions we have left open. Answering the list of questions we have raised (or in some cases, finding the answers in the literature) is future work.

We have an implemented system for voting with anonymous voting rules that we used for checking a number of the factual propositions in this paper. The present version of the software implements strategizing, under the assumption that the rest of the profile is known to the strategizer. Our intention is to extend this implementation to an epistemic model checker for voting under partial uncertainty about the profile. The software is available on the internet as a literate Haskell program Eijck and Sietsma (2012).

Acknowledgement Thanks to Krzysztof Apt, Ulle Endriss and Sunil Simon for enlightening discussions on the topic of this paper. The participants of the Lorentz workshop on Modeling Strategic Reasoning (Leiden, Feb 20–24, 2012) and three anonymous LAMAS 2012 reviewers also gave useful feedback.

References

K. Arrow. *Social Choice and Individual Values*. Wiley, New York, 1951, second edition: 1963.

J.-C. d. Borda. *Mémoire sur les élections au scrutin*. Histoire de l'Académie Royale des Sciences, Paris, 1781.

S. J. Brams and P. C. Fishburn. Voting procedures. In K. Arrow, A. Sen, and K. Suzumura, editors, *Handbook of Social Choice and Welfare*, volume I, chapter 4. Elsevier, 2002.

H. Chernoff. Rational selection of decision functions. *Econometrica*, 22:422–443, 1951.

S. Chopra, E. Pacuit, and R. Parikh. Knowledge-theoretic properties of strategic voting. In *Proceedings of the 9th European Conference on Logics in Artificial Intelligence (JELIA-2004)*, pages 18–30, 2004.

M. l. M. d. Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, Paris, 1785.

A. Copeland. A "reasonable" social welfare function. Seminar on Mathematics in Social Sciences, 1951.

J. v. Eijck and F. Sietsma. Basic voting theory. Literate Haskell Program, available from `www.cwi.nl/~jve/software/voting/`, 2012.

T. Hare. *The Election of Representatives, Parliamentary and Municipal: A Treatise*. Longman, Green, London, 1861.

S. Jukna. *Extremal Combinatorics, with Applications in Computer Science — Second Edition*. Texts in Theoretical Computer Science. Springer, 2011.

J. S. Mill. *Considerations of a Representative Government*. Parker, Son, and Bourn, London, 1861. Electronically available from Project Gutenberg.

M. J. Osborne. *An Introduction to Game Theory*. Oxford University Press, New York, Oxford, 2004.

D. Saari. *Basic Geometry of Voting*. Springer, 1995.

A. Sen. *Collective Choice and Social Welfare*. Holden-Day, 1970.

A. D. Taylor. *Social Choice and the Mathematics of Manipulation*. Cambridge University Press, 2005.

H. Young. Social choice scoring functions. *SIAM Journal on Applied Mathematics*, 28(4):824–836, 1975.

Playing extensive form games in parallel

Sujata Ghosh, R. Ramanujam and Sunil Simon

Indian Statistical Institute, SETS Campus, Taramani, Chennai
Institute of Mathematical Sciences, CIT Campus, Taramani, Chennai
CWI, Amsterdam
sujata@isichennai.res.in, jam@imsc.res.in, s.e.simon@cwi.nl

Abstract

Consider a player playing against different opponents in two extensive form games simultaneously. Can she then have a strategy in one game using information from the other? The famous example of playing chess against two grandmasters simultaneously illustrates such reasoning. We consider a simple dynamic logic of extensive form games with sequential and parallel composition in which such situations can be expressed. We present a complete axiomatization and show that the satisfiability problem for the logic is decidable.

1 Motivation

How can any one of us [1] expect to win a game of chess against a Grandmaster (GM)? The strategy is simple: play simultaneously against two Grandmasters! If we play black against GM 1 playing white, and in the parallel game play white against GM 2 playing black, we can do this simply. Watch what GM 1 plays, play that move in the second game, get GM 2's response, play that *same*

[1] By "us" we mean poor mortals who know how to play the game but lack expertise.

move as our response in game 1, and repeat this process. If one of the two GMs wins, we are assured of a win in the other game. In the worst case, both games will end in a draw.

Note that the strategy construction in this example critically depends on several features:

- Both games need to be played in *lock-step synchrony*; if they are slightly out of step with each other, or are sequentialized in some way, the strategy is not applicable. So concurrency is critically exploited.
- The strategy cannot be constructed *a priori*, as we do not know what moves would be played by either of the GMs. Such reasoning is intrinsically different from the discussion of the existence of winning strategies in determined games. In particular, strategic reasoning as in normal form games is not applicable.
- The common player in the two games acts as a conduit for transfer of information from one game to the other; thus *game composition* is essential for such reasoning. The example illustrates that playing several instances of the same game may mean something very different from repeated games.
- The common player can be a resource bounded agent who cannot analyse the entire game structure and compute the winning strategy (even if it exists). The player thus mimics the moves of an "expert" in order to win one of the constituent games.

In general, when extensive form games are played in parallel, with one player participating in several games simultaneously, such an information transfer from one game to the other is possible. In general, since strategies are structured in extensive form games, they can make use of such information in a non-trivial manner.

In the context of agent-based systems, agents are supposed to play several interactive roles at the same time. Hence when interaction is modelled by games (as in the case of negotiations, auctions, social dilemma games, market games, etc.) such parallel games can assume a great deal of importance. Indeed, a prominent feature of an agent in such a system is the ability to *learn* and transferring strategic moves from one game to the other can be of importance as one form of learning.

Indeed, sequential composition of games can already lead to interesting situations. Consider player A playing a game against B, and after the game is over, playing another instance of the *same* game against player C. Now each of the leaf nodes of the first game carries important historical information about play in the game, and A can strategize differently from each of these nodes in the second game, thus reflecting learning again. Negotiation games carry many such instances of history-based strategizing.

What is needed is an algebra of game composition in which the addition of a parallel operator can be studied in terms of how it interacts with the other operators like choice and sequential composition. This is reminiscent of process calculi, where equivalence of terms in such algebras is studied in depth.

In this paper, we follow the seminal work of Parikh (Parikh (1985)) on **propositional game logic**. We use dynamic logic for game expressions but extended with parallel composition; since we wish to take into account game structure, we work with extensive form games embedded in Kripke structures rather than with effectivity functions. In this framework, we present a complete axiomatization of the logic and show that the satisfiability problem for the logic is decidable.

The interleaving operator has been looked at in the context of program analysis in terms of dynamic logic Abrahamson (1980). The main technical difficulty addressed in the paper is that parallel composition is not that of sequences (as typically done in process calculi) but that of trees. The main modality of the logic is an assertion of the form $\langle g, i\rangle\alpha$ which asserts, at a state s, that a tree t in the "tree language" associated with g is enabled at s, and that player i has a strategy (subtree) in it to ensure α. Parallel composition is not compositional in the standard logical sense: the semantics of $g_1|g_2$ is not given in terms of the semantics of g_1 and g_2 considered as wholes, but by going into their structure. Therefore, defining the enabled-ness of a strategy as above is complicated. Note that the branching structure we consider is quite different from the intersection operator in dynamic logic Harel (1984), Danecki (1984), Lange and Lutz (2005) and is closer to the paradigm of concurrent dynamic logic Peleg (1987).

For ease of presentation, we first present the logic with only sequential and parallel composition and discuss technicalities before considering iteration, which adds a great deal of complication. Note that the dual operator, which is important in Parikh's game logic is not relevant here, since we wish to consider games between several players played in parallel.

Related work

Games have been extensively studied in temporal and dynamic logics. For concurrent games, this effort was pioneered by work on Alternating time temporal logic (ATL) Alur et al. (2002), which considers selective quantification over paths. Various extension of ATL was subsequently proposed, these include ones in which strategies can be named and explicitly referred to in the formulas of the logic van der Hoek et al. (2005), Ågotnes (2006), Walther et al. (2007). Parikh's work on propositional game logics Parikh (1985) initiated the study of game structures in terms of algebraic properties. Pauly Pauly (2001) has built on this to reason about abilities of coalitions of players. Goranko draws parallels between Pauly's coalition logic and ATL Goranko (2001). Van Benthem uses dynamic logic to describe games and strategies van Benthem (2002). Strategic reasoning in terms of a detailed notion of agency has been studied in the *stit* framework Horty (2001), Broersen (2010), Broersen et al. (2006).

Somewhat closer in spirit is the work of van Benthem et al. (2008) where van Benthem and co-authors develop a logic to reason about simultaneous games in terms of a parallel operator. The reasoning is based on powers of players in terms of the outcome states that can be ensured. Our point of departure is in considering extensive form game trees explicitly and looking at interleavings of moves of players in the tree structure.

2 Preliminaries

2.1 Extensive form games

Let $N = \{1, \ldots, n\}$ denote the set of players, we use i to range over this set. For $i \in N$, we often use the notation $\bar{i}$ to denote the set $N \setminus \{i\}$. Let Σ be a finite set of action symbols representing moves of players, we let a, b range over Σ. For a set X and a finite sequence $\rho = x_1 x_2 \ldots x_m \in X^*$, let $last(\rho) = x_m$ denote the last element in this sequence.

Game trees:

Let $\mathbb{T} = (S, \Rightarrow, s_0)$ be a tree rooted at s_0 on the set of vertices S and $\Rightarrow : (S \times \Sigma) \to S$ is a *partial* function specifying the edges of the tree. The tree $\mathbb{T}$ is said to be finite if S is a finite set. For a node $s \in S$, let $\vec{s} = \{s' \in S \mid s \stackrel{a}{\Rightarrow} s' \text{ for some } a \in \Sigma\}$, $moves(s) = \{a \in \Sigma \mid \exists s' \in S \text{ with } s \stackrel{a}{\Rightarrow} s'\}$ and $E_T(s) = \{(s, a, s') \mid s \stackrel{a}{\Rightarrow} s'\}$. By $E_T(s) \times x$ we denote the set $\{((s, x), a, (s', x)) \mid (s, a, s') \in E_T(s)\}$. The set $x \times E_T(s)$ is defined similarly. A node s is called a leaf node (or terminal node) if $\vec{s} = \emptyset$. The depth of a tree is the length of the longest path in the tree.

An extensive form game tree is a pair $T = (\mathbb{T}, \widehat{\lambda})$ where $\mathbb{T} = (S, \Rightarrow, s_0)$ is a tree. The set S denotes the set of game positions with s_0 being the initial game position. The edge function $\Rightarrow$ specifies the moves enabled at a game position and the turn function $\widehat{\lambda} : S \to N$ associates each game position with a player. Technically, we need player labelling only at the non-leaf nodes. However, for the sake of uniform presentation, we do not distinguish between leaf nodes and non-leaf nodes as far as player labelling is concerned. An extensive form game tree $T = (\mathbb{T}, \widehat{\lambda})$ is said to be finite if $\mathbb{T}$ is finite. For $i \in N$, let $S^i = \{s \mid \widehat{\lambda}(s) = i\}$ and let $frontier(T)$ denote the set of all leaf nodes of T. Let $S_T^L = frontier(T)$ and $S_T^{NL} = S \setminus S_T^L$. For a tree $T = (S, \Rightarrow, s_0, \widehat{\lambda})$ we use $head(T)$ denote the depth one tree generated by taking all the outgoing edges of s_0.

A play in the game T starts by placing a token on s_0 and proceeds as follows: at any stage if the token is at a position s and $\widehat{\lambda}(s) = i$ then player i picks an action which is enabled for her at s, and the token is moved to s' where $s \stackrel{a}{\Rightarrow} s'$. Formally a play in T is simply a path $\rho : s_0 a_1 s_1 \cdots$ in $\mathbb{T}$ such that for all $j > 0$, $s_{j-1} \stackrel{a_j}{\Rightarrow} s_j$. Let $Plays(T)$ denote the set of all plays in the game tree T.

2.2 Strategies

A strategy for player $i \in N$ is a function μ^i which specifies a move at every game position of the player, i.e. $\mu^i : S^i \to \Sigma$. A strategy μ^i can also be viewed as a subtree of T where for each player i node, there is a unique outgoing edge and for nodes belonging to players in $\bar{\imath}$, every enabled move is included. Formally we define the strategy tree as follows: For $i \in N$ and a player i strategy $\mu^i : S^i \to \Sigma$ the strategy tree $T_{\mu^i} = (S_{\mu^i}, \Rightarrow_{\mu^i}, s_0, \widehat{\lambda}_{\mu^i})$ associated with μ is the least

subtree of T satisfying the following property: $s_0 \in S_{\mu^i}$,

- For any node $s \in S_{\mu^i}$,
  - if $\widehat{\lambda}(s) = i$ then there exists a unique $s' \in S_{\mu^i}$ and action a such that $s \stackrel{a}{\Rightarrow}_{\mu^i} s'$.
  - if $\widehat{\lambda}(s) \neq i$ then for all s' such that $s \stackrel{a}{\Rightarrow} s'$, we have $s \stackrel{a}{\Rightarrow}_{\mu^i} s'$.

Let $\Omega^i(T)$ denote the set of all strategies for player i in the extensive form game tree T. A play $\rho : s_0 a_0 s_1 \cdots$ is said to be consistent with μ^i if for all $j \geq 0$ we have $s_j \in S^i$ implies $\mu^i(s_j) = a_j$.

2.3 Composing game trees

We consider sequential and parallel composition of game trees. In the case of sequences, composing them amounts to concatenation and interleaving. Concatenating trees is less straightforward, since each leaf node of the first is now a root of the second tree. Interleaving trees is not the same as a tree obtained by interleaving paths from the two trees, since we wish to preserve choices made by players.

Sequential composition:

Suppose we are given two finite extensive form game trees $T_1 = (S_1, \Rightarrow_1, s_1^0, \widehat{\lambda}_1)$ and $T_2 = (S_2, \Rightarrow_2, s_2^0, \widehat{\lambda}_2)$. The sequential composition of T_1 and T_2 (denoted $T_1; T_2$) gives rise to a game tree $T = (S, \Rightarrow, s_0, \widehat{\lambda})$, defined as follows: $S = S_1^{NL} \cup S_2$, $s_0 = s_1^0$,

- $\widehat{\lambda}(s) = \widehat{\lambda}_1(s)$ if $s \in S_1^{NL}$ and $\widehat{\lambda}(s) = \widehat{\lambda}_2(s)$ if $s \in S_2$.
- $s \stackrel{a}{\Rightarrow} s'$ iff:
  - $s, s' \in S_1^{NL}$ and $s \stackrel{a}{\Rightarrow}_1 s'$, or
  - $s, s' \in S_2$ and $s \stackrel{a}{\Rightarrow}_2 s'$, or

– $s \in S_1^{NL}, s' = s_2^0$ and there exists $s'' \in S_1^L$ such that $s \stackrel{a}{\Rightarrow}_1 s''$.

In other words, the game tree $T_1; T_2$ is generated by pasting the tree T_2 at all the leaf nodes of T_1. The definition of sequential composition can be extended to a set of trees $\mathcal{T}_2$ (denoted $T_1; \mathcal{T}_2$) with the interpretation that at each leaf node of T_1, a tree $T_2 \in \mathcal{T}_2$ is attached.

Parallel composition:

The parallel composition of T_1 and T_2 (denoted $T_1 | T_2$) yields a set of trees. A tree $t = (S, \Rightarrow, s_0, \widehat{\lambda})$ in the set of trees $T_1 | T_2$ provided: $S \subseteq S_1 \times S_2$, $s_0 = (s_1^0, s_2^0)$,

- For all $(s, s') \in S$:
  - $E_T((s, s')) = E_{t_1}(s) \times s'$ and $\widehat{\lambda}(s, s') = \widehat{\lambda}_1(s)$, or
  - $E_T((s, s')) = s \times E_{t_2}(s')$ and $\widehat{\lambda}(s, s') = \widehat{\lambda}_2(s')$.
- For every edge $s_1 \stackrel{a}{\Rightarrow}_1 s_1'$ in t_1, there exists $s_2 \in S_2$ such that $(s_1, s_2) \stackrel{a}{\Rightarrow} (s_1', s_2)$ in t.
- For every edge $s_2 \stackrel{a}{\Rightarrow}_2 s_2'$ in t_2, there exists $s_1 \in S_1$ such that $(s_1, s_2) \stackrel{a}{\Rightarrow} (s_1, s_2')$ in t.

3 Examples

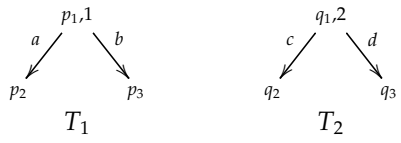


Figure 1: atomic games

Consider the trees T_1 and T_2 given in Figure 1. The sequential composition of T_1 and T_2 (denoted $T_1; T_2$) is shown in Figure 2. This is obtained by pasting the tree T_2 at all the leaf nodes of T_1.

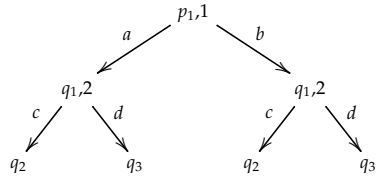
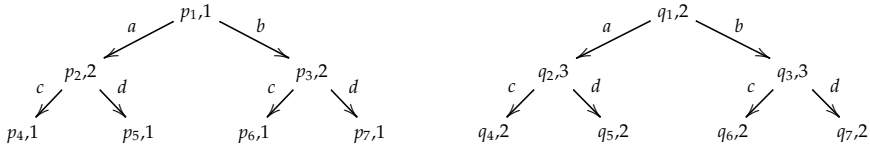


Figure 2: $T_1; T_2$

Now consider two finite extensive form game trees T_4 and T_5 given in figure 3. Each game is played between two players, player 2 is common in both games.



(a) T_4 (b) T_5

Figure 3: Atomic games

Note that we are talking about different instances of the same game (as evident from the similar game trees) played between different pairs of players with a player in common. Consider the interleaving of T_4 and T_5 where player 1 moves first in T_4, followed by 2 and 3 in T_5, and then again coming back to the game T_4, with the player 2-moves. This game constitutes a valid tree in the set of trees defined by $T_4|T_5$ and is shown in Figure 4.

Due to space constraints, we have not provided the names for each of the states in the parallel game tree, but they are quite clear from the context. The game starts with player 1 moving from p_1 in T_4 to p_2 or p_3. Then the play moves to the game T_5, where player 2 moves to q_2 or q_3, followed by the moves of player 3. After that, the play comes back to T_4, where player 2 moves once again.

These games clearly represent toy versions of "playing against two Grandmasters simultaneously". Players 1 and 3 can be considered as the Grandmasters, and 2 as the poor mortal. Let us now describe the copycat strategy that can be

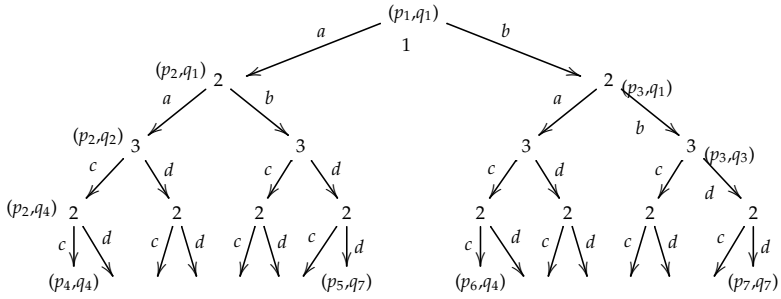
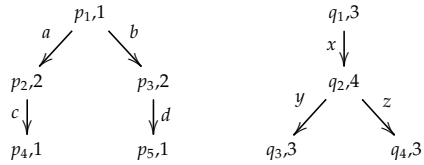


Figure 4: Game tree T

used by player 2, when the two games are played in parallel. The simultaneous game (figure 4), starts with player 1 making the first move a, say in the game tree T_4 (from (p_1, q_1)) to move to (p_2, q_1). Player 2 then copies this move in game T_5, to move to (p_2, q_2). The game continues in T_5, with player 3 moving to (p_2, q_4), say. Player 2 then copies this move in T_4 (playing action c) to move to (p_4, q_4). This constitutes a play of the game, where player 2 copies the moves of players 1 and 3, respectively.

Evidently, if player 1 has a strategy in T_4 to achieve a certain objective, whatever be the moves of player 2, following the same strategy, player 2 can attain the same objective in T_5.

Parallel composition can also be performed with respect to games structures which are not the same. Consider the game trees T_6 and T_7 given in Figure 5.



(a) T_6 (b) T_7

Figure 5: Atomic games

An interleaved game where each game is played alternatively starting from the game T_6 can be represented by the game tree in Figure 6.

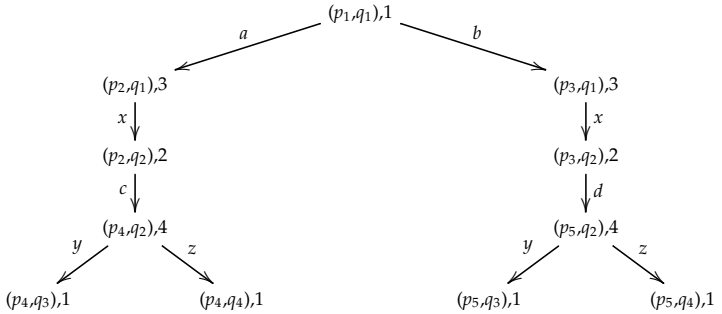


Figure 6: A game tree in $T_6|T_7$

4 The logic

For a finite set of action symbols Σ, let $\mathcal{T}(\Sigma)$ be a countable set of finite extensive form game trees over the action set Σ which is closed under subtree inclusion. That is, if $T \in \mathcal{T}(\Sigma)$ and T' is a subtree of T then $T' \in \mathcal{T}(\Sigma)$. We also assume that for each $a \in \Sigma$, the tree consisting of the single edge labelled with a is in $\mathcal{T}(\Sigma)$. Let $\mathbb{H}$ be a countable set and h, h' range over this set. Elements of $\mathbb{H}$ are referred to in the formulas of the logic and the idea is to use them as names for extensive form game trees in $\mathcal{T}(\Sigma)$. Formally we have a map $\nu : \mathbb{H} \to \mathcal{T}(\Sigma)$ which given any name $h \in \mathbb{H}$ associates a tree $\nu(h) \in \mathcal{T}(\Sigma)$. We often abuse notation and use h to also denote $\nu(h)$ where the meaning is clear from the context.

4.1 Syntax

Let P be a countable set of propositions, the syntax of the logic is given by:

$$\Gamma := h \mid g_1; g_2 \mid g_1 \cup g_2 \mid g_1|g_2$$

$$\Phi := p \in P \mid \neg\alpha \mid \alpha_1 \vee \alpha_2 \mid \langle g, i\rangle\alpha$$

where $h \in \mathbb{H}$ and $g \in \Gamma$.

In Γ, the atomic construct h specifies a finite extensive form game tree. Composite games are then constructed using the standard dynamic logic operators along with the parallel operator. $g_1 \cup g_2$ denotes playing g_1 or g_2. Sequential composition is denoted by $g_1; g_2$ and $g_1 | g_2$ denotes the parallel composition of games.

The main connective $\langle g, i \rangle \alpha$ asserts at state s that a tree in g is enabled at s and that player i has a strategy subtree in it at whose leaves α holds.

4.2 Semantics

A model $M = (W, \rightarrow, \widehat{\lambda}, V)$ where W is the set of states (or game positions), $\rightarrow \subseteq W \times \Sigma \times W$ is the move relation, $V : W \rightarrow 2^P$ is a valuation function and $\widehat{\lambda} : W \rightarrow N$ is a player labelling function. These can be thought of as standard Kripke structures whose states correspond to game positions along with an additional player labelling function. An extensive form game tree can be thought of as *enabled* at a certain state, say s of a Kripke structure, if we can embed the tree structure in the tree unfolding of the Kripke structure rooted at s. We make this notion more precise below.

Enabling of trees:

For a game position $u \in W$, let T_u denote the tree unfolding of M rooted at u. We say the game h is enabled at a state u if the structure $\nu(h)$ can be embedded in T_u with respect to the enabled actions and player labelling. Formally this can be defined as follows:

Given a state u and $h \in \mathbb{H}$, let $T_u = (S^s_M, \Rightarrow_M, \widehat{\lambda}_M, s)$ and $\nu(h) = T_h = (S_h, \Rightarrow_h, \widehat{\lambda}_h, s_{h,0})$. The restriction of T_u with respect to the game tree h (denoted $T_u \restriction h$) is the subtree of T_s which is generated by the structure specified by T_h. The restriction is defined inductively as follows: $T_u \restriction h = (S, \Rightarrow, \widehat{\lambda}, s_0, f)$ where $f : S \rightarrow S_h$. Initially $S = \{s\}$, $\widehat{\lambda}(s) = \widehat{\lambda}_M(s)$, $s_0 = s$ and $f(s_0) = s_{h,0}$.

For any $s \in S$, let $f(s) = t \in S_h$. Let $\{a_1, \ldots, a_k\}$ be the outgoing edges of t, i.e. for all $j : 1 \leq j \leq k$, $t \stackrel{a_j}{\Rightarrow}_h t_j$. For each a_j, let $\{s^1_j, \ldots, s^m_j\}$ be the nodes in S^s_M such that

$s \overset{a_j}{\Rightarrow}_M s^l_j$ for all $l : 1 \leq l \leq m$. Add nodes $s^1_j, \ldots, s^m_j$ to S and the edges $s \overset{a_j}{\Rightarrow} s^l_j$ for all $l : 1 \leq l \leq m$. Also set $\widehat{\lambda}(s^l_j) = \widehat{\lambda}_M(s^l_j)$ and $f(s^l_j) = t_j$.

We say that a game h is enabled at u (denoted $enabled(h, u)$) if the tree $T_u \restriction h = (S, \Rightarrow, \widehat{\lambda}, s_0, f)$ satisfies the following properties: for all $s \in S$,

- $moves(s) = moves(f(s))$,
- if $moves(s) \neq \emptyset$ then $\widehat{\lambda}(s) = \widehat{\lambda}_h(f(s))$.

Interpretation of atomic games:

To formally define the semantics of the logic, we need to first fix the interpretation of the compositional games constructs. In the dynamic logic approach, for each game construct g and player i we would associate a relation $R^i_g \subseteq (W \times 2^W)$ which specifies the outcome of a winning strategy for player i. However due to the ability of being able to interleave game positions, in this setting we need to keep track of the actual tree structure rather just the "input-output" relations, which is closer in spirit to what is done in process logics Harel et al. (1982) . Thus for a game g and player i we define the relation $R^i_g \subseteq 2^{(W \times W)^*}$. For a pair $\mathbf{x} = (u, w) \in W \times W$ and a set of sequences $Y \in 2^{(W \times W)^*}$ we define $(u, w) \cdot Y = \{(u, w) \cdot \rho \mid \rho \in Y\}$. For $j \in \{1, 2\}$ we use $\mathbf{x}[j]$ to denote the j-th component of $\mathbf{x}$.

For each atomic game h and each state $u \in W$, we define $R^i_h(u)$ in a bottom-up manner in such a way that whenever h is enabled at u, $R^i_h(u)$ encodes the set of all available strategies (cf. Section 2.2) for player i in the game h enabled at u. The collection of all such strategies that a player i can have, whenever the game h is enabled at some state $u \in W$ is given by R^i_h.

Let $h = (S, \Rightarrow, s_0, \widehat{\lambda})$ be a depth 1 tree with $moves(s_0) = \{a_1, \ldots, a_k\}$ and for all $s \neq s_0$, $moves(s) = \emptyset$. For $i \in N$ and a state $u \in W$, we define $R^i_h(u) \subseteq 2^{(W \times W)^*}$ as follows:

- If $\widehat{\lambda}(s_0) = i$ then $R^i_h(u) = \{X_j \mid enabled(h, u)$ and $X_j = \{(u, w_j)\}$ where $u \overset{a_j}{\rightarrow} w_j\}$.
- if $\widehat{\lambda}(s_0) \in \bar{i}$ then $R^i_h(u) = \{\{(u, w_j) \mid enabled(h, u)$ and $\exists a_j \in moves(s_0)$ with $u \overset{a_j}{\rightarrow} w_j\}\}$.

For $g \in \Gamma$, let $R^i_g = \bigcup_{u \in W} R^i_g(u)$.

For a tree $h = (S, \Rightarrow, s_0, \widehat{\lambda})$ such that $depth(h) > 1$, we define $R^i_h(u)$ as,

- if $\widehat{\lambda}(s_0) = i$ then $R^i_h(u) = \{\{(u,w) \cdot Y\} \mid \exists X \in R^i_{head(h)}$ with $(u,w) \in X, u \xrightarrow{a_j} w$ and $Y \in R^i_{h_{a_j}}\}$
- if $\widehat{\lambda}(s_0) \in \bar{i}$ then $R^i_h(u) = \{\{(u,w) \cdot Y \mid \exists X \in R^i_{head(h)}$ with $(u,w) \in X, u \xrightarrow{a_j} w$ and $Y \in R^i_{h_{a_j}}\}\}$.

Remark: Note that a set $X \in R^i_h$ can contain sequences such as $(u,w)(v,x)$ where $w \neq v$. Thus in general sequence of pairs of states in X need not represent a subtree of T_u for some $u \in W$. We however need to include such sequences since if h is interleaved with another game tree h', a move enabled in h' could make the transition from w to v. A sequence $\varrho \in X$ is said to be legal if whenever $(u,w)(v,x)$ is a subsequence of ϱ then $w = v$. A set $X \subseteq 2^{(W \times W)^*}$ is a valid tree if for all sequence $\varrho \in X$, ϱ is legal and X is prefix closed. For X which is a valid tree we have the property that for all $\varrho, \varrho' \in X$, $first(\varrho)[1] = first(\varrho')[1]$. We denote this state by $root(X)$. We also use $frontier(X)$ to denote the frontier nodes, i.e. $frontier(X) = \{last(\varrho)[2] \mid \varrho \in X\}$.

For a game tree h, although every set $X \in R^i_h$ need not be a valid tree, we can associate a tree structure with X (denoted $\mathfrak{T}(X)$) where the edges are labelled with pairs of the form (u,w) which appears in X. Conversely given $W \times W$ edge labelled finite game tree $\mathfrak{T}$, we can construct a set $X \subseteq 2^{(W \times W)^*}$ by simply enumerating the paths and extracting the labels of each edge in the path. We denote this translation by $\mathfrak{f}(\mathfrak{T})$. We use these two translations in what follows:

Interpretation of composite games:

For $g \in \Gamma$ and $i \in N$, we define $R^i_g \subseteq 2^{(W \times W)^*}$ as follows:

- $R^i_{g_1 \cup g_2} = R^i_{g_1} \cup R^i_{g_2}$.
- $R^i_{g_1 ; g_2} = \{\mathfrak{f}(\mathfrak{T}(X); \mathcal{T}) \mid X \in R^i_{g_1}$ and $\mathcal{T} = \{\mathfrak{T}(X_1), \ldots, \mathfrak{T}(X_k)\}$ where $\{X_1, \ldots, X_k\} \subseteq R^i_{g_2}\}$.

- $R^i_{g_1|g_2} = \{\mathfrak{f}(\mathfrak{T}(X_1)|\mathfrak{T}(X_2)) \mid X_1 \in R^i_{g_1} \text{ and } X_2 \in R^i_{g_2}\}$.

The truth of a formula $\alpha \in \Phi$ in a model M and a position u (denoted $M, u \models \alpha$) is defined as follows:

- $M, u \models p$ iff $p \in V(u)$.
- $M, u \models \neg\alpha$ iff $M, u \not\models \alpha$.
- $M, u \models \alpha_1 \vee \alpha_2$ iff $M, u \models \alpha_1$ or $M, u \models \alpha_2$.
- $M, u \models \langle g, i\rangle\alpha$ iff $\exists X \in R^i_g$ such that X constitutes a valid tree, $root(X) = u$ and for all $w \in frontier(X)$, $M, w \models \alpha$.

A formula α is satisfiable if there exists a model M and a state u such that $M, u \models \alpha$.

Let h_1 and h_2 be the game trees T_4 and T_5 given in Figure 3. The tree in which the moves of players are interleaved in lock-step synchrony is one of the trees in the semantics of $h_1|h_2$. This essentially means that at every other stage if a depth one tree is enabled then after that the same tree structure is enabled again, except for the player labelling. Given the (finite) atomic trees, we can write a formula α_{LS} which specifies this condition. If the tree h is a minimal one, i.e. of depth one given by $(S, \Rightarrow, s_0, \widehat{\lambda})$, α_{LS_h} can be defined as, $\bigwedge_{a_j \in moves(s_0)}(\langle a_j\rangle\top \wedge [a_j](\wedge_{a_j \in moves(s_0)}\langle a_j\rangle\top)$.

If player 1 has a strategy (playing a, say) to achieve certain objective ϕ in the game h_1, player 2 can play (copy) the same strategy in h_2 to ensure ϕ. This phenomenon can be adequately captured in the interleaved game structure, where player 2 has a strategy (viz. playing a) to end in those states of the game $h_1|h_2$, where player 1 can end in h_1. So we have that, whenever h_1 and $h_1|h_2$ are enabled and players can move in lock-step synchrony with respect to the game h_1 (or, h_2), $\langle h_1, 1\rangle\phi \rightarrow \langle h_1|h_2, 2\rangle\phi$ holds.

5 Axiom system

The main technical contribution of this paper is a sound and complete axiom system. Firstly, note that the logic extends standard PDL (without iteration).

For $a \in \Sigma$ and $i \in N$, let T^i_a be the tree defined as: $T^i_a = (S, \Rightarrow, s_0, \widehat{\lambda})$ where $S = \{s_0, s_1\}$, $s_0 \stackrel{a}{\Rightarrow} s_1$, $\widehat{\lambda}(s_0) = i$ and $\widehat{\lambda}(s_1) \in N$. Let t^i_a be the name denoting this tree, i.e. $\nu(t^i_a) = T^i_a$. For each $a \in \Sigma$ we define,

- $\langle a \rangle \alpha = \bigwedge_{i \in N} (\mathbf{turn}_i \supset \langle t^i_a, i \rangle \alpha)$.

From the semantics it is easy to see that we get the standard interpretation for $\langle a \rangle \alpha$, i.e. $\langle a \rangle \alpha$ holds at a state u iff there is a state w such that $u \stackrel{a}{\rightarrow} w$ and α holds at w.

Enabling of trees: The crucial observation is that the property of whether a game is enabled can be described by a formula of the logic. Formally, for $h \in \mathbb{H}$ such that $\nu(h) = (S, \Rightarrow, s_0, \widehat{\lambda})$ and $moves(s_0) \neq \emptyset$ and an action $a \in moves(s_0)$, let h_a be the subtree of T rooted at a node s' with $s_0 \stackrel{a}{\Rightarrow} s'$. The formula $h^{\surd}$ (defined below) is used to express the fact that the tree structure $\nu(h)$ is enabled and $head^{\surd}_h$ to express that $head(\nu(h))$ is enabled. This is defined as,

- If $\nu(h)$ is atomic then $h^{\surd} = \top$ and $head^{\surd}_h = \top$.
- If $\nu(h)$ is not atomic and $\widehat{\lambda}(s_0) = i$ then
  - $h^{\surd} = \mathbf{turn}_i \wedge (\bigwedge_{a_j \in moves(s_0)} (\langle a_j \rangle \top \wedge [a_j] h^{\surd}_{a_j}))$.
  - $head^{\surd}_h = \mathbf{turn}_i \wedge (\bigwedge_{a_j \in moves(s_0)} \langle a_j \rangle \top)$.

Due to the ability to interleave choices of players, we also need to define for a composite game expression g, the initial (atomic) game of g and the game expression generated after playing the initial atomic game (or in other words the residue). We make this notion precise below:

Definition of init

- $init(h) = \{h\}$ for $h \in \mathrm{G}$
- $init(g_1; g_2) = init(g_1)$ if $g_1 \neq \epsilon$ else $init(g_2)$.
- $init(g_1 \cup g_2) = init(g_1) \cup init(g_2)$.
- $init(g_1 | g_2) = init(g_1) \cup init(g_2)$.

Definition of residue

- $h\backslash h = \epsilon$ and $\epsilon\backslash h = \epsilon$.
- $(g_1; g_2)\backslash h = \begin{cases} (g_1\backslash h); g_2 & \text{if } g_1 \neq \epsilon. \\ (g_2\backslash h) & \text{otherwise.} \end{cases}$
- $(g_1 \cup g_2)\backslash h = \begin{cases} (g_1\backslash h) \cup (g_2\backslash h) & \text{if } h \in \mathit{init}(g_1) \text{ and } h \in \mathit{init}(g_2). \\ g_1\backslash h & \text{if } h \in \mathit{init}(g_1) \text{ and } h \notin \mathit{init}(g_2). \\ g_2\backslash h & \text{if } h \in \mathit{init}(g_2) \text{ and } h \notin \mathit{init}(g_1). \end{cases}$
- $(g_1|g_2)\backslash h = \begin{cases} (g_1\backslash h|g_2) \cup (g_1|g_2\backslash h) & \text{if } h \in \mathit{init}(g_1) \text{ and } h \in \mathit{init}(g_2). \\ (g_1\backslash h|g_2) & \text{if } h \in \mathit{init}(g_1) \text{ and } h \notin \mathit{init}(g_2). \\ (g_1|g_2\backslash h) & \text{if } h \in \mathit{init}(g_2) \text{ and } h \notin \mathit{init}(g_1). \end{cases}$

The translation used to express the property of enabling of trees in terms of standard PDL formulas also suggest that the techniques developed for proving completeness of PDL can be applied in the current setting. We base our axiomatization of the logic on the "reduction axioms" methodology of dynamic logic. The most interesting reduction axiom in our setting would naturally involve the parallel composition operator. Intuitively, for game expressions g_1, g_2, a formula α and a player $i \in N$ the reduction axiom for $\langle g_1|g_2, i\rangle\alpha$ need to express the following properties:

- There exists an atomic tree $h \in \mathit{init}(g_1|g_2)$ such that $\mathit{head}(\nu(h))$ is enabled.
- Player i has a strategy in $\mathit{head}(\nu(h))$ which when composed with a strategy in the residue ensures α. We use $\mathit{comp}^i(h, g_1, g_2, \alpha)$ to denote this property and formally define it inductively as follows:

Suppose $h = (S, \Rightarrow, s_0, \widehat{\lambda})$ where $A = \mathit{moves}(s_0) = \{a_1, \ldots, a_k\}$.

- If $h \in \mathit{init}(g_1)$, $h \in \mathit{init}(g_2)$ and
  - $\widehat{\lambda}(s_0) = i$ then $\mathit{comp}^i(h, g_1, g_2, \alpha) = \bigvee_{a_j \in A}(\langle a_j\rangle\langle(h_{a_j}; (g_1\backslash h))|g_2\rangle\alpha \vee \langle a_j\rangle\langle g_1|(h_{a_j}; (g_2\backslash h))\rangle\alpha)$.
  - $\widehat{\lambda}(s_0) \in \bar{i}$ then $\mathit{comp}^i(h, g_1, g_2, \alpha) = \bigwedge_{a_j \in A}([a_j]\langle(h_{a_j}; (g_1\backslash h))|g_2\rangle\alpha \vee [a_j]\langle g_1|(h_{a_j}; (g_2\backslash h))\rangle\alpha)$.

- If $h \in init(g_1)$, $h \notin init(g_2)$ and
  - $\widehat{\lambda}(s_0) = i$ then $comp^i(h, g_1, g_2, \alpha) = \bigvee_{a_j \in A}(\langle a_j \rangle \langle (h_{a_j}; (g_1 \backslash h)) | g_2 \rangle \alpha)$.
  - $\widehat{\lambda}(s_0) \in \bar{i}$ then $comp^i(h, g_1, g_2, \alpha) = \bigwedge_{a_j \in A}([a_j] \langle (h_{a_j}; (g_1 \backslash h)) | g_2 \rangle \alpha)$.
- if $h \in init(g_2)$, $h \notin init(g_1)$ and
  - $\widehat{\lambda}(s_0) = i$ then $comp^i(h, g_1, g_2, \alpha) = \bigvee_{a_j \in A}(\langle a_j \rangle \langle g_1 | (h_{a_j}; (g_2 \backslash h)) \rangle \alpha)$.
  - $\widehat{\lambda}(s_0) \in \bar{i}$ then $comp^i(h, g_1, g_2, \alpha) = \bigwedge_{a_j \in A}([a_j] \langle g_1 | (h_{a_j}; (g_2 \backslash h)) \rangle \alpha)$.

Note that the semantics for parallel composition allows us to interleave subtrees of g_2 within g_1 (and vice versa). Therefore in the definition of $comp^i$ at each stage after an action a_j, it is important to perform the sequential composition of the subtree h_{a_j} with the residue of the game expression.

The axiom schemes

A*1* Propositional axioms:

(a) All the substitutional instances of tautologies of PC.

(b) $\mathbf{turn}_i \equiv \bigwedge_{j \in \bar{i}} \neg \mathbf{turn}_j$.

A*2* Axiom for single edge games:

(a) $\langle a \rangle(\alpha_1 \vee \alpha_2) \equiv \langle a \rangle \alpha_1 \vee \langle a \rangle \alpha_2$.

(b) $\langle a \rangle \mathbf{turn}_i \supset [a] \mathbf{turn}_i$.

A*3* Dynamic logic axioms:

(a) $\langle g_1 \cup g_2, i \rangle \alpha \equiv \langle g_1, i \rangle \alpha \vee \langle g_2, i \rangle \alpha$.

(b) $\langle g_1; g_2, i \rangle \alpha \equiv \langle g_1, i \rangle \langle g_2, i \rangle \alpha$.

(c) $\langle g_1 | g_2, i \rangle \alpha \equiv \bigvee_{h \in init(g_1 | g_2)} head_h^{\vee} \wedge comp^i(h, g_1, g_2, \alpha)$.

A*4* $\langle h, i \rangle \alpha \equiv h^{\vee} \wedge \downarrow_{(h,i,\alpha)}$.

For $h \in \mathbb{H}$ with $\nu(h) = T = (S, \Rightarrow, s_0, \widehat{\lambda})$ we define $\downarrow_{(h,i,\alpha)}$ as follow:

- $\downarrow_{(h,i,\alpha)} = \begin{cases} \alpha & \text{if } moves(s_0) = \emptyset. \\ \bigvee_{a\in\Sigma} \langle a\rangle\langle h_a, i\rangle\alpha & \text{if } moves(s_0) \neq \emptyset \text{ and } \widehat{\lambda}(s_0) = i. \\ \bigwedge_{a\in\Sigma} [a]\langle h_a, i\rangle\alpha & \text{if } moves(s_0) \neq \emptyset \text{ and } \widehat{\lambda}(s_0) \in \bar{i}. \end{cases}$

Inference rules

$$(MP)\ \frac{\alpha,\ \alpha \supset \beta}{\beta} \qquad (NG)\ \frac{\alpha}{[a]\alpha}$$

Axioms (A1) and (A2) are self explanatory. Axiom (A3) constitutes the reduction axioms for the compositional operators. Note that unlike in PDL sequential composition in our setting corresponds to composition over trees. The following proposition shows that the usual reduction axiom for sequential composition remains valid.

Proposition 1. *The formula* $\langle g_1; g_2, i\rangle\alpha \equiv \langle g_1, i\rangle\langle g_2, i\rangle\alpha$ *is valid.*

Proof. Suppose $\langle g_1; g_2, i\rangle\alpha \supset \langle g_1, i\rangle\langle g_2, i\rangle\alpha$ is not valid. This means there exists a model M and a state u such that $M, u \models \langle g_1; g_2, i\rangle\alpha$ and $M, u \not\models \langle g_1, i\rangle\langle g_2, i\rangle\alpha$. From semantics we get $\exists X \in R^i_{g_1;g_2}$ such that X is a valid tree, $root(X) = u$ and for all $w \in frontier(X)$ we have $M, u \models \alpha$. By definition, X is of the form $\mathfrak{f}(\mathfrak{T}(Y); \mathcal{T})$ where $Y \in R^i_{g_1}$ and $\mathcal{T} = \{\mathfrak{T}(X_1), \ldots, \mathfrak{T}(X_k)\}$ with $\{X_1, \ldots, X_k\} \subseteq R^i_{g_2}\}$. Since X is a valid tree we have $Y, X_1, \ldots, X_k$ are valid trees. Thus we get that for all $j : 1 \leq j \leq k$, $M, root(X_j) \models \langle \xi_2, i\rangle\alpha$ and from semantics we have $M, u \models \langle g_1, i\rangle\langle g_2, i\rangle\alpha$ which gives the required contradiction.

A similar argument which makes use of the definition of R^i_g and the semantics shows that $\langle g_1, i\rangle\langle g_2, i\rangle\alpha \supset \langle g_1; g_2, i\rangle\alpha$ is valid.

□

5.1 Completeness

To show completeness, we prove that every consistent formula is satisfiable. Let α_0 be a consistent formula, and $CL(\alpha_0)$ denote the subformula closure of α_0. In addition to the usual subformula closure we also require the following: if $\langle h, i\rangle\alpha \in CL(\alpha_0)$ then $g^{\surd}, \downarrow_{(h,i,\alpha)} \in CL(\alpha_0)$ and if $\langle g_1 | g_2, i\rangle\alpha \in CL(\alpha_0)$ then $\bigwedge_{h\in init(g_1|g_2)} head^{\surd}_h, comp^i(h, g_1, g_2, \alpha) \in CL(\alpha_0)$.

Let $AT(\alpha_0)$ be the set of all maximal consistent subsets of $CL(\alpha_0)$, referred to as atoms. We use u, w to range over the set of atoms. Each $u \in AT(\alpha_0)$ is a finite set of formulas, we denote the conjunction of all formulas in u by $\widehat{u}$. For a nonempty subset $X \subseteq AT(\alpha_0)$, we denote by $\widetilde{X}$ the disjunction of all $\widehat{u}, u \in X$. Define a transition relation on $AT(\alpha_0)$ as follows: $u \xrightarrow{a} w$ iff $\widehat{u} \wedge \langle a \rangle \widehat{w}$ is consistent. Let the model $M = (W, \longrightarrow, V)$ where $W = AT(\alpha_0)$ and the valuation function V is defined as $V(w) = \{p \in P \mid p \in w\}$. Once the model is defined, the semantics (given earlier) specifies relation R^i_g. The following lemma asserts the consistency condition on elements of R^i_g.

Lemma 1. *For all $i \in N$, for all $h \in \mathbb{H}$, for all $X \subseteq (W \times W)^*$ with $\mathcal{X} = frontier(X)$, for all $u \in W$ the following holds:*

1. *if X is a valid tree with $root(X) = u$ and $X \in R^i_h$ then $\widehat{u} \wedge \langle h, i \rangle \widetilde{\mathcal{X}}$ is consistent.*
2. *if $\widehat{u} \wedge \langle h, i \rangle \widetilde{\mathcal{X}}$ is consistent then there exists a X' which is a valid tree with $frontier(X') \subseteq \mathcal{X}$ and $root(X') = u$ such that $X' \in R^i_h$.*

Proof. A detailed proof is given in the appendix. It essentially involves showing that the game h is enabled at the state u and that there is a strategy for player i in $T_u \restriction h$ represented by the tree X whose frontier nodes are $\mathcal{X}$. The strategy tree X is constructed in stages starting at u. For any path of the partially constructed strategy tree if the paths ends in a position of player i then the path is extended by guessing a unique outgoing edge. If the position belongs to a player in $\bar{i}$ then all edges are taken into account. □

Lemma 2. *For all $i \in N$, for all $g \in \Gamma$, for all $X \subseteq (W \times W)^*$ with $\mathcal{X} = frontier(X)$ and $u \in W$, if $\widehat{u} \wedge \langle h, i \rangle \widetilde{\mathcal{X}}$ is consistent then there exists X' which is a valid tree with $frontier(X') \subseteq \mathcal{X}$ and $root(X') = u$ such that $X' \in R^i_h$.*

Proof is given in the appendix.

Lemma 3. *For all $\langle g, i \rangle \alpha \in CL(\alpha_0)$, for all $u \in W$, $\widehat{u} \wedge \langle g, i \rangle \alpha$ is consistent iff there exists $X \in R^i_g$ which is a valid tree with $root(X) = u$ such that $\forall w \in frontier(X)$, $\alpha \in w$.*

Proof. ($\Rightarrow$) Follows from lemma 2.

($\Leftarrow$) Suppose there exists $X \in R^i_g$ which is a valid tree with $root(X) = u$ such that $\forall w \in frontier(X)$, $\alpha \in w$. We need to show that $\widehat{u} \wedge \langle g, i \rangle \alpha$ is consistent, this is done by induction on the structure of g.

- The case when $g = h$ follows from lemma 1. For $g = g_1 \cup g_2$ the result follows from axiom (A3a).

- $g = g_1; g_2$: Since $X \in R^i_{g_1;g_2}$, $\exists Y$ with $root(Y) = u$ and $frontier(Y) = \{v_2, \ldots, v_k\}$, there exist sets $X_1, \ldots, X_k$ where for all $j : 1 \leq j \leq k$, $root(X_j) = v_j$, $\bigcup_{j=1,\ldots,k} frontier(X_j) = frontier(X)$, $X_j \in R^i_{g_2}$ and $Y \in R^i_{g_1}$. By induction hypothesis, for all j, $\widehat{v_j} \wedge \langle g_2 \rangle \alpha$ is consistent. Since v_j is an atom and $\langle g_2, i \rangle \alpha \in CL(\alpha_0)$, we get $\langle g_2, i \rangle \alpha \in v_j$. Again by induction hypothesis we have $\widehat{u} \wedge \langle g_1, i \rangle \langle g_2, i \rangle \alpha$ is consistent. Hence from (A3b) we have $\widehat{u} \wedge \langle g_1; g_2, i \rangle \alpha$ is consistent.

- $g = g_1 | g_2$: Let $h \in init(g_1 | g_2)$, and $h = (S, \Rightarrow, s_0, \widehat{\lambda})$. We have three cases depending on whether h is the initial constituent game in g_1 and g_2. We look at the case when $h \in init(g_1)$ and $h \notin init(g_2)$, the arguments for the remaining cases are similar. Let $A = moves(s_0) = \{a_1, \ldots, a_k\}$. By semantics, since $enabled(h, u)$ holds we have $moves(u) = A$. We also get there exists $Y_j \in R^i_{t_{a_j};(g_1 \backslash h)|g_2}$ where $\bigcup_{j=1,\ldots,k} frontier(Y_j) = frontier(X)$. Suppose $\widehat{\lambda}(s_0) = \bar{\imath}$, by performing a second induction on the depth of X we can argue that $\widehat{u} \wedge (\bigwedge_{a_j \in A}([a_j]\langle (t_{a_j}; (g_1 \backslash h)) | g_2 \rangle \alpha)$ is consistent. Therefore from axiom (A3c) we have $\widehat{u} \wedge \langle g_1 | g_2 \rangle \alpha$ is consistent.

□

This leads us to the following theorem from which we can deduce the completeness of the axiom system.

Theorem 1. *For all formulas α_0, if α_0 is consistent then α_0 is satisfiable.*

Dedidability: Given a formula α_0, let $\mathfrak{H}(\alpha_0)$ be the set of all atomic game terms appearing in α_0. Let $\mathfrak{T}(\alpha_0) = \{v(h) \mid h \in \mathfrak{H}(\alpha_0)\}$ and $\mathfrak{m} = \max_{T \in \mathfrak{T}(\alpha_0)} |T|$. For any finite tree T, we define $|T|$ to be the number of vertices and edges in T. It can be verified that $|CL(\alpha_0)|$ is linear in $|\alpha_0|$ and therefore we have $|AT(\alpha_0)| = O(2^{|\alpha_0|})$. The states of the model M constitutes atoms of α_0 and therefore we get that if α_0 is satisfiable then there is a model whose size is at most exponential in $|\alpha_0|$. The relation R^i_g can be explicitly constructed in time $O(2^{|M|^{\mathfrak{m}}})$. Thus we get the following corollary.

Corollary 1. *The satisfiability problem for the logic is decidable.*

6 Discussion

Iteration

An obvious extension of the logic is to add an operator for (unbounded) iteration of sequential composition. The semantics is slightly more complicated since we are dealing with trees. One needs to define it in terms of a least fixed point operator (as seen in Parikh (1985)). Under this interpretation, the standard dynamic logic axiom for iteration remains valid: $\langle g^*, i\rangle\alpha \equiv \alpha \vee \langle g, i\rangle\langle g^*, i\rangle\alpha$.

We also have the familiar induction rule for dynamic logic which asserts that when α is invariant under g so it is with the iteration of g.

$$(IND)\quad \frac{\langle g, i\rangle\alpha \supset \alpha}{\langle g^*, i\rangle\alpha \supset \alpha}$$

Note that the completeness proof (in the presence of interleaving) gets considerably more complicated now. Firstly, the complexity of $g\backslash h$ is no longer less than that of g so we cannot apply induction directly for parallel composition. In general when we consider $g_1^* \| g_2^*$, the interleaving critically depends on how many iterations are chosen in each of the components. The technique is to consider a graph for every g as follows: add an edge labelled h from g to $g\backslash h$. This is a finite graph, and we can show that the enabling of g at a state s corresponds to the existence of an embedding of this graph at s. In effect, the unfolding of the parallel composition axiom asserts the existence of this subgraph, and the rest of the proof uses the induction rule as in the completeness proof for dynamic logic. We omit the detailed proof here since it is technical and lengthy.

Strategy specifications

Throughout the paper we have been talking of existence of strategies in compositional games. It would be more interesting to specify strategies explicitly in terms of their properties as done in Ramanujam and Simon (2008). In the presence of parallel composition, this adds more value to the analysis since apart from specifying structural conditions which ensures the ability for players to copy moves, we can also specify the exact sequence of moves which are copied across games. The basic techniques used here can be extended to deal with

strategy specification. However, it would be more interesting to come up with compositional operators for strategy specifications which can naturally exploit the interleaving semantics.

Acknowledgements We thank the anonymous referees for their valuable comments and suggestions. The second author thanks the Netherlands Institute for Advanced Study in the Humanities and Social Sciences for its support.

A Appendix

Lemma 1. For all $i \in N$, for all $h \in \mathbb{H}$, for all $X \subseteq (W \times W)^*$ with $\mathcal{X} = \mathit{frontier}(X)$, for all $u \in W$ the following holds:

1. if X is a valid tree with $\mathit{root}(X) = u$ and $X \in R^i_h$ then $\widehat{u} \wedge \langle h, i\rangle\widetilde{\mathcal{X}}$ is consistent.
2. if $\widehat{u} \wedge \langle h, i\rangle\widetilde{\mathcal{X}}$ is consistent then there exists a X' which is a valid tree with $\mathit{frontier}(X') \subseteq \mathcal{X}$ and $\mathit{root}(X') = u$ such that $X' \in R^i_h$.

Proof. Let $h = (S, \Rightarrow, s_0, \widehat{\lambda})$. If $\mathit{moves}(s_0) = \emptyset$ then from axiom (A4) we get $\langle h, i\rangle\alpha \equiv \beta \wedge \alpha$ and the lemma holds. Let $\mathit{moves}(s_0) = \{a_1, \ldots, a_k\}$ and $\widehat{\lambda}(s_0) = i$.

Suppose $X \in R^i_h$, since X is a valid tree and $\mathit{enabled}(\mathit{head}(h), u)$ holds, there exist sets $Y_1, \ldots, Y_k$ such that for all $j : 1 \leq j \leq k$, $w_j = \mathit{root}(Y_j)$ and $u \xrightarrow{a_j} w_j$. Since u is an i node we have that the strategy should choose a w_j such that $u \xrightarrow{a_j} w_j$ and $X' \in R^i_{h_{a_j}}$ where $X = (u, w_j) \cdot X'$. By induction hypothesis we have $\widehat{w_j} \wedge \langle h_{a_j}, i\rangle\widetilde{\mathcal{X}}$ is consistent. Hence from axiom (A4) we conclude $\widehat{u} \wedge \langle h, i\rangle\widetilde{\mathcal{X}}$ is consistent.

Suppose $\widehat{u} \wedge \langle h, i\rangle\widetilde{\mathcal{X}}$ is consistent. From axiom (A4) it follows that there exists $w_1, \ldots, w_k$ such that for all $j : 1 \leq j \leq k$, we have $u \xrightarrow{a_j} w_j$ and hence $\mathit{enabled}(h, u)$ holds. Let $\mathcal{X} = \{v_1, \ldots, v_m\}$, from axiom (A4) we have $\widehat{u} \wedge (\bigvee_{a\in\Sigma} \langle a\rangle\langle h_a, i\rangle\widetilde{\mathcal{X}})$ is consistent. Hence we get that there exists w_j such that $u \xrightarrow{a_j} w_j$ and $\widehat{w_j} \wedge \langle h_a, i\rangle\widetilde{\mathcal{X}}$ is consistent. By induction hypothesis there exists X' which is a valid tree

with $frontier(X') \subseteq \mathcal{X}$, $root(X') = w_j$ and $X' \in R^i_{h_a}$. By definition of R^i we get $(u, w_j) \cdot X' \in R^i_h$.

Let $\widehat{\lambda}(s_0) = \bar{\imath}$ and suppose $X \in R^i_h$. Since $enabled(head(h), u)$ holds and X is a valid tree, there exist sets $Y_1, \ldots, Y_k$ such that for all $j : 1 \leq j \leq k$, $w_j = root(Y_j)$ and $u \xrightarrow{a_j} w_j$. Since u is an $\bar{\imath}$ node, any strategy of i need to have all the branches at u (by definition of strategy). Thus we get: for all w_j with $u \xrightarrow{a_j} w_j$, there exists X_j with $root(X_j) = w_j$ such that $X_j \in R^i_h$ and $X = \bigcup_{j=1,\ldots,k}(u, w_j) \cdot X_j$. By induction hypothesis and the fact that $\mathcal{X}_j = frontier(X_j) \subseteq \mathcal{X}$, we have $\widehat{w}_j \wedge \langle h, i\rangle\widetilde{\mathcal{X}}$ is consistent. Hence from axiom (A4) we get $\widehat{u} \wedge \langle h, i\rangle\widetilde{\mathcal{X}}$ is consistent.

Likewise, using axiom (A4) we can show that if $\widehat{u} \wedge \langle h, i\rangle\widetilde{\mathcal{X}}$ is consistent then there exists a X' which is a valid tree with $frontier(X') \subseteq \mathcal{X}$ and $root(X') = u$ such that $X' \in R^i_h$. □

Lemma 2. For all $i \in N$, for all $g \in \Gamma$, for all $X \subseteq (W \times W)^*$ with $\mathcal{X} = frontier(X)$ and $u \in W$, if $\widehat{u} \wedge \langle h, i\rangle\widetilde{\mathcal{X}}$ is consistent then there exists X' which is a valid tree with $frontier(X') \subseteq \mathcal{X}$ and $root(X') = u$ such that $X' \in R^i_h$.

Proof. By induction on the structure of g.

- $g = h$: The claim follows from Lemma 1 item 2.
- $g = g_1 \cup g_2$: By axiom (A3a) we get $\widehat{u} \wedge \langle g_1, i\rangle\widetilde{\mathcal{X}}$ is consistent or $\widehat{u} \wedge \langle g_2, i\rangle\widetilde{\mathcal{X}}$ is consistent. By induction hypothesis there exists X_1 which is a valid tree with $frontier(X_1) \subseteq \mathcal{X}$ and $root(X_1) = u$ such that $(u, X_1) \in R^i_h$ or there exists X_2 which is a valid tree with $frontier(X_2) \subseteq \mathcal{X}$ and $root(X_2) = u$ such that $X_2 \in R^i_h$. Hence we have $X_1 \in R^i_{g_1 \cup g_2}$ or $X_2 \in R^i_{g_1 \cup g_2}$.
- $g = g_1; g_2$: By axiom (A3b), $\widehat{u} \wedge \langle g_1, i\rangle\langle g_2, i\rangle\widetilde{\mathcal{X}}$ is consistent. Hence $\widehat{u} \wedge \langle g_1, i\rangle(\bigvee(\widehat{w} \wedge \langle g_2, i\rangle\widetilde{\mathcal{X}}))$ is consistent, where the join is taken over all $w \in \mathcal{Y} = \{w \mid w \wedge \langle g_2, i\rangle\widetilde{\mathcal{X}} \text{ is consistent }\}$. So $\widehat{u} \wedge \langle g_1, i\rangle\widetilde{\mathcal{Y}}$ is consistent. By induction hypothesis, there exists Y' which is a valid tree with $\mathcal{Y}' = frontier(Y') \subseteq \mathcal{Y}$ and $root(Y') = u$ such that $(u, Y') \in R^i_{g_1}$. We also have that for all $w \in \mathcal{Y}$, $\widehat{w} \wedge \langle g_2, i\rangle\widetilde{\mathcal{X}}$ is consistent. Therefore we get for all $w_j \in \mathcal{Y}' = \{w_1, \ldots, w_k\}$, $\widehat{w}_j \wedge \langle g_2, i\rangle\widetilde{\mathcal{X}}$ is consistent. By induction hypothesis, there exists X_j which

is a valid tree with $\mathcal{X}_j = \mathit{frontier}(X_j) \subseteq \mathcal{X}$ and $\mathit{root}(\mathcal{X}_j) = w_j$ such that $X_j \in R^i_{g_2}$. Let X' be the tree in $Y'; \{X_j \mid j = 1, \ldots, k\}$ obtained by pasting X_j to the leaf node w_j in Y'. We get $X' \in R^i_{g_1;g_2}$.

- $g = g_1|g_2$: Note that for all $g \in \Gamma$ and $h \in \mathit{head}(g)$, the complexity of $g\backslash h$ is less than that of g. Therefore by making use of axiom (A3c) we can show that there exists X' with $\mathit{frontier}(X') \subseteq \mathcal{X}'$ and $\mathit{root}(X') = u$ such that $X' \in R^i_h$.

□

References

K. Abrahamson. *Decidability and expressiveness of logics of processes*. PhD thesis, Dept. of Computer Science, Univ. of Washington, 1980.

T. Ågotnes. Action and knowledge in alternating time temporal logic. *Synthese*, 149(2):377–409, 2006.

R. Alur, T. A. Henzinger, and O. Kupferman. Alternating-time temporal logic. *Journal of the ACM*, 49:672–713, 2002.

J. Broersen. CTL.STIT: Enhancing ATL to express important multi-agent system verification properties. In *Proceedings of AAMAS-2010*. ACM Press, 2010.

J. Broersen, A. Herzig, and N. Troquard. Embedding Alternating-time Temporal Logic in strategic STIT logic of agency. *Journal of Logic and Computation*, 16(5):559–578, 2006.

R. Danecki. Nondeterministic propositional dynamic logic with intersection is decidable. In *Proc. 5th Symposium in Computation Theory*, Lecture Notes in Computer Science, pages 34–53. Springer, 1984.

V. Goranko. Coalition games and alternating temporal logics. In *Proceedings of TARK-2001*, pages 259–272, 2001.

D. Harel. Dynamic logic. *Handbook of Philosophical Logic*, 2:496–604, 1984.

D. Harel, D. Kozen, and R. Parikh. Process logic: Expressiveness, decidability, completeness. *Journal of Computer and System Sciences*, 25(2):144–170, 1982.

J. Horty. *Agency and Deontic Logic*. Oxford University Press, 2001.

M. Lange and C. Lutz. 2-EXPTIME lower bounds for propositional dynamic logics with intersection. *Journal of Symbolic Logic*, 70(4):1072–1086, 2005.

R. Parikh. The logic of games and its applications. *Annals of Discrete Mathematics*, 24:111–140, 1985.

M. Pauly. *Logic for Social Software*. PhD thesis, Univ. of Amsterdam, 2001.

D. Peleg. Concurrent dynamic logic. *Journal of the ACM*, 34(2):450–479, 1987.

R. Ramanujam and S. Simon. Dynamic logic on games with structured strategies. In *Proceedings of KR-08*, pages 49–58. AAAI Press, 2008.

J. van Benthem. Extensive games as process models. *Journal of Logic Language and Information*, 11:289–313, 2002.

J. van Benthem, S. Ghosh, and F. Liu. Modelling simultaneous games with dynamic logic. *Synthese (Knowledge, Rationality and Action)*, 165:247–268, 2008.

W. van der Hoek, W. Jamroga, and M. Wooldridge. A logic for strategic reasoning. *Proceedings of AAMAS-2005*, pages 157–164, 2005.

D. Walther, W. van der Hoek, and M. Wooldridge. Alternating-time temporal logic with explicit strategies. In *Proceedings of TARK-2007*, pages 269–278, 2007.

Short Sight in Extensive Games

Davide Grossi and Paolo Turrini

University of Liverpool, University of Luxembourg
d.grossi@liverpool.ac.uk, paolo.turrini@uni.lu

Abstract

The paper introduces a class of games in extensive form where players take strategic decisions while not having access to the terminal histories of the game, hence being unable to solve it by standard backward induction. This class of games is studied along two directions: first, by providing an appropriate refinement of the subgame perfect equilibrium concept, a corresponding extension of the backward induction algorithm and an equilibrium existence theorem; second, by showing that these games are a well-behaved subclass of a class of games with possibly unaware players recently studied in the literature.[1]

1 Introduction

In the past decade the multi agent systems (MAS) community has witnessed several attempts to relax the strong assumptions underpinning game-

[1] This version slightly extends substantially similar versions of this paper which have appeared in the proceedings of AAMAS 2012 and as a technical report of the Computer Science Department of the University of Liverpool (ULCS-11-005).

theoretical models, such as common knowledge of the game structure, logical omniscence and unbounded computational power, to mention a few. Along these lines Joseph Halpern's invited talk at AAMAS 2011—*Beyond Nash-Equilibrium: Solution Concepts for the 21st Century*[2]—highlighted several research challenges that arise when attempting to provide more realistic versions of the Nash equilibrium solution concept. Among those challenges, the issue of unawareness seems to stand out, viz. the observation that in real games, like for instance chess, players take decisions even if they cannot possibly have access to the whole game form. Halpern himself extensively contributed to the research on players' unawareness: in Halpern and Rêgo (2006) and its extension Halpern and Rêgo (2007), a game-theoretical analysis of unawareness in extensive games is presented, where players have access to only part of the terminal histories of a game tree as they ignore, at some nodes, some of the actions available to their fellow players. The same phenomenon has been studied, although by different means, by Yossi Feinberg in Feinberg (2004; 2005).

All the aforementioned models of unawareness in games make a common assumption: players might be unaware of some branches of the game tree, but they do have access to a subset of the terminal histories, that is, they have a full representation of at least some possible endings of the game. With the present work we would like to push Halpern's stance further, by lifting this assumption and present a model of players who not only might not see a part of the terminal nodes of a game tree but who might not even see any such nodes. As happens in real games like chess, but also in a number of occasions where individuals are confronted with a large game structure, decisions are taken on the basis of a stepwise evaluation of foreseeable intermediate positions. As the game proceeds, it often reveals earlier decisions to be wrong.[3] The following example provides a concrete motivating scenario representing this special kind of unawareness, which we will be calling *short sight*.

Example 1 (A chess scenario)**.** In Figure 1 Black is to move. He has three options at his disposal: moving the black king to g7 (shortly ♔g7), moving it to e7 (♔e7), or moving the pawn one square further to h2 (h2). Let us assume that Black has to move under pressing time constraints or that he is not well-versed in evaluating key positions on the chessboard. He will then take into consideration only a few possible developments of the play—for instance

[2]Published in Halpern (2011).

[3]To say it with Watson (1998), "Chess is a draw that is only made competitive by human error" .

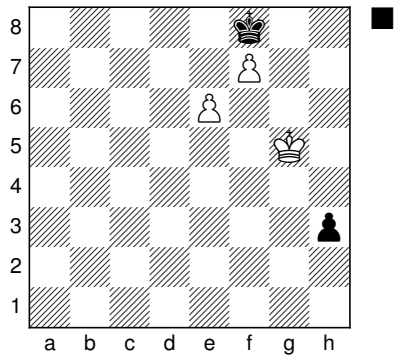


Figure 1: Black to move

what he would be able to reach in two moves (i.e., some plays up to two steps ahead)—and he will base his decisions on somewhat 'coarse' evaluations—for instance, gaining material advantage.

If this is the case, in a situation such as the one displayed in Figure 1, he will prefer to queen his pawn as quickly as possible.[4] Comparing the moves ♔g7, ♔e7 and h2, the latter clearly leads to material advantage while the formers do not. So Black will go for h2. However after h2 White can move its king to f6 (♔f6). Now Black is in trouble because after the white king is in f6 Black has only one move at its disposal — he must queen his pawn (h1) — as ♔e7 and ♔g7 are now illegal. Black's material advantage in the resulting position (one queen against two pawns) is no consolation: after e7 Black is checkmated.

In the example Black loses for two reasons: 1) he has partial view even on the immediate development of the game; 2) he bases his decision on an evaluation criterion—reaching material advantage—which turns out to be counterproductive. These observations exemplify the characteristics of short sight in extensive games: 1) players may be aware of only part of the game structure and may not be able to calculate the consequences of their actions up to the terminal nodes; 2) at each choice point, players base their decisions evaluating the positions they can foresee according to (possibly faulty) criteria. The paper

[4] The black pawn reaching h1 can be queened, i.e. turned into a strong major piece, giving its owner an often decisive advantage.

will incorporate the characteristic features of short sight in a standard treatment of extensive games, studying their properties and their relation with models of players' unawareness to be found in the literature—in particular the ones in Halpern and Rêgo (2006).

Outline of the paper. Section 2 introduces the basic terminology and facts to be used later on in the paper. It mainly concernes the notion of extensive game and preference relation and it presents standard solution concepts, such as the subgame perfect equilibrium. Section 3 equips extensive games with a description of players' limited view at each history and presents corresponding solution concepts for the new models. In particular it defines a backward induction algorithm for games with short sight and proves an equilibrium existence theorem for this class of games. Section 4 discusses the relation between games with short sight and games with awareness as studied by Halpern and Rêgo. Concretely, it shows that games with short sight are a special type of games with awareness. Section 5 concludes the paper pointing to several possible developments.

2 Preliminaries

The section introduces the basic terminology and notation to be used in the rest of the paper.

2.1 Game forms and games

The structures we will be working with are extensive games which, unlike the games in strategic or normal form, take the sequential structure of decisions into account Osborne and Rubinstein (1994). We start out introducing extensive games forms of perfect information (henceforth simply "extensive game forms" or "game forms"), where players have full knowledge of the possible courses of events. The following definition is adapted from Osborne and Rubinstein (1994).

Definition 2.1 (Extensive game forms)**.** An *extensive game form* is a tuple $\mathcal{G} = (N, H, \mathsf{t}, \Sigma_i, o)$ where:

- N is a non-empty set of players;

- H is a non-empty set of sequences, called *histories*, such that:
  - The empty sequence $\emptyset$ is a member of H;
  - If $(a^k)_{k=1,\dots,K} \in H$ and $L < K$ then $(a^k)_{k=1,\dots,L} \in H$;
  - If an infinite sequence $(a^k)_{k=1}^{\omega}$ is such that $(a^k)_{k=1,\dots,L} \in H$ for every $L < \omega = |\mathbb{N}|$ then $(a^k)_{k=1}^{\omega} \in H$;

  A history $h \in H$ is called *terminal* if it is infinite or it is of the form $(a^k)_{k=1,\dots,K}$ with $K < \omega$ and there is no a^{K+1} such that $(a^k)_{k=1,\dots,K+1} \in H$. The set of terminal histories is denoted Z. Each component of a history is called an *action*. The set of all actions is denoted A. The set of actions following a history h is denoted with $A(h)$. Formally $A(h) = \{a \mid (h, a) \in H\}$. If h is a prefix of h' we write $h \lhd h'$.

- $\mathrm{t} : H \backslash Z \to N$ is a function, called *turn function*, assigning players to non-terminal histories, with the idea that player i moves at history h whenever $\mathrm{t}(h) = i$;

- Σ_i is a non-empty set of strategies $\sigma_i : \{h \in H \backslash Z \mid \mathrm{t}(h) = i\} \to A$ for each player i that assign an action to any non-terminal history whose turn to play is i's; we refer to $\sigma_{\mathrm{t}(h)}(h)$ as the action *prescribed* by strategy σ at history h for the player who moves at h;

- $o : \prod_{i \in N} \Sigma_i \to Z$ is a bijective *outcome* function from the set of strategy profiles to the set of terminal histories.

For any set of histories $A \subseteq H$ we denote $l(A)$ the length of its longest history. The notation can also be used with game forms, where $l(\mathcal{G}) = l(H)$, for H being the set of histories of game form $\mathcal{G}$. If H is a finite set $\mathcal{G}$ is called a *finite* game form. Extensive game forms equipped with preference relations, i.e. a family of orders on terminal histories for each player, are referred to as extensive games (or simply as games).

Definition 2.2 (Extensive games)**.** An extensive game is a tuple $\mathcal{E} = (\mathcal{G}, \succeq_i)$ where $\mathcal{G}$ is an extensive game form and $\succeq_i \subseteq Z^2$ is a total preorder[5] over Z, for each player i.

An extensive game $\mathcal{E} = (\mathcal{G}, \succeq_i)$ is called *finite* if $\mathcal{G}$ is finite.

[5] I.e., a reflexive, transitive and total binary relation.

2.2 Preferences and evaluation criteria

In Definition 2.2 players' preferences are given by a total preorder over the set of terminal nodes. However situations such as the one described in Example 1 suggest that, in presence of short sight, decisions need to be taken even when terminal nodes are not accessible. For this reason we assume here that players hold preferences about foreseeable intermediate nodes according to general criteria which remain stable throughout the game. The idea is that players are endowed with some kind of 'theory' that allows them to conceptualize and evaluate game positions. For instance, in Example 1 *Black* evaluates the positions that he can calculate according to the general criterion of material advantage.

Priority sequences

To model the intuition above we follow a simple strategy. We take evaluation criteria to consist of preferences defined over properties of game positions, and we take properties to be sets of game positions, i.e., sets of histories.

Definition 2.3 (Priority sequences)**.** Let $\mathcal{G} = (N, H, \mathrm{t}, \Sigma_i, o)$ be an extensive game form. A priority sequence, or P-sequence, for $\mathcal{G}$ is a tuple $P = (\mathcal{H}, >)$ where:

- $\mathcal{H} \subseteq \wp(H)$ and $\mathcal{H}$ is finite, i.e., the set of properties $\mathcal{H}$ is a finite set of sets of histories. Elements of $\mathcal{H}$ are denoted $\mathbf{H}, \mathbf{H}', \ldots$.
- $> \subseteq \mathcal{H}^2$ is a strict linear order[6] on the properties in $\mathcal{H}$. To say that $\mathbf{H}$ is preferred to $\mathbf{H}'$, for $\mathbf{H}, \mathbf{H}' \in \mathcal{H}$, we write: $\mathbf{H} > \mathbf{H}'$.

P-sequences express a fixed priority between a finite set of relevant criteria. In our understanding they represent a general theory that a player can use to assess game positions. P-sequences and their generalisation to graphs have been object of quite some recent studies in the logic of preference, such as Liu (2011) from which Definition 2.3 is adapted. Given a P-sequence, a preference over histories can be derived in a natural way:

Definition 2.4 (Preferences)**.** Let $\mathcal{G} = (N, H, \mathrm{t}, \Sigma_i, o)$ be an extensive game form and $P = (\mathcal{H}, >)$ a P-sequence for $\mathcal{G}$. The preference relation $\succeq^P \subseteq H^2$ over the set

[6] I.e. an irreflexive, transitive, asymmetric and total binary relation.

of histories of $\mathcal{G}$ induced by P is defined as follows:

$$h \succeq^P h' \iff \forall \mathbf{H} \in \mathcal{H} : [\text{ if } h' \in \mathbf{H} \text{ then } h \in \mathbf{H} \\ \text{or } \exists \mathbf{H}' \in \mathcal{H} : [h \in \mathbf{H}' \text{ and } h' \notin \mathbf{H}' \text{ and } \mathbf{H}' \succ \mathbf{H}]]$$

In words, a history h is at least as good as a history h' according to P, if and only if, either all properties occurring in P that are satisfied by h' are also satisfied by h or, if that is not the case and there is some property that h' has but h has not, then there exists some other better property which h satisfies and h' does not. This 'recipe' yields preferences of a standard type:

Fact 1. *Let $\mathcal{G}$ be an extensive game form and $P = (\mathcal{H}, \succ)$ a P-sequence for $\mathcal{G}$. The relation $\succeq^P$ has the following properties:*

1. *It is a total pre-order;*
2. *$\succeq^P$ contains at most $2^{|\mathcal{H}|}$ sets of equally preferred elements.*[7]

Sketch of proof. 1. That $\succeq^P$ is reflexive follows directly from Definition 2.4. Transitivity is established by the following argument: assume $h \succeq^P h'$ and $h' \succeq^P h''$. By Definition 2.4 we have four possible cases: i) all properties satisfied by h' are also satisfied by h and all properties satisfied by h'' are also satisfied by h', hence $h \succeq^P h''$; ii) all properties satisfied by h' are also satisfied by h and for some property $\mathbf{H}$ enjoyed by h'' but not by h' there exists another property $\mathbf{H}'$ such that $\mathbf{H}' \succ \mathbf{H}$ and h' satisfies $\mathbf{H}$ but h'' does not. Hence for some property $\mathbf{H}$ enjoyed by h but not by h'' there exists another property $\mathbf{H}'$ such that $\mathbf{H} \succ \mathbf{H}'$ and h satisfies $\mathbf{H}$ but h'' does not, from which we conclude $h \succeq^P h''$. iii) More schematically, for all $\mathbf{H}$: $\exists \mathbf{H}' \in \mathcal{H} : [h' \in \mathbf{H}' \text{ and } h'' \notin \mathbf{H} \text{ and } \mathbf{H}' \succ \mathbf{H}]]$ and $\forall \mathbf{H} \in \mathcal{H} : [\text{ if } h' \in \mathbf{H} \text{ then } h \in \mathbf{H}]$. The proof is analogous to the one of ii). iv) For all $\mathbf{H}$: $\exists \mathbf{H}' \in \mathcal{H} : [h \in \mathbf{H}' \text{ and } h' \notin \mathbf{H} \text{ and } \mathbf{H}' \succ \mathbf{H}]]$ and $\exists \mathbf{H}'' \in \mathcal{H} : [h' \in \mathbf{H}'' \text{ and } h'' \notin \mathbf{H}'' \text{ and } \mathbf{H}'' \succ \mathbf{H}]]$ follows from the transitivity of relation $\succ$ (Definition 2.3). As for totality, suppose not $h \succeq^P h'$. But then, by totality of $\succ$ (Definition 2.3) $\exists \mathbf{H} \in \mathcal{H} : [h' \in \mathbf{H} \text{ and } h \notin \mathbf{H} \text{ and } \forall \mathbf{H}' \in \mathcal{H} : [h \in \mathbf{H}' \text{ and } h' \notin \mathbf{H}' \text{ implies } \mathbf{H} \succ \mathbf{H}']]$, which implies that $h' \succeq^P h$.

2) Equivalence classes in $\succeq^P$ are determined by the set of properties in $\mathcal{H}$ that they satisfy, hence by elements of $\wp(\mathcal{H})$. As some of these sets might be empty, $2^{|\mathcal{H}|}$ is an upper bound. □

[7] I.e., sets of elements h, h' such that $h \succeq^P h'$ and $h' \succeq^P h$.

Intuitively, P-sequences yield total preorders consisting of a finite set of equally preferred elements which form a linear hierarchy from the set of most preferred elements to the set of least preferred elements. Notice that P-sequences are flexible enough to represent a variety of players' preferences, from natural cases where the most preferred property in the P-sequence contains some of the terminal histories of the game, to cases where all terminal histories of the game are equally "disliked" by not appearing in any of the properties in the P-sequence.

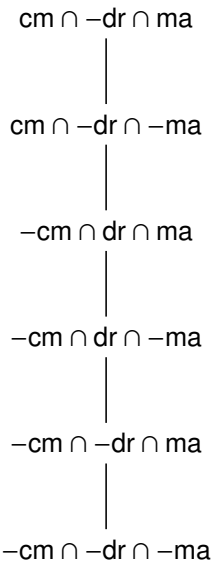
Remark 1 (Linearity of P-sequences and total pre-orders)**.** *The linearity of P-sequences could be viewed as an unrealistic constraint for modeling the players' criteria for assessing positions in a game. On the other hand, allowing for non-linear preference structures would induce, by Definition 2.4, preferences which fail to be total. This would take us out of the realm of (standard) game theory, whose key tenet about players' preferences is that they be total.*

Example 2. As an illustration, recall Example 1. We could model Black's evaluation criteria by the following simple P-sequence (let cm denote the set of histories where White is checkmated, dr the set of histories where the game is a draw, and ma the set of histories where Black has material advantage): $cm \succ \mathsf{dr} \succ \mathsf{ma}$. This P-sequence yields the total preorder over histories depicted at the top of the right page.[8] Here, we have assumed that no history can be a checkmate and a draw at the same time. In words, Black prefers most of all positions where White is checkmated and at the same time he retains material advantage, then positions where White is checkmated without material advantage, and so according to the above P-sequence. The worst positions are the ones where none of the properties occurring in the P-sequence are satisfied.

It is worth observing that the elements of a P-sequence can be represented by set-theoretic compounds of properties [9]. The link to logic should here be evident as sets of histories—our properties—could be seen as denotations of formulae in some logical language (e.g. propositional logic). Our exposition abstracts from the logical aspect which could, however, add a further interesting syntactic dimension to our account.

[8] The total preorder is represented as a Hasse diagram consisting of linearly ordered equivalence classes. Standard set-theoretic notation for inclusion and complementation is used.

[9] As an anonymous reviewer pointed out, there may be situations in which two properties $\mathbf{H}$ and $\mathbf{H'}$ that, when occurring together, outweigh a third one $\mathbf{H''}$, while $\mathbf{H''}$ would be preferred over both $\mathbf{H}$ and $\mathbf{H'}$ when they occur alone (e.g., centre control *together with* an exposed opponent's king may outweigh material disadvantage). In our framework this is handled by stating that $\mathbf{H} \cap \mathbf{H'} \succ \mathbf{H''} \succ \mathbf{H} \cup \mathbf{H'}$.



Games with priorities

Henceforth we will be working with game forms that are endowed with a family of P-sequences, one for each player:

Definition 2.5 (Prioritized games)**.** Let $\mathcal{G}$ be a game form and let P_i be a family of P-sequences for $\mathcal{G}$, one for each player $i \in N$. A *prioritized game* is a tuple $\mathcal{G}^P = (\mathcal{G}, P_i)$.

Clearly, each prioritized game $\mathcal{G}^P = (\mathcal{G}, P_i)$ defines a game in extensive form (Definition 2.2) $\mathcal{E}_{\mathcal{G}^P} = (\mathcal{G}, Z^2 \cap \succeq^{P_i})$. So, when attention is restricted to terminal histories, prioritized games yield standard extensive form games. What they add to the them is information by means of which players can systematically rank non-terminal histories also without having access to terminal histories.

2.3 Subgame-perfect equilibrium

In this section we adapt the notion of subgame-perfect equilibrium to prioritized games. The adaptation is straightforward since each prioritized game univocally determines an extensive one. It is nevertheless worth it to introduce all the notions in details, as they will be our stepping stone for the definition of an analogous solution concept in games with short sight.

We first need to introduce the notion of subgame.

Definition 2.6 (Subgames of prioritized games)**.** Take a finite prioratized game $\mathcal{G}^P = ((N, H, \mathrm{t}, \Sigma_i, o), P_i)$. Its subgame from history h is a prioritized game $\mathcal{G}^P_h = ((N|_h, H|_h, \mathrm{t}|_h, \Sigma_i|_h, o|_h), P_i|_h)$ such that:

- $H|_h$ is the set of sequences h' for which $(h, h') \in H$;
- $\Sigma_i|_h$ is the set of strategies for each player available at h. It consists of elements $\sigma_i|_h$ such that $\sigma_i|_h(h') = \sigma_i(h, h')$ for each $h' \in H|_h$ with $\mathrm{t}(h, h') = i$;
- $\mathrm{t}|_h$ is such that $\mathrm{t}|_h(h') = \mathrm{t}(h, h')$ for each $h' \in H|_h$;
- $o|_h : \prod_{i \in N} \Sigma_i|_h \to Z|_h$ is the outcome function of $\mathcal{G}^P_h$, where $Z|_h$ is the set of sequences h' for which $(h, h') \in Z$;
- $P_i|_h = P_i$.

Now we are ready to introduce subgame perfect equilibria.

Definition 2.7 (Subgame perfect equilibrium)**.** Let $\mathcal{G}^P$ be a finite prioritized game. A strategy profile σ^* is a *subgame perfect equilibrium* if for every player $i \in N$ and every nonterminal history $h \in H \backslash Z$ for which $\mathrm{t}(h) = i$ we have that:

$$o|_h(\sigma^*_i|_h, \sigma^*_{-i}|_h) \succeq^{P_i} o|_h(\sigma_i, \sigma^*_{-i}|_h)$$

for every strategy σ_i available to player i in the subgame $\mathcal{G}^P_h$ that differs from $\sigma^*_i|_h$ only in the action it prescribes after the initial history of $\mathcal{G}^P_h$.

The definition of subgame perfect equilibrium is normally given in its stronger version, without the requirement that σ_i for player i in the subgame $\mathcal{G}^P_h$ differs from $\sigma^*_i|_h$ only in the action it prescribes after the initial history of $\mathcal{G}^P_h$. However the formulation we have given is equivalent to the stronger version for the case of finite games, as proved in (Osborne and Rubinstein 1994, Lemma 98.2). This property of the subgame perfect equilibria is known as *the one deviation property*.

By Kuhn's theorem[10] we can then conclude that all finite prioritized games have at least one subgame perfect equilibrium.

Remark 2. *The existence of subgame perfect equilibria in finite extensive games is usually proven constructively via the well-known backward induction (BI) algorithm. It might be worth recalling that the algorithm solves the game by extending the total preorder on the terminal histories of the game to a total preorder over all histories, where for every player each history is as preferred as the terminal history it leads to under the assumption that the other players play 'rationally'. So the result of the algorithm is a total preorder over all histories consisting of a finite set of equivalence classes, viz. the sort of preference structures also determined by P-sequences (Fact 1). The key difference, however, is that while the order determined by BI is consistent with the order on the terminal nodes, in the sense that keeping on choosing the best option guarantees the best outcome in the game, no such guarantee exist in the order yielded by a P-sequence—as Example 1 neatly shows.*

3 Short sight in games

In this section we introduce and discuss the notion that has motivated the present work: short sight.

3.1 Players' sights

The following definition introduces a simple device to capture what and how deep each player can see in the game at each choice point.

Definition 3.1 (Sight function)**.** Let $\mathcal{G}^P = ((N, H, \mathrm{t}, \Sigma_i, o), P_i)$ be a prioritized game. A (short) sight function for $\mathcal{G}^P$ is a function

$$\mathrm{s} : H \backslash Z \to 2^H \backslash \emptyset$$

associating to each non-terminal history h a finite subset of all the available histories at h. That is:

1. $\mathrm{s}(h) \in 2^{H|_h} \backslash \emptyset$ and $|\mathrm{s}(h)| < \omega$, i.e. the sight at h consists of a finite nonempty set of histories extending h;

[10] We adopt the terminology of (Osborne and Rubinstein 1994, Proposition 99.2) and refer to the result stating that every finite extensive game has a subgame perfect equilibrium as *Kuhn's theorem*.

2. $h' \in \mathrm{s}(h)$ implies that $h'' \in \mathrm{s}(h)$ for every $h'' \lhd h'$, i.e. players' sight is closed under prefixes.

Intuitively, the function associates to any choice point those histories that the player playing at that choice point can see. Notice that how this set of histories is determined is left open. In other words, the set constitutes the view that the player playing at that non-terminal history has of the remaining of the game. It could be, for instance, all the histories of length at least d, or all histories that start with a given action a, or similar constraints.

The intuition is that $\mathrm{s}(h)$ is the limited view of $\mathrm{t}(h)$ after history h. Such intuition is supported by the fact that $\mathrm{s}(h)$ inherits the moves and the turns from $\mathcal{G}^P$ but not necessarily the terminal nodes. That the view is limited can be noticed by the conditions required in Definition 3.1, which together imply that $l(\mathrm{s}(h)) < \omega$, i.e. players can only see finitely many steps ahead. Several extra conditions, besides the one given in Definition 3.1, might be natural for short sight, e.g.: requiring that the sight increases as the play proceeds, in the sense that what player i can see from h is at least as much as from any history hh'. The present work will not deal with these extra conditions and will limit itself to a general account.

We now define the class of games with short sight.

Definition 3.2 (Games with short sight)**.** A game with short sight is a tuple $\mathcal{S} = (\mathcal{G}^P, \mathrm{s})$ where $\mathcal{G}^P$ is a prioritized game and s a sight function for $\mathcal{G}^P$.

It is clear that each game with short sight yields a family of finite extensive games, one for each non-terminal history:

Fact 2. *Let $\mathcal{S} = (\mathcal{G}^P, \mathrm{s})$ be a prioritized game with short sight, with $\mathcal{G}^P = ((N, H, \mathrm{t}, \Sigma_i, o), P_i)$. Let also h be a finite non-terminal history. Consider the tuple:*

$$\mathcal{E}\lceil_h = (N\lceil_h, H\lceil_h, \mathrm{t}\lceil_h, \Sigma_i\lceil_h, o\lceil_h, \succeq_i \lceil_h)$$

where:

- $N\lceil_h = N$;
- $H\lceil_h = \mathrm{s}(h)$. *The set $Z\lceil_h$ denotes the histories in $H\lceil_h$ of maximal length, i.e., the terminal histories in $H\lceil_h$;*
- $\mathrm{t}\lceil_h = H\lceil_h \backslash Z\lceil_h \to N$ *so that* $\mathrm{t}\lceil_h(h') = \mathrm{t}(h, h')$;

- $\Sigma_i\lceil_h$ *is the set of strategies for each player available at* h *and restricted to* $\mathrm{s}(h)$*. It consists of elements* $\sigma_i\lceil_h$ *such that* $\sigma_i\lceil_h(h') = \sigma_i(h, h')$ *for each* $(h', \sigma_i(h, h')) \in H\lceil_h$ *with* $\mathrm{t}\lceil_h(h') = i$ *;*
- $o\lceil_h \colon \prod_{i\in N} \Sigma_i\lceil_h \to Z\lceil_h$*;*
- $\succeq_i \lceil_h = \succeq^{P_i} \cap (Z\lceil_h)^2$*.*

Tuple $\mathcal{E}\lceil_h$ *is a finite extensive game.*

Remark 3. *It is worth noticing that each* finite *extensive game* $\mathcal{E}_{\mathcal{G}^P}$ *determined by a prioritized game* $\mathcal{G}^P$ *(recall Definition 2.5) is equivalent (modulo the sight function) to the game with short sight built on* $\mathcal{G}^P$ *such that, for each* h*,* $\mathcal{E}_{\mathcal{G}^P}\lceil_h = \mathcal{E}_{\mathcal{G}^P|_h}$*. That is, at each non-terminal history, the game determined by the sight function corresponds to the whole subgame at* h*.*

Remark 4. *The definition of a game with short sight might look odd at first since players are endowed with a sight, which restricts their awareness of the game structure, but, at the same time, their preferences are expressed through P-sequences which are built by assuming access to the set of all histories. It must be clear that this representation is the one of an 'external observer' which knows the players' sights and how players evaluate the terminal histories within their sights according to their P-sequences. The representation is such that players can be considered aware of their preferences only in as much as they have access to the relevant histories.*

3.2 Solving games with short sight

In games with short sight the course of the play is such that at each node players are confronted with decisions to be taken on the grounds of what they can foresee of the game. The purpose of this section is to provide a model of rationality for such situations, i.e. what players should do given the history of the play and their sight.

Subgame perfect equilibria

As we are dealing with self-interested agents, it is natural to think that they will try to get the most out of the information they possess, choosing their best strategy at each choice node. This leads us to a simple adaptation of the notion of subgame perfect equilibrium (Definition 2.7).

Definition 3.3 (Sight-compatible subgame perfection)**.** Take a game with short sight $\mathcal{S} = (\mathcal{G}^P, \mathrm{s})$ and, for each finite history h, let $\mathcal{E}\lceil_h$ be the extensive game yielded by s at h (as defined in Fact 2). A sight-compatible subgame perfect equilibrium of $\mathcal{S}$ is a profile of strategies $\sigma^* \in \prod_{i \in N} \Sigma_i$ such that for every nonterminal history h there exists a strategy profile $\sigma\lceil_h$ that is a subgame perfect equilibrium of $\mathcal{E}\lceil_h$ and such that $\sigma_{\mathrm{t}(h)}\lceil_h(h) = \sigma^*_{\mathrm{t}(h)}(h)$.

Three aspects of the equilibrium definition are worth mentioning. First, each restriction $\mathcal{E}\lceil_h$ prunes the game tree at the bottom (considering the extensions of h) and at the top (considering only the sight-compatible extensions of h). Second, each player i determines his best move supposing that his opponents behave rationally with respect to their P-sequences and relative to the part of the game that i can see. This might be considered a conservative—or safe, depending on the circumstances—way for i to play, by attributing to the opponents the ability to see at least as much as i sees. Third, the definition of subgame perfect equilibrium in games with short sight does not require an explicit finiteness assumption. A finiteness assumption—the finiteness of the histories constituting the sight—is built in Definition 3.1. This brings us to the next section.

An equilibrium existence theorem

Let us start with the following observation:

Fact 3. *Let $\mathcal{S} = (\mathcal{G}^P, \mathrm{s})$ be a game with short sight and h one of its finite non-terminal histories. Then $\mathcal{E}\lceil_h$ has a subgame perfect equilibrium.*

Proof. The fact is a direct consequence of Fact 2 and Kuhn's theorem (Osborne and Rubinstein 1994, Proposition 99.2). □

We can now prove the existence of sight-compatible subgame perfect equilibria (Definition 3.3) for each game with short sight.

Theorem 1. *Every game with short sight has a sight-compatible subgame perfect equilibrium.*

Proof. Let $\mathcal{S} = (\mathcal{G}^P, \mathrm{s})$ be a game with short sight and let σ^* be a strategy profile such that, for each non-terminal history h:

$$\sigma^*_{\mathrm{t}(h)}(h) = \sigma^{BI(\mathcal{E}\lceil_h)}_{\mathrm{t}(h)}$$

where $\sigma^{BI(\mathcal{E}\lceil_h)}$ denotes the strategy profile constructed by the standard backward induction algorithm on the extensive game $\mathcal{E}\lceil_h$ determined by the sight function at history h. The result follows then directly by the construction—via backward induction—of subgame perfect equilibria for each $\mathcal{E}\lceil_h$ (Kuhn's theorem) as $\sigma^{BI(\mathcal{E}\lceil_h)}_{\mathrm{t}(h)}$ is the action dictated to player $\mathrm{t}(h)$ by its backward induction strategy and therefore the action dictated by a subgame perfect equilibrium of $\mathcal{E}\lceil_h$. □

An algorithm for solving games with short sight

By building on the standard backward induction algorithm (*BI*), we can define an algorithm which solves each finite game with short sight by constructing a terminal history, the one determined a sight-compatible subgame perfect equilibrium of the game.

Definition 3.4 (BI-path in games with short sight)**.**

Input: A finite game with short sight $\mathcal{S} = (\mathcal{G}^P, \mathrm{s})$

Output: A terminal history $(x_0, \ldots, x_n)$ of $\mathcal{G}^P$

Method: 1. Define $h := \emptyset$;

2. Run *BI* over $\mathcal{E}\lceil_h$ and set $h := \left(h, \sigma^{BI(\mathcal{E}\lceil_h)}_{\mathrm{t}(h)}\right)$;

3. If $h \in Z$ then return h, otherwise repeat step 2.

It is easy to see that the algorithm terminates and constructs indeed a history consisting of actions dictated by a sight-compatible subgame perfect equilibrium. Intuitively, the algorithm starts at the root and solves $\mathcal{E}\lceil_\emptyset$. This yields a terminal history in $Z\lceil_\emptyset$, and their initial fragments of length 1 are taken as the first moves of the histories returned by the algorithm. Each of these first moves determine, in turn, as many extensive games via the sight function. These are solved in the same way, determining a set of histories of length 2, and so on, until terminal histories of $\mathcal{G}^P$ are built.

Remark 5. *Before concluding this section, it is worth stressing an important aspect of sight-compatible subgame perfect equilibrium. In games with short sight, players could be considered as having preferences not only on the terminal histories within their sight, but also over the non-terminal ones—due to the fact their preferences are determined by P-sequences. However, preferences over non-terminal histories are disregarded when solving the game, as players proceed by backward induction from the terminal histories*

in their sight, therefore possibly overruling any preference they have over intermediate histories.

4 Short sight and unawareness

This section is devoted to establishing the precise relationship between games with short sight and games with possibly unaware players elaborated by Halpern and Rêgo in Halpern and Rêgo (2006). As already pointed out, the models focused upon in Halpern and Rêgo (2006) feature players that can always observe at least some of the terminal histories of the actual game being played. In the same paper, in order to overcome this limitation, Halpern and Rêgo generalize their models to allow players to hold false beliefs about the game being played although, it must be mentioned, they do not provide an equilibrium analysis of that class of games. Essentially, at each node of a game each player might believe to be playing a completely different game from the one that he or she is actually playing. These generalized models are extremely abstract and can incorporate several forms of unawareness. Even though the intuitive understanding of short sight is rather different from that of false belief, the models in Halpern and Rêgo (2006) can be formally related to our models. To establish this relationship we proceed as follows:

1. We formally introduce games with possibly unaware players and lack of common knowledge of the underlying game, the most general model of unawareness provided in Halpern and Rêgo (2006). We will refer to this class of models simply as *games with awareness* (Subsection 4.1).

2. We provide a canonical representation of games with short sight as games with awareness. In short, we are going to build a class of the latter models where, at each position of the actual game being played, players *believe to be playing a game that corresponds to their own sight*. We show, moreover, that the canonical representation is of the right kind, i.e. it obeys the axioms of the general models of Halpern and Rêgo (Subsection 4.2).

3. We provide the axioms that exactly characterize games with short sight as games with awareness (Subsection 4.3).

4.1 Games with awareness

Halpern and Rêgo work with finite extensive games endowed with information sets and probability measures Halpern and Rêgo (2006). As the games structures dealt with in our paper do not model epistemic aspects such as knowledge and belief, the comparison to which this section is devoted will concern the somewhat more fundamental level of the finite extensive games with perfect information upon which Halpern and Rêgo base their models.

To each extensive game $\mathcal{E} = ((N, H, \mathsf{t}, \Sigma_i, o), \succeq_i)$, Halpern and Rêgo (2006) associates an *augmented game* ${}^{+}\mathcal{E}$ that specifies the level of awareness of each player at each node of the original game. The following definition is adapted from Halpern and Rêgo (2006).

Definition 4.1 (Augmented game)**.** Let $\mathcal{E} = (\mathcal{G}, \succeq_i)$ be a finite extensive game and, for each history h (not necessarily belonging to the set of histories of $\mathcal{G}$), let $\overline{h}$ be the subsequence of h consisting of the moves in h that are made by actions available in $\mathcal{G}$. The *augmented game* ${}^{+}\mathcal{E} = (((N, H, \mathsf{t}, \Sigma_i, o), \succeq_i), Aw_i)$ based on $\mathcal{G}$ is such that:

A1 $(N, H, \mathsf{t}, \Sigma_i, o), \succeq_i)$ is a finite extensive game;

A2 $Aw_i : H \to 2^{H'}$ is the awareness function of each player i, that maps each history to a set of histories (in $2^{H'}$) of some arbitrary finite extensive game $\mathcal{E}'$. For each $h \in H$ the set $Aw_i(h)$ consists of histories in H' and their prefixes.

A11 $\{\overline{z} \mid z \in Z\} \subseteq Z$, i.e. the terminal histories of the game ${}^{+}\mathcal{E}$ correspond to terminal histories of $\mathcal{E}$; moreover if z' is a terminal history of ${}^{+}\mathcal{E}$ then $z' \in Z$, i.e. terminal histories of which players are aware are terminal histories of the game $\mathcal{E}$ upon which ${}^{+}\mathcal{E}$ is based.

A12 for each terminal history $z \in Z$ such that $\overline{z} \in Z$ we have that $z \succeq_i \overline{z}$ and $\overline{z} \succeq_i z$ for each $i \in N$, i.e. players' preferences are inherited from game $\mathcal{E}$ upon which ${}^{+}\mathcal{E}$ is based.

The items in the definition keep the original names of axioms A1, A2, A11 and A12 given in Halpern and Rêgo (2006) for games with lack of common knowledge.

We can now formally introduce a game with awareness in its most general form.

Definition 4.2 (Games with awareness)**.** Let $\mathcal{E}$ be a finite extensive game. A *game with awareness* based on $\mathcal{E}$ is a tuple $\mathcal{E}^{Aw} = (\Gamma, \mathcal{E}^m, \mathcal{F})$, where:

- Γ is a countable set of augmented games each one based on some (possibly different) game $\mathcal{E}'$;
- $\mathcal{E}^m$ is a distinguished augmented game based on $\mathcal{E}$;
- $\mathcal{F}$ is a mapping that associates to each augmented game ${}^+\mathcal{E}' \in \Gamma$ and history h' of ${}^+\mathcal{E}'$ an augmented game $\mathcal{E}_{h'}$. This game is the game the player whose turn is to play believes to be the true underlying game when the history is h'.[11]

The definition spells out the crucial feature of a game with awareness, namely the fact that each player at each history is associated to a game that he belives to be the current game. This can be distinct from the current game being played, which is instead observed by an omniscent modeller. Specifically, while each ${}^+\mathcal{E}'$ is the point of view of some player at some history (the precise relation is given by the $\mathcal{F}$ mapping), $\mathcal{E}^m$ is the point of view of the omniscent modeller, who can actually see the game that is being played and the players' awareness level. Definition 4.2 is extremely abstract and can be refined by imposing several reasonable constraints, especially with respect to $\mathcal{E}^m$, the point of view of the modeller. The following definition, adapted from Halpern and Rêgo (2006), takes care of that.

Definition 4.3 (Games with awareness: constraints)**.** The class of games with awareness is refined by the following constraints, for each $\mathcal{E}^{Aw} = (\Gamma, \mathcal{E}^m, \mathcal{F})$:

M1 $N^m = N$, i.e. the modeller is aware of all the players;

M2 $A \subseteq A^m$ and $\{\overline{z} : z \in Z^m\} = Z$, i.e. the modeller is aware of all the moves available to the players and knows the terminal histories of the game;

M3 If $\mathrm{t}^m(h) \in N$ then $A^m(h) = A(\overline{h})$, i.e. the modeller is aware of the possible courses of the events;

C1 $\{\overline{h'} \mid h' \in H_h\} = Aw_i(h)$, i.e. the awareness function shows exactly the histories that can be observed.

[11] Henceforth, to reduce clutter in notation, we use the subscript $_{h'}$ to index the elements of game tuple $\mathcal{E}_{h'}$, i.e. the game that player t(h') believes to be playing at history h'. For instance $H_{h'}$ is the set of histories that player t(h') believes to be the set of histories that are available when he is in h'.

The constraints just discussed hold for all games with awareness. The following part lays a first bridge between these structures and games with short sight.

4.2 Canonical representation

In Rêgo and Halpern ((to appear) a canonical representation is provided of a finite extensive game as a game with awareness. For the present purposes, which are not concerned with epistemic aspects, a finite extensive game $\mathcal{E}$ is representable as a tuple $(\{\mathcal{E}^m\}, \mathcal{E}^m, \mathcal{F})$ where $\mathcal{E}^m = (((N, H, \mathrm{t}, \Sigma_i, o), \succeq_i), Aw_i)$ with $Aw_i(h) = H$ for all $h \in H$ and $\mathcal{F}(\mathcal{E}^m, h) = \mathcal{E}^m$. Essentially, all players and the modeller are aware of the game and agree on it. Likewise in this section we provide a canonical representation of games with short sight in terms of the general models introduced above (Definitions 4.2 and 4.3).

Definition 4.4 (Canonical representation of short sight)**.** Take a finite prioritized game with short sight $(\mathcal{G}^P, \mathrm{s})$ where $\mathcal{G}^P = ((N, H, \mathrm{t}, \Sigma_i, o), P_i)$. Let also h be a finite non-terminal history and $\mathcal{E}\lceil_h$ the resulting extensive game as in Definition 2. The *canonical representation* of $(\mathcal{G}^P, \mathrm{s})$ consists of the tuple

$$\mathcal{E}^{(\mathcal{G}^P, \mathrm{s})} = (\{\{(\mathcal{E}\lceil_h, Aw_i\lceil_h) \mid h \in H\}, \mathcal{E}^m\}, \mathcal{E}^m, \mathcal{F})$$

where:

1. $\mathcal{E}^m = (((N, H, \mathrm{t}, \Sigma_i, o), \succeq_i), Aw_i)$ with $Aw_i(h) = H\lceil_h = \mathrm{s}(h)$;
2. $Aw_i\lceil_h(h') = Aw_i(h, h')$;
3. for each ${}^+\mathcal{E} \in \Gamma$, ${}^+\succeq_i = ({}^+Z \times {}^+Z) \cap \succeq^{P_i}$;
4. $\mathcal{F}(\mathcal{E}^m, h) = (\mathcal{E}\lceil_h, Aw_i\lceil_h)$;
5. $\mathcal{F}((\mathcal{E}\lceil_h, Aw_i\lceil_h), h') = (\mathcal{E}\lceil_{(h,h')}, Aw_i\lceil_{(h,h')})$.

In words, a game with short sight can be represented as a game with awareness where at each choice point players believe to be playing the game induced by their sight. Specifically, the first item says that the modeller knows the structure of the game and the sight of the players at each point. The second item says that players' sight in each augmented game agrees with their sight in the original game. The third item says that every augmented game is consistent with the P-sequence in its terminal nodes. The fourth and fifth item say that the awareness function returns the sight of the players at each decision point.

The following result shows that the above representation of games with short sight yields the right sort of games with awareness.

Theorem 2. *Let* $(\mathcal{G}^P, \mathrm{s})$ *be a game with short sight.* $\mathcal{E}^{(\mathcal{G}^P,\mathrm{s})}$ *is a game with awareness.*

Proof. We first need to check that Γ^* is made by a countable set of augmented games and then that they satisfy the axioms given in Definition 4.3. As for the first part we need to show that the axioms of Definition 4.1 are satisfied: [A1] we know that the game $(\mathcal{G}^P, \mathrm{s})$ is finite (Definition 4.4) and that each $\mathcal{E}\lceil_h$ for h being a history of $\mathcal{E}$ is a finite extensive game (Proposition 2); [A2] Aw_i is well defined, as it associates each history of each augmented game exactly the sight of the player who moves at that history. Players' sight is closed under prefixes by Definition 3.1; [A11-12] Notice that by the construction in Proposition 2 for each history $h \in H$ we have that $\overline{h} = h$. By reflexivity of preferences we obtain the desired result. As for the second part we need to show that the following axioms are satisfied: [M1-M3, C1] Consequence of Definition 4.4 . □

4.3 Characterization result

In this section we provide the constraints that a game with awareness needs to satisfy in order to be the canonical representation of some game with short sight. Before doing this we introduce the auxiliary notion of game pruning.

Definition 4.5 (Game pruning)**.** Let $\mathcal{E} = ((N, H, \mathrm{t}, \Sigma_i, o), \succeq_i)$ be a finite extensive game. The game $\mathcal{E}' = ((N', H', \mathrm{t}', \Sigma_i', o'), \succeq_i')$ is a *pruning* of game $\mathcal{E}$ whenever

- $N = N'$;
- $H' \subseteq H$ and H' is a finite set of histories closed under prefixes;
- for each $h' \in H'$, $\mathrm{t}'(h') = \mathrm{t}(h')$;
- $\Sigma_i' = \{\sigma_i \in \Sigma_i \mid \sigma_i : h' \to A \text{ for } h' \in H' \text{ with } \mathrm{t}'(h') = i \text{ and there is a } h'' \in H' \text{ with } h' \lhd h''\}$;
- for each $\sigma' \in \Sigma'$, $o'(\sigma') = z'$ whenever $z' \in Z'$ and is obtained by executing σ'.[12]

[12]Formally, for $z = (z_1, z_2, \ldots, z_{l(z)})$ and $\forall i \in \{1, 2, \ldots, l(z)\}$ we have that $\sigma_{\mathrm{t}(z_i)}(z_i) = z_{i+1}$.

A game pruning of an extensive game $\mathcal{E}$ is just $\mathcal{E}$ deprived of some histories, preserving the structure of strategies and turn function and defining the outcome function accordingly. Notice that a game pruning of a game is nothing but what we called a sight (recall Definition 3.1), defined at the root of the game.

The following definition makes use of game prunings, isolating a class of games with awareness with which we will be able to exactly characterize games with short sight.

Definition 4.6 (Coherence)**.** Let $\mathcal{E}^{Aw} = (\Gamma, \mathcal{E}^m, \mathcal{F})$ be a game with awareness based on a finite extensive game $\mathcal{E} = ((N, H, \mathrm{t}, \Sigma_i, o), \succeq_i)$. We call $\mathcal{E}^{Aw}$ *coherent* if it satisfies the following constraints:

K1 the game $\mathcal{E}^m$ is the tuple $(((N, H, \mathrm{t}, \Sigma_i, o), \succeq_i), Aw_i)$ with $Aw_i(h) = H'$ for H' being the set of histories of some game ${}^{+}\mathcal{E} \in \Gamma$;

K2 the set Γ comprises $\mathcal{E}^m$ and for each $h \in H$ a set of |H| augmented games of the form $(\mathcal{E}'|_h, Aw'_i)$, with $\mathcal{E}'$ being a *pruning* of $\mathcal{E}$, and $Aw'_i(h') = Aw_i(h, h')$;

K3 there exists a total preorder $\succeq_i^H$ on H extending $\succeq_i$ such that for each ${}^{+}\mathcal{E} \in \Gamma$ we have that ${}^{+}\succeq_i = \succeq_i^H \cap ({}^{+}Z \times {}^{+}Z)$, i.e. histories get the same preferences across augmented games;

K4 $\mathcal{F}(\mathcal{E}^m, h') = (\mathcal{E}'|_{h'}, Aw'_i)$, for $\mathcal{E}'$ being the pruning of $\mathcal{E}$ associated to h';

K5 for each $(\mathcal{E}'|_h, Aw'_i) \in \Gamma$ we have that $\mathcal{F}((\mathcal{E}'|_h, Aw'_i), h') = (\mathcal{E}'|_{(h,h')}, Aw''_i)$, where $Aw''_i(h'') = Aw_i(h, h', h'')$.

The constraints deal with the game form structure and the preferences of coherent games with awareness. Axiom K1 states that the modeller has a perfect view of the game and of the awareness of each player at each history. Notice that by K1, awareness of players agrees at each decision point.[13] Axiom K2 states that players can only see a part of the real game being played. Axiom K3 deals instead with the preference relations and ensures that histories are evaluated according to the same criteria if observed from different points. Axioms K4-5 state that what players believe to be true in the real game at a point coincides with their awareness level at that point. Notice the resemblance of these axioms with the conditions on Definition 4.4.

We first prove the following lemma:

[13] The requirement looks rather strong, but notice that for decision making purposes the only awareness level that matters is the one of the player who is to move.

Proposition 1 (P-sequence existence)**.** *Let $\mathcal{E}^{Aw} = (\Gamma, \mathcal{E}^m, \mathcal{F})$ be a game with awareness that is coherent. We can construct a finite game with short sight $\mathcal{G}^P = (\mathcal{G}, P_i)$ such that $Z \times Z \cap \succeq^{P_i} = \succeq_i$ where Z and $\succeq_i$ are the terminal histories and the preference relation for player i in any game ${}^+\mathcal{E} \in \Gamma$.*

(sketch). Let $((N, H, \mathrm{t}, \Sigma_i, o) \succeq_i)$ be the game $\mathcal{E}$ upon which $\mathcal{E}^m$ is based. Consider its game form $(N, H, \mathrm{t}, \Sigma_i, o)$. We construct the desired P-sequence as follows. Let $\succeq_i^H$ be the total preorder required by axiom K3 (Definition 4.6) and let $\succ_i^H$ indicate its strict counterpart. Let moreover $\mathcal{H} = \{[h] \mid h' \in [h] \iff h' \succeq_i^H h \text{ AND } h \succeq_i^H h'\}$. Intuitively, $\mathcal{H}$ is the set of all equivalence classes induced by the relation $\succeq_i^H$. The desired P-sequence $(\mathcal{H}, \succ)$ is so defined for each $\mathbf{H}, \mathbf{H}' \in \mathcal{H}$:

$$\mathbf{H} \succ \mathbf{H}' \text{ if and only if for some } x \in \mathbf{H}, y \in \mathbf{H}' \text{ we have } x \succ_i^H y$$

We need to show (i) that $(\mathcal{H}, \succ)$ is indeed a P-sequence and (ii) that it displays the required properties. As for (i) set $\mathcal{H}$ is clearly a finite set of subsets of H. We are left to show that the relation $\succ$ is (a) irreflexive (b) transitive (c) asymmetric and (d) total. (a) Suppose not, then for some $\mathbf{H} \in \mathcal{H}$ and $x, y \in \mathbf{H}$ we would have $x \succ_i^H y$, leading to contradiction. Claims (b) - (c) - (d) can be proven by a similar procedure. (ii) For any two histories h', h and ${}^+\mathcal{E} \in \Gamma$ with preference relation $\succeq_i$ and with h', h among the terminal histories of ${}^+\mathcal{E}$ we need to show that: $h \succeq_i h'$ if and only if $h \succeq^{(\mathcal{H}, \succ)} h'$. Both directions are straightforward. □

We are now ready to formulate our main result.

Theorem 3 (Correspondence)**.** *Let $\mathcal{E}^{Aw} = (\Gamma, \mathcal{E}^m, \mathcal{F})$ be a coherent game with awareness based on $\mathcal{E}$. There exists a finite game with short sight $(\mathcal{G}^P, \mathrm{s})$ such that its canonical representation $\mathcal{E}^{(\mathcal{G}^P, \mathrm{s})}$ is such that $\mathcal{E}^{Aw} = \mathcal{E}^{(\mathcal{G}^P, \mathrm{s})}$.*

Proof. We proceed by construction. Let $((N, H, \mathrm{t}, \Sigma_i, o) \succeq_i)$ be the game $\mathcal{E}$. Consider its game form $(N, H, \mathrm{t}, \Sigma_i, o)$. To construct the game $(\mathcal{G}^P, \mathrm{s})$ first use Proposition 1 to obtain the desired P-sequence P_i for each player. As for the sight function we simply impose the following: for every history $h \in H$, and every player $i \in N$ we have that $\mathrm{s}(h) = Aw_i(h)$, where $Aw_i(h) = H'$ is the awareness function as appears in $\mathcal{E}^m$. The requirements of Definition 3.1 are satisfied as a consequence of the fact that $\mathrm{s}(h)$ is always the set of histories of some finite game following h (Definition 4.6). Now the fact that $\mathcal{E}^{Aw} = \mathcal{E}^{(\mathcal{G}^P, \mathrm{s})}$ follows from Definitions 4.4 and 4.6. □

Theorems 2 and 3 have established a precise link between the most general class of games with awareness introduced in Halpern and Rêgo (2006)—i.e., games with awareness and lack of common knowledge of the game structure—and the class of games with short sight, namely that the latter is a special subclass of the former. This puts the results presented in Section 3 in an interesting light. In fact, Halpern and Rêgo (2006) did not develop any equilibrium analysis of games with awareness and lack of common knowledge of the game structure. The notion of sight compatible subgame perfect equilibrium can therefore be viewed as a first principled generalization of subgame perfection to a specific form of unawareness—short sight.

5 Conclusions

Inspired by Joseph Halpern's invited talk at AAMAS 2011—*Beyond Nash-Equilibrium: Solution Concepts for the 21st Century*—and moving from simple considerations concerning real life game playing (Example 1), the paper has proposed a class of games where players are characterized by two key features: 1) they have only partial access to the game structure including, critically, having possibly no access to terminal nodes; 2) they play according to extrinsic evaluation criteria, which have here been modeled as sequences of properties of histories (Definition 2.3). The paper has shown thas such games 1) always possess an appropriate refinement of the subgame perfect equilibrium concept (Theorem 1); 2) are an interesting—because of the above equilibrium properties—subclass of the most general class of games with awareness proposed by Halpern and Rêgo (Theorems 2 and 3) which, although introduced in Halpern and Rêgo (2006), had not yet been object of investigation from the point of equilibrium analysis.

Future work will focus on weakening two assumptions. First, the fact that in solving games with short sight we have presupposed that players only consider their own sight (Definition 3.3) and that the evaluative components of the game—the P-sequences—are common knowledge. Dropping these assumptions could open up interesting avenues of research concerning learning methods by means of which players could infer other players' evaluation criteria and sights, i.e., other players' types. This would bring the game-theoretical method of equilibrium analysis close to established game-playing techniques in artificial intelligence and some of its recent developments such as the theory

of *general game playing* Genesareth et al. (2005). Second, it is clear that the granularity of their evaluation criteria has direct impact on players' performance in a game with short sight. We have currently defined P-sequences as sequences of sets of histories. A more refined approach would take into consideration the (formal) language by means of which players express their evaluation criteria. Methods from logic could then be used to compare the expressivity of different languages for P-sequences, possibly correlating such expressivity to players' performance in the games.

Acknowledgements. Paolo Turrini acknowledges the support of the National Research Fund of Luxembourg for the Trust Games project (1196394), cofunded under the Marie Curie Actions of the European Commission (FP7-COFUND).

References

Y. Feinberg. Subjective reasoning—games with unawareness. Technical Report Research Paper Series 1875, Stanford Graduate School of Business, 2004.

Y. Feinberg. Games with incomplete unawareness. Technical report research paper series, Stanford Graduate School of Business, 2005.

M. Genesareth, N. Love, and B. Pell. General game playing: Overview of the aaai competition. *AAAI Magazine*, 2005.

J. Halpern. Beyond nash-equilibrium: Solution concepts for the 21st century. In K. Apt and E. Grädel, editors, *Lectures in Game Theory for Computer Scientists*, pages 264–289. Cambridge University Press, 2011.

J. Y. Halpern and L. C. Rêgo. Extensive games with possibly unaware players. In *In Proceedings of the 5th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2006)*, pages 744–751, 2006.

J. Y. Halpern and L. C. Rêgo. Extensive games with possibly unaware players. *CoRR*, abs/0704.2014, 2007.

F. Liu. A two-level perspective on preference. *Journal of Philosophical Logic*, 40 (3):421–439, 2011.

M. Osborne and A. Rubinstein. *A course in Game Theory*. The MIT Press, 1994.

L. C. Rêgo and J. Y. Halpern. Generalized solution concepts in games with possibly unaware players. *International Journal of Game Theory*, (to appear).

J. Watson. *Secrets of Modern Chess Strategy: Advances since Nimzowitch*. Gambit Publications, 1998.

Making Choices in Social Situations

Meiyun Guo and Jeremy Seligman

Institute of Logic and Intelligence, Southwest University, Chongqing, China.
guomy007@swu.edu.cn, j.seligman@auckland.ac.nz

Abstract

We propose a general account of decision making in social situations based on an analysis of the role of three concepts: knowledge, preference and freedom of choice. The normative aspect of decision making is sharply contrasted with the descriptive aspect, as is the distinction between *a priori* and *a posteriori* rationality. As a partial validation of the analysis, we apply our account to the theory of strategic games with both pure and mixed (probabilistic) strategies, showing that the concept of a dominated strategy and Nash equilibrium are correctly predicted by more general norms. Our account is purely model-theoretic but uses discrete relational structures that are well-suited for future application of the techniques of modal logic.

1 Introduction

Modern decision theory has developed into an interdisciplinary subject pursued by researchers from economics, psychology, philosophy, mathematics, and statistics Peterson (2009). And, since the middle of 20th century logicians have been interested in the norms of rational decision Jeffrey (1965). Social de-

cision theory extends the theory of individual decisions to the case in which the decisions of more than one agent interact. This is closely related to the field of game theory Osberne and Rubinstein (1994). In recent years, techniques from modal logic (especially epistemic and preference logic) have been widely used to study game theory, e.g, van Benthem (2001), Pauly (2002), van Benthem et al. (2006), van Benthem (2007), Bonanno (2008). A recent survey is van der Hoek and Pauly (2006). In particular van Benthem et al. (2006), van Benthem (2007) and Bonanno (2008) give detailed logical accounts of game-theoretic concepts such as Nash Equilibrium and the procedure of iterated deletion of strictly (or weakly) dominated strategy in strategic-form games. van Benthem (2001) and Pauly (2002) concentrate on individual power and the ability of coalitions in extensive-form games.

Our approach is to start with a very general account of decision-making, first in the single-agent case, and then in a social setting, and only later to consider the application to game theory. In this respect, our work is similar to Lorini and Schwarzentruber (2010), which starts from a general theory of agency. However, we make a sharp distinction between the descriptive and normative aspects of decision-making, with the aim of showing precisely which assumptions are made along the way. Also, unlike Lorini and Schwarzentruber (2010), we do not begin with an explicit model of actions. Instead, our account is based on only three conceptual primitives: knowledge, preferences and freedom of choice. Each of these is understood as a counterfactual relation between possible decision states. The novelty here is the 'freedom of choice' relation, which holds between one's current situation u and some other possible situation v when one could have been in situation v instead, if one had chosen differently. Finally, our approach is entirely model-theoretic. We do not consider formal languages in this paper.

In Section 2, we introduce these three relations and show how they may be used to define some central concepts such as independence, determinism, value and decision itself. Section 3 aims to identify the norms of rationality that apply to decision-making. We start with a series of arguments about the irrationality of certain decisions under different idealising assumptions, and gradually remove those assumptions to produce a general norm of decision-making. Along the way, we find it important to distinguish between *a priori* and *a posteriori* rationality. The section ends with a discussion of the typical assumptions about the preference relation, such as the reflexivity, transitivity and totality of the 'at least as good as' relation, which we show to be largely irrelevant to the norms of decision-making.

Section 4 extends the account from the single-agent case to decisions made in a social setting. Here the concept of freedom to choice is extended from individuals to groups, and this enables us to describe when an agent's decisions are independent of other agents. The Reduction Lemma (3) identifies sufficient conditions for the reduction of knowledge to freedom of choice, i.e., when every agent is ignorant only of the other agents' choices.

Section 5 gives a standard presentation of strategic games, with both pure and mixed strategies, and shows how they can be understood as social decision frames. This ends in Representation Theorem 1, which identifies the class of social decision frames arising from games. Furthermore, within this class of frames, we show how the general norms of *a priori* rational decisions and the concept of *a posteriori* rationality characterise the basic game theoretic concepts of weakly dominated strategy, best response and Nash equilibrium.

2 The Components of Decision: Freedom, Ignorance and Preference

We model decisions within a structure that is used to represent our knowledge, preferences, and freedom to decide. A *decision frame* F is defined to be a relational structure $\langle W, \sim, \approx, \leq \rangle$ where $\sim$ and $\approx$ are equivalence relations and $\leq$ is an arbitrary binary relation.

The elements of W represent possible decisions and we interpret $\sim$ as representing your freedom to choose: $u \sim v$ means that in situation u you *could have* been in situation v, if you had made a (possibly different) choice. If $u \nsim v$ then no matter what you decided, the result would not have been your being in situation v; contextual factors operating in u are such that this is simply impossible. For example, suppose you chose to take your umbrella with you when you left your home this morning. Now, in situation u, it is raining and you are happy to have the umbrella. You could have also worn a raincoat, or left both at home. In the first case, you would be in situation v_1, wearing a raincoat in the rain; in the second you would very wet, in situation v_2. Both $u \sim v_1$ and $u \sim v_2$. All three situations lay within your freedom of choice. But situation v_3 in which you are without raincoat and umbrella, and still dry because it isn't raining; that is not. The equivalence class (u) of situation u therefore represents what was possible, given your freedom of choice. We say that some state-of-affairs

$X \subseteq W$ is *independent* of your choices if $(u) \subseteq X$ iff $u \in X$. In other words, if X occurs in some situation u then you could not have chosen in such a way that X would not have occurred and if X does not occurs in u, you could not have chosen in such a way that X would have occurred. Whether it rains is, contrary to popular opinion, quite independent of whether you carry an umbrella.

The second component of a decision frame, the relation $\approx$, represents your knowledge concerning the consequences of your decision. As usual, we interpret $u \approx v$ to mean that you are unable to distinguish between u and v on the basis of what you know, and consider $[u]$, the $\approx$-equivalence class of u to represent your knowledge state in u. That $[u]$ is not always $\{u\}$ represents the gap between your knowledge of the consequences of your decision, and what those consequences really are, which is determined in part by the contextual factors in operation when you make the decision. Another way of thinking of $[u]$ is as the most specific state-of-affairs that you ensured to be the case by choosing as you did when you chose u. We will therefore also refer to $[u]$ as your *decision* in situation u.

We will also be interested in the effects of acquiring new knowledge. Just as your knowledge is represented by an equivalence relation, so also is the information that you might acquire. If E is such an equivalence relation, then learning E is a matter of learning to discriminate between situations that are not E-related, or, in other words, in any one situation, learning which of the E-equivalence classes it belongs to. The effect is to update your knowledge from $\approx$ to $\approx \cap E$.[1]

When making decisions, we are particularly interested in the effect of knowing your own freedom to choose, given the (originally unknown) contextual factors in operation when you made your decision. This is given by $\approx \cap \sim$. The equivalence class $(u)[u] = (u) \cap [u]$ of this relation may still fail to be the singleton $\{u\}$ for situations u in which the outcome of your decision depends on an essentially non-deterministic process. Even with full knowledge of what you could have achieved, given operant contextual factors, there may still be something further that influences the outcome.[2] We will say that u is *determin-*

[1] This approach is taken, for example, in van Benthem and Ştefan Minică (2009). Note that this update assume that the change in your knowledge has no effect on the contextual factors that might influence the outcome of your decision. In the more complicated setting of social decision frames, considered in Section 4, this is not necessarily true.

[2] The kind of non-determinism involved here is not just the randomness of the roll of a die, about which you at least have some information in the form of a probability distribution, but something about which you may have no information at all, such as the actions of other free agents subsequent

istic iff $(u)[u] = \{u\}$. Likewise, a frame will be said to be deterministic when every situation in it is deterministic.

Finally, $u \leq v$ represents your regarding v as at least as good as u. A strict preference $u < v$ can be defined as $u \leq v$ and $v \not\leq u$, meaning that you regard v as better than u. Likewise, indifference between u and v, written $u \bowtie v$, is defined by $u \leq v$ and $v \leq u$. There is also the possibility $u\#v$ that neither $u \leq v$ nor $v \leq u$. We interpret this as some sort of conflict in your preferences between u and v. Although preference is taken as our primitive notion, rather than some notion of value, the relation $\bowtie$ holds between situations of equal value, and so we can define your *values* to be the sets $\bowtie (u) = \{v \in W\} u \bowtie v$. We will need to consider the size of the set of your values, and so define the *value size* of a decision frame to be the cardinality of the set of its values.

We intend a decision frame only to model the facts of decision making, not the norms. So no assumption is made that $\leq$ is either reflexive or transitive or total, although, as is often claimed and we will be shown below, failure of one's preferences to have these properties leads to undesirable consequences.[3] Note, however, that the definitions ensure that strict preference ($<$) is asymmetric and both indifference ($\bowtie$) and conflict (#) are symmetric. In the special case that $\leq$ is a linear order (as is often assumed in the literature on decision theory and game theory), we will say that the frame is *linear*.

3 Norms of Decision Making

We will start by comparing events in a frame in which we have complete knowledge of all possible decisions (so $\approx$ is the identity relation) and there are no limitations on our freedom to decide (so $\sim$ is the universal relation). In such circumstances, the basic norm of preference applies: do not choose one thing when there is something better.

to your making your decision.

[3]Failure of transitivity can also lead to vague or overlapping values, since in this case, $\bowtie (u)$ is not an equivalence class, and it is possible for $\bowtie u \cap \bowtie (v)$ without $u \bowtie v$.

Ideal Decisions

If $\approx$ is the identity relation and $\sim$ is the universal relation then if u is rational there is no $v > u$.

3.1 Coping with ignorance

Typically, however, we are not blessed with such ideal conditions. In particular, we may be ignorant of the consequences of our decisions and so unsure of which situation we are in. This lack of complete knowledge is represented by a non-trivial $\approx$ relation. If we are actually in situation u, then we know only that we are in one of the situations of the equivalence class $[u]$. Now suppose there is another situation v that we prefer to u. We cannot be faulted in our decision, if we do not know that it would result in our being in u. And even if we know v to be better than u, even this is not enough. Because, in state v we would not know we are in that state, but only in the class $[v]$. To judge whether we are rational in making our decision, we must therefore compare the whole of $[u]$ with the whole of $[v]$.

An example will help to think about this. Suppose there is a pile of envelopes in front of you. Some of the envelopes are coloured red and some are coloured green. You know that all the envelopes are initially empty and then see me place a bank note (of suitably large denomination) in one of the red ones. Now you are asked to pick one. Given that all the envelopes look the same apart from their colour, your only decision is the choice between red and green. Assuming that you want to get some money, although there is no guarantee of getting it, you should not be completely indifferent; you should choose a red envelope.

After choosing a red envelope, you open it and find it empty. Call the resulting situation r_0. There was another possible situation r_1 in which you picked the envelope with the money. But at the time of picking a red envelope, you did not know whether you were in situation r_0 or r_1, which we represent with $r_0 \approx r_1$. Despite your disappointment, your rationality cannot be faulted, since if you have chosen a green envelope, you would be bound to do no better, and there was at least a chance of getting the money.

This form of reasoning can be captured by defining defining the relation of *known preference* to be

$$u \leq_K v \text{ iff } u' \leq v' \text{ for all } u' \approx u \text{ and all } v' \approx v.$$

Strict known preference ($u <_K v$), indifference ($u \approx_K v$) and conflict ($u\#_K v$) can be defined as before. Now suppose that faced with the pile of envelopes, you had chosen a green one, putting you in situation g, with again no money. The reason that the decision you would have made in g is worse than the one you actually made in r_0, despite the similar outcome, is that $g <_K r_1$, which is to say $g \leq_K r_1$ (you knew your actual decision to be at least as good as choosing green) and $r_1 \not\leq_K g$ (you did not know that choosing green would be at least as good as your actual decision, because of the possibility that you would be in state r_1, with the money, instead of in state r_0).

By adjusting for the role of knowledge (or the lack of it) in our deliberations, we have a second norm:

Known Decisions

If $\sim$ is the universal relation then if u is rational there is no $v >_K u$.

Note that the norm of Ideal Decisions is a special case of Known Decisions, and they are equivalent under the assumption of complete knowledge ($\approx$ is the identity relation).

3.2 Limited freedom to choose

Next, we relax the assumption that you are able to choose without restriction. In conditions of complete knowledge, this a relatively straightforward change. The rationality of your decision u is not defeated by a decision situation v that you could not possibly have attained.

Free Decisions

If $\approx$ is the identity relation then if u is rational there is no $v \sim u$ such that $v > u$.

Again, Ideal Decisions is a special case of Free Decisions, and they are equivalent under the assumption of unrestricted freedom ($\sim$ is the universal relation).

3.3 The interaction of knowledge and freedom

Finally, we must consider the interaction between knowledge and freedom that occurs when we drop both idealising assumptions. In the general case, the existence of a preferred v that one could have achieved is not sufficient to make your decision u irrational. This is for the same reason of ignorance that we have already considered. v must not only be attainable, it must also be known to be a better decision than u. Moreover, v be be known to be attainable.

Known Free Decisions

If u is rational there is no $v \sim_K u$ such that $v >_K u$.

Here $\sim_K$ is defined in a parallel way to $>_K$, namely,

$$u \sim_K v \text{ iff } u' \sim v' \text{ for all } u' \approx u \text{ and all } v' \approx v.$$

Yet this too is relatively uninteresting. It is merely the restriction of Known Decisions to the class of attainable decisions. Nonetheless, it is a generalisation: both Free Decisions and Known Decisions are special cases.

More interesting, is to consider contextual factors, whose effect on your freedom is unknown. Returning to our previous example, and simplifying a little, suppose there are just four envelopes, which we'll refer to as 1 and 2 (red), 3 and 4 (green). Although we presented the example as a choice between four envelopes, there are really not four but eight possible decision situations for you to consider, namely each of the situations $u(n, m)$ in which you chose envelope n (1,2,3, or 4) and the money is in envelope m (1 or 2). Suppose that the actual situation is $u = u(3, 2)$: you stupidly chose envelope 3 (green) and the money is in envelope 2 (red). This is not shown to be irrational by the norm of Known Free Decisions. Although there is a situation $v = u(2, 2)$ that you know to be better than u ($v >_K u$), and you had the freedom to be in that situation ($v \sim u$), you did not *know* that you had that freedom. You cannot distinguish between the actual situation $u = u(3, 2)$ and the situation $u(3, 1)$, in which the money is in red envelope 1 ($u \approx u(3, 1)$), and in that situation, you would not have had the freedom to have been in situation v, so $v \nsim u(3, 1)$ and hence $v \sim_K u$.

Of course, had the money been in envelope 1, and you in $u(3, 1)$ there is *another* situation you could have been in which would have been better for you, namely

$u(1,1)$, and this is why your decision was stupid. Unknown limitations on our freedom to choose serve as *ceteris paribus* conditions on our preferences. For any decision u, we can identify the operating limiting condition as (u): within this class, you are free to make alternative choices, and so comparisons are possible. As argued above, for an alternative v to be judged as better than your actual decision u, you are limited to comparing what you would know about v to what you know about u. That means that you must not only compare u to v but also other possibilities u' and v' that you cannot distinguish from u and v ($u \approx u'$ and $v \approx v'$). But such a comparison is only needed when u' and v' are also genuine alternatives, that is when $u' \sim v'$.

This form of reasoning can be captured by restricting our comparisons to situations related by the $\sim$ relations, so defining the relation of *free preference* to be

$$u \leq_F v \text{ iff } u' \leq v' \text{ for all } u' \approx u \text{ and all } v' \approx v \text{ such that } u' \sim v'.$$

Our final expression of a norm of rational decisions that incorporates the effects of ignorance, limitations of freedom and preference is therefore the following:

Free Known Decisions

If u is rational there is no $v \sim u$ such that $v >_F u$.

Free Known Decisions is a generalisation of all previously mentioned norms.

3.4 *A priori* and *a posteriori* rationality

A further issue to be considered is the distinction between *a priori* and *a posteriori* norms of decision making. Typically, as we have seen, a decision takes place in conditions of uncertainty. You do not know what the result of your decision will be because it depends on contextual factors over which you have neither knowledge or control. Applying the kind of reasoning outlined above, you will view any decision u that fails to obey the norm of Free Known Decisions as irrational. In this case we will say that the decision is *a priori irrational* because its irrationality is based only on what is known in advance of the decision being made. Moreover, to be precise, we will talk of $\leq_F$ as *a priori free preference*, and then define

u is *a priori rational* iff there is no $v \sim u$ such that $v >_F u$.

Thus the norm of Free Known Decisions is just that *a priori* rationality is necessary for rationality. We will allow for there being further conditions for rational decisions that are not considered here, so that *a priori* considerations may not be sufficient. In particular, in the social setting, which we consider in the next section, consideration of the rationality of other agents may be relevant.

One can also look at decisions from the perspective of what is known after the decisions, looking back on what one chose and asking whether one could have done any better, after gaining information about those contextual factors that influenced the outcome. This is obviously a graded concept but here we will confine our attention to the limiting case: complete knowledge of the circumstances of one's decision. This is modelled by the relation $\approx'=\approx \cap \sim$. Or, in other words, after discovering the limitations of one's freedom to choose, one can distinguish between between u and v when $u \nsim v$. We therefore define *a posteriori free preference* to be

$u \leq_{F'} v$ iff $u' \leq v'$ for all $u' \approx u$ and all $v' \approx v$ such that $u \sim u' \sim v' \sim v$.

Moreover, we will talk of *a posteriori* rationality, which can be defined as

u is *a posteriori rational* iff there is no $v \sim u$ such that $v >_{F'} u$.

In a deterministic frame, this can be simplified:

Lemma 1 (Deterministic *a posteriori* rationality)**.** *In a deterministic frame, u is* a posteriori *rational iff there is no $v \sim u$ such that $v > u$.*

Proof. In a deterministic frame, $u' \approx u$ and $v' \approx v$ and $u \sim u' \sim v' \sim v$ if and only if $u' = u$ and $v' = v$ and $u \sim v$. □

Unlike *a priori* rationality, *a posteriori* rationality is not necessary for rationality since it makes use of information that is not available to you at the time you make the decision. Nonetheless, it captures the goal of decision-making: to choose so that you could not have (in fact) done better by making another choice, given the context.

3.5 Constraints on the preference order

Finally, we come to the much-discussed issue of whether $\leq$ should be transitive and the less discussed, but also commonly assumption, that it should be reflexive. Another issue is whether your preferences should be total, i.e. for every u and v, either $u \leq v$ or $v \leq u$.

First, reflexivity. As noted above, we already have that $<$ is irreflexive, by definition. There are therefore two possibilities for any situation u: either $u \bowtie u$ or $u\#u$. If we think of the situations as 'possible worlds' in which every aspect of your life is completely specified, then we could rule out the second case. But if the situations we model have a variety of different consequences for you, it is quite possible that you are in a state of conflict rather than indifference.

Next, transitivity. The usual argument that your preferences should be transitive is that otherwise, you could be faced with a series of free decisions after which you end up with less desirable goods than you started with. If, for example, you prefer A to B and B to C and C to A, then you would trade A for B and then B for C, with the effect that you now have C instead of A even though you prefer A to C. So let u be your decision to refuse the first trade, keeping A, v_1 your decision to accept the first and refuse the second, so keeping B, and v_2 your decision to accept both trades, resulting in having C. Then $u < v_1 < v_2$ but $v_2 < u$, and all three situations are within the same $\sim$-class. Assuming that you have complete knowledge of the consequences of these trades, so that $[u] = \{u\}$, $[v_1] = \{v_1\}$, and $[v_2] = \{v_2\}$, all three decisions are *a priori* irrational. The norm of Free Known Decisions applies but is somewhat unhelpful since there is no decision you can make that is rational. A proper account of transitivity should therefore consider preference *change*, which is beyond the scope of the present analysis.

Lastly, totality. As with reflexivity, totality is only plausible if the situations are completely specified. This is not an assumption we wish to make, as for many applications to ordinary decision-making, it is simply false. Moreover, there is a technical reason for avoiding this assumption. Given any decision frame $F = \langle W, \sim, \approx, \leq \rangle$, we can remove the indeterminacy in F by taking a quotient $F' = \langle W', \sim', \approx', \leq' \rangle$, in which W' is the set of equivalence classes $(u)[u]$ for each $u \in W$, with $\sim'$ and $\approx'$ defined in the obvious way[4] and $(u)[u] \leq (v)[v]$ iff $u' \leq v'$ for all $u' \in (u)[u]$ and $v' \in (v)[v]$. The resulting frame F' is deterministic and

[4] $(u)[u] \sim' (v)[v]$ iff $(u) = (v)$, and $(u)[u] \approx' (v)[v]$ iff $[u] = [v]$.

is just as good as F for the purposes of practical decision making: u is *a priori* rational in F iff $(u)[u]$ is *a priori* rational in F'. Yet, clearly, conflict can arise in comparing $(u)[u]$ with $(v)[v]$, even if the underlying order $\leq$ is total.

Nonetheless, in comparing our account with game theory (in Section 5.3 and 6), it will be necessary to consider frames in which all three conditions hold. So we say that a decision frame is *linear* iff $\leq$ is a linear order (reflexive, transitive, and total).

4 Social Decision Frames

So far we have only considered the single-agent case. In moving to a social setting, nothing changes from the perspective of individual rationality, but there is the possibility of systematic relationships between agents that can have consequences for decision-making. We define a *social decision frame* to be a structure $F = \langle W, A, \sim, \approx, \leq \rangle$ in which for each $a \in A$, $F_a = \langle W, \sim_a, \approx_a, \leq_a \rangle$ is a decision frame. We write $[u]_a$ for the $\approx_a$ equivalence class of u and $(u)_a$ for its $\sim_a$ equivalence class. We also write $\overline{a}$ for $A \setminus \{a\}$, the set of agents other than a.

Borrowing from epistemic logic, the *common knowledge* of a group of agents $G \subseteq A$ is defined by

$$\approx_G = [\bigcup_{a \in G} \approx_a]^*$$

i.e., the transitive closure of the the union of the indistinguishability relations of all agents in G. Put another way, this is the greatest amount of knowledge compatible with the restriction that it is possessed by every agent in the group. We write $[u]_G$ for the $\approx_G$ equivalence class of u, so that $X \subseteq W$ is common knowledge among members of G iff $[u]_G \subseteq X$.

Analogously, the *joint freedom* of the group is represented by the relation

$$\sim_G = [\bigcup_{a \in G} \sim_a]^*$$

We write $(u)_G$ for the $\sim_G$-equivalence class of u, which is the the set of situations v that could have occurred, given the capacities of all the agents in G, had they chosen otherwise. Our notion of independence extends to groups, so that a state-of-affairs X is independent of the group G iff $(u)_G \subseteq X$ for all $u \in X$. This

holds when the group G lacks the (joint) power to influence whether or not X occurs.

This idea can then be applied to other agent's decisions. We say that agent a's decision in situation u is *independent* of a group G iff $[u]_a$ is independent of G, i.e., iff $(u)_G \subseteq [u]_a$. We can also interpret this as saying that a is ignorant of the all of the decisions of agents in G. When a's decisions are independent of all other agents $(A \setminus \{a\})$ in situation u, we say that a is *isolated* in u, and when every agent is isolated in every situation, we say that the frame F is *isolated*. Isolation is a common assumption of game theory, as will be shown in Section 5.3.

An important issue that arises for social decision-making is the possibility of ignorance about other agent's capacities and preferences. Within game theory, this kind of ignorance is called 'incomplete information' and it has been known since Harsanyi (1967/68) that its effect can be modelled by introducing a new agent, 'Nature', with the freedom to choose between different games, and about whose choice agents may be ignorant. Again, we will postpone a proper treatment of this issue here, noting only that this kind of ignorance is possible in a social decision frame, without such a manoeuvre. The reason for this is that game theory generated the space of possible decisions from the capacities of agents. We can capture this property in the general case with two additional concepts.

Firstly, if in situation u, agent a had the freedom to be in situation v ($u \sim_a v$) and agent b had the freedom to be in situation w ($u \sim_b w$) , then there is another situation u' in which agent b had the freedom to be in v ($u' \sim_b v$) and agent a had the freedom to be in w ($u' \sim_a w$). The thought is that whatever choice a could make in situation u so as to make v possible is a freedom of a and whatever choice b could make in situation u so as to make w possible is a freedom of b, so it should be possible for the two agents to exercise those freedom simultaneously. In this case, we say that the two agents have 'unordered' freedoms. More precisely, we say that a group G of agents is *unordered* iff $\sim_a;\sim_b = \sim_b;\sim_a$ for all $a, b \in G$, and a frame F is unordered if the group A of all agents is unordered. The following lemma about unordered agents will prove useful.

Lemma 2 (Order of Choices)**.** *If G is an unordered group of agents and $H \subset G$, then*

$$\sim_G = \sim_H;\sim_{(G \setminus H)}$$

Proof. If $u \sim_G v$ then there is an ordering $a_1, \ldots, a_n$ of G such that $u \sim_{a_1}; \ldots; \sim_{a_n} v$. Choose a permutation $b_1, \ldots, b_m, c_1, \ldots, c_p$ of this ordering such that the b's

are all in H and the c's are all in $G \setminus H$. Since G is an unordered group, $u \sim_{b_1}; \ldots; \sim_{b_m}; \sim_{c_1}; \ldots; \sim_{c_p} v$, and hence $u \sim_H; \sim_{(G \setminus H)} v$. The converse follows because $H, G \setminus H \subseteq G$. □

Secondly, we say that the frame F is *connected* if $u \sim_A v$ for all $u, v \in W$, or, equivalently, that $(u)_A = W$ for all $u \in W$. For F to be connected, there can be no external source of possibility in addition to the joint capacity of the set of all agents. Together, the assumptions that F is unordered and connected will be shown to be necessary assumptions in the characterisation of strategy games (see Theorem 1). In particular, the combination of the above mentioned properties of frames enable a reduction of knowledge to freedom:

Lemma 3 (Reduction of Knowledge)**.** *In a connected isolated unordered deterministic frame,* $[u]_a = (v)_{\overline{a}}$ *for all* $v \approx_a u$.

Proof. That $(v)_{\overline{a}} \subseteq [u]_a$ is just a restatement of the isolation of the frame. To show that $[u]_a \subseteq (v)_{\overline{a}}$, suppose that $w \approx_a u$. By connectedness of the frame, $w \sim_A v$ and so by Lemma 2, there is a w' such that $w \sim_a w' \sim_{\overline{a}} v$. But then $w' \approx_a v$, since the frame is isolated. We assumed that $w \approx_a u$ and $v \approx_a u$, so $w \approx_a w'$ by transitivity. Determinism then ensures that $w = w'$ and so $w \sim_{\overline{a}} v$, as required. □

Lemma 4 (Reduction of Freedom)**.** *In a connected isolated unordered deterministic frame,* $(u)_a = \bigcap_{b \in \overline{a}} [v]_a$ *for all* $v \sim_a u$.

Proof. By Lemma 3, it is enough to show $(u)_a = \bigcap_{b \in \overline{a}} (u)_{\overline{b}}$. Firstly, $(u)_a \subseteq (u)_{\overline{b}}$ for each $b \neq a$ because $\{a\} \subseteq \overline{b} = A \setminus \{b\}$. For the converse inclusion, suppose $v \in \bigcap_{b \in \overline{a}} (u)_{\overline{b}}$, that is, for each $b \neq a$ such that $v \sim_{\overline{b}} u$. So by Lemma 2, there is some w such that $v \sim_a w \sim_{\overline{ab}} u$. Then there is a path $w_0 = w, \ldots, w_i, \ldots, w_m = u$ and agents $b_0, \ldots, b_{m-1}$ $(\neq a, b)$ such that $w_i \sim_{b_i} w_{i+1}$ and for $i \leq i \leq m$. Without loss of generality, by Lemma 2, we can assume that $b_0, \ldots, b_{m-1}$ are all distinct. Now we prove $w = w_i$ by induction, so that $w = w_m = u$ and hence $v \sim_a u$, as required. The base case $w = w_0$ holds by definition. Suppose $w = w_i$, so $v \sim_a w_i \sim_{b_i} w_{i+1}$. But also $w_{i+1} \sim_{\overline{b_i}} w_m = u$ because b_i does not occur again in the list $b_{i+1}, \ldots, b_{m-1}$. Now $b_i \neq a$, so $v \sim_{\overline{b_i}} u$ since $v \in \bigcap_{b \in \overline{a}} (u)_{\overline{b}}$. Thus $w = w_i \sim_{\overline{b_i}} v \sim_{\overline{b_i}} u \sim_{\overline{b_i}} w_{i+1}$ and hence $w_i \sim_{\overline{b_i}} w_{i+1}$. Then, by isolation, $w \approx_{b_i} w_{i+1}$ and by determinism, $w = w_{i+1}$. □

Finally, the concept of value can be extended to the social setting in a straightforward way. We say that a subset of W is a *value* in a social decision frame iff

it is a value for one of the agent, ie., a $\bowtie_a$-equivalence class. The *value size* of a social decision frame is the cardinality of the set of its values.

5 Strategic games

We will show that standard models of strategic games used in game theory can be accommodated in our framework. Consider a game in which each player $a_i$from a finite set $\{a_1, \ldots, a_n\}$ of players has a finite set S_i of possible strategies. We will call these *pure strategies* to foreshadow the 'mixed strategy' case to be considered shortly. A *pure strategy profile* is a sequence $s_1, \ldots, s_n$ with $s_i \in S_i$ Each pure strategy profile determines an outcome $O(s_1, \ldots, s_n)$ in a set Ω of outcomes. Each player a assigns a real-valued *utility* $u_a(\omega)$ to each outcome $\omega \in \Omega$.

5.1 Pure strategy games

In a pure strategy game, the contextual factors influencing one's decisions are completely determined by the strategies of the other players. So once you have chosen your strategy, all the relevant facts regarding your decision have been fixed. We therefore define the set W of social decision states to be the set of pure strategy profiles, namely

$$W = S_1 \times \ldots \times S_n$$

In a game situation, an agent's preferences are assumed to be determined entirely by the utility of the game outcome. If one has other preferences, such as wanting your daughter to beat you at chess at least some of the time, this makes the decision situation different from the one specified by the rules of the game.[5] The *game utility* $U_a(s_1, \ldots, s_n)$ of profile $s_1, \ldots, s_n$ to player a is given by

$$U_a(s_1, \ldots, s_n) = u_a(O(s_1, \ldots, s_n))$$

[5]Of course, games are used to model many situations that are not formal games, but in all cases, the model is only as good as the utility function, and so we must assume this is accurate.

and this defines a preference order by [6]

$$w \leq_a v \quad \text{iff} \quad U_a(w) \leq U_a(v)$$

Each player is assumed to be aware only of her own actions, so define

$$w \approx_{a_i} v \text{ iff } w_i = v_i$$

Yet each player is able to choose freely between pure strategies, and so between any two outcomes in which the pure strategies of all other agents is fixed. So we define

$$w \sim_{a_i} v \text{ iff } w_j = v_j \text{ for all } j \neq i$$

Denote the resulting social decision frame as $P(A, S, \Omega, O, u)$. It is clearly transitive. Any model constructed in this way we call a *pure strategy game model*.

5.2 Mixed strategy games

Now consider 'mixed' strategies, in which player a_i makes a non-deterministic choice of strategy, modelled by a probability function $\delta\colon S_i \to [0, 1]$. If δ takes values in $\{0, 1\}$ we say that it is a *deterministic* strategy.[7] Let Δ_i be the set of a_i's mixed strategies. A *mixed strategy profile* is a sequence $\delta_1, \ldots, \delta_n$ such that δ_i is in Δ_i.

Generalising the pure strategy case, we define a social decision state to be a mixed strategy profile, setting

$$W = \Delta_1 \times \ldots \times \Delta_n$$

It may seem a little strange to define the decision state to be such an obviously indeterminate entity. The obvious objection to doing this is that it does not determine an outcome. But in decision situations in general, it can be very difficult to determine outcomes. They are only relevant to our decisions insofar as our beliefs about them allow us to determine a preference relation between

[6] It is not normal to distinguish between game utility and utility simpliciter but we do so for theoretical reasons, explained below.

[7] Deterministic strategies are equivalent to pure strategies, in a sense to be made clear below.

decision states. And the same holds in mixed-strategy games. The game utility $U_a(\delta_1, \ldots, \delta_n)$ of a profile $\delta_1, \ldots, \delta_n$ to agent a is computed as the expected utility

$$U_a(\delta_1, \ldots, \delta_n) = \sum_{s \in S_1 \times \ldots \times S_n} u_a(O(s_1, \ldots, s_n)) \prod_{i=1}^{n} \delta_i(s_i)$$

Recall that $S_1 \times \ldots \times S_n$ is the set of pure strategy profiles, and so for any $s \in S_1 \times \ldots \times S_n$, the probability that the actions modelled by s are actually carried out by agents with mixed strategy profile δ is $\prod_{i=1}^{n} \delta_i(s_i)$, assuming that these actions are independent.[8] Using this more sophisticated account of game utility, we define the relations $\leq$, $\sim$ and $\approx$ as before and denote the resulting *mixed strategy* social decision frame as $M(A, S, \Omega, O, u)$.

5.3 Games as social decision frames

Examining the above constructions, we see that there are many details that are irrelevant to our construction, so we will simplify a bit. We use the term 'strategy' to generalise over pure and mixed strategies. Given any finite set A, sets D_a of strategies for each $a \in A$, let $W(A, D)$ be the set of social decision situations given above:

$$W(A, D) = \prod_{a \in A} D_a$$

In other words, each social decision situation is just a strategy profile. Now let U_a be the real-valued function with domain $W(A, D)$ that represents a's utilities. Define the relations $\leq$, $\sim$ and $\approx$ on $W(A, D)$ in the familiar way:

$w \leq_a v$ iff $U_a(w) \leq U_a(v)$
$w \approx_a v$ iff $w_a = v_a$
$w \sim_a v$ iff $w_b = v_b$ for all $b \neq a$ in A

Call the resulting social decision frame $G(A, D, U)$ a *strategic game frame*.

Lemma 5 (Game Frames)**.** *Every pure or mixed strategy game frame is a strategic game frame.*

[8] Since A and each S_i is finite, the sum and product here are both well-defined, despite the fact that the set of game states W is uncountable.

Proof. Suppose $P(A, S, \Omega, O, u)$ is a pure strategy game frame. When U is defined by $U_a(w) = u_a(O(w))$ and $D_{a_i} = S_i$ then clearly $P(A, S, \Omega, O, u) = G(A, D, U)$.

Now suppose $M(A, S, \Omega, O, u)$ is a mixed strategy game frame. Let Δ_i be the set of a_i's mixed strategies and let U be the game utility function defined in the construction of $M(A, S, \Omega, O, u)$. Now let $D_{a_i} = \Delta_i$. It is easy to check that $M(A, S, \Omega, O, u) = G(A, D, U)$. □

A useful notation for discussing strategies in game frames is the following. Given $w \in W(A, D)$, $a \in A$ and $d \in D$, let $w[^a_d]$ be the strategy profile defined by

$$w[^a_d]_b = \begin{cases} d & \text{if } a = b \\ w_b & \text{otherwise} \end{cases}$$

noting that $w \sim_a v$ iff $w = v[^a_{w_a}]$. The notation helps us to prove the following small but useful lemma:

Lemma 6 (Joint Freedom in Game Frames)**.** *In a strategic game frame, $G(A, D, U)$, for any $G \subseteq A$ and $u, v \in W(A, D)$,*

$$u \sim_G v \quad \textit{iff} \quad u_b = v_b \textit{ for all } b \notin G.$$

Proof. Suppose $u_b = v_b$ for all $b \notin G$. G is finite, so let $G = \{a_0, ..., a_{n-1}\}$ and define $u_0, ..., u_n$ inductively by: $u_0 = u$ and $u_{i+1} = u_i[^{a_i}_{v_{a_i}}]$. Then $u_i \sim_{a_i} u_{i+1}$ and

$$u_{nb} = u[^{a_0}_{v_{a_0}}]...[^{a_{n-1}}_{v_{a_{n-1}}}]_b = \begin{cases} v_{a_i} & \text{if } b = a_i \\ u_b & \text{if } b \notin G \end{cases} = v_b$$

So $u_n = v$; hence $u \sim_G v$. Conversly, suppose $u \sim_G v$, so there are $u_0, ..., u_n$ and $a_0, ..., a_{n-1} \in G$ such that $a_0 = u$, $u_n = v$ and $u_i \sim_{a_i} u_{i+1}$. But then $u_{ib} = u_{i+1b}$ for all $b \neq a_i$, and so for all $b \notin G$. Thus $u_b = u_{0b} = u_{nb} = v_b$ for all $b \notin G$. □

There are many social decision frames that are not strategic game frames, so we can ask what properties of social decision frames characterise the class of game frames. The answer is given by the following representation theorem.

Theorem 1 (Representation)**.** *A social decision frame is isomorphic to a strategic game frame iff it is connected, isolated, unordered, deterministic, linear, with a value size* $\leq 2^{\aleph_0}$ *and a finite number of agents.*

The theorem follows from the following lemmas.

Lemma 7 (Left-to-right)**.** *Every strategic game frame is connected, isolated, unordered, deterministic, linear, with a value size* $\leq 2^{\aleph_0}$ *and has a finite number of agents.*

Proof. The set A of agents in a strategic game frame $G(A, D, U)$ is finite by definition. And the function U_a maps $W(A, D)$ to $\mathbb{R}$, and so the a cannot have more values that the cardinality of $\mathbb{R}$, namely $2^{\aleph_0}$.

Determinism: Suppose that $u \sim_a v$ and $u \approx_a v$. Then $v_b = u_b$ for all $b \neq a$ in A (definition of $\sim_a$) and $v_a = u_a$ (definition of $\approx_a$). So $u = v$.

Isolation: We need to show that $(u)_{\overline{a}} \subseteq [u]_a$. So suppose that $v \sim_{\overline{a}} u$. Then by Lemma 6, $v_b = u_b$ for all $b \notin \overline{a}$. But $a \notin \overline{a}$, so $v_a = u_a$. Hence $v \approx_a u$, as required.

Connectedness: For any $u, v \in W$, by Lemma 6, $u \sim_A v$ holds vacuously.

Unordered: Suppose $a \neq b$ and $u \sim_a w \sim_b v$. Then by Lemma 6, $w = u[^a_{w_a}]$ and $v = w[^b_{v_b}]$, so $v = u[^a_{w_a}][^b_{v_b}]$. But $u[^a_{w_a}][^b_{v_b}] = u[^b_{v_b}][^a_{w_a}]$ because $a \neq b$. So let $w' = u[^b_{v_b}]$. Then $u \sim_b w' \sim_a v$, as required.

□

Lemma 8 (Existence)**.** *In any connected isolated unordered social decision frame, for any function* $d\colon A \to W$*, there is a* $u \in W$ *such that* $da \approx_a u$ *for all* $a \in A$*.*

Proof. Suppose for contradiction that $\bigcap_{a\in A}[da]_a = \emptyset$. Let G be a minimal subset of A such that $\bigcap_{a\in G}[da]_a = \emptyset$. Clearly $G \neq \emptyset$ so let $b \in G$ and $H = G \setminus \{b\}$. Then $\bigcap_{a\in H}[da]_a \neq \emptyset$, so there is a v such that $da \approx_a v$ for all $a \in H$. By connectedness, $db \sim_A v$, and by Lemma 2, there is a u such that $db \sim_H u \sim_{A\setminus H} v$. But as $H \subseteq \overline{b}$ and $A \setminus H \subseteq \overline{a}$ for each $a \in H$, we get that $db \sim_{\overline{b}} u \sim_{\overline{a}} v$, and by isolation, that $db \approx_b u \approx_a v$, for all $a \in H$. By the choice of v, also $da \approx_a v$ and so $da \approx_a u$, for all $a \in H$. Hence $u \in \bigcap_{a_i\in G}[da]_a$, which is a contradiction. □

Lemma 9 (Uniqueness)**.** *In any connected isolated unordered deterministic social decision frame with a finite number of agents, for any function* $d\colon A \to W$*, if* $da \approx_a u$ *and* $da \approx_a v$ *for all* $a \in A$*, then* $u = v$*.*

Proof. Let the finite set of agents be $A = \{a_0, ..., a_n\}$ and define A_i inductively as follows: $A_0 = A$ and $A_{i+1} = A_i \setminus \{a_i\}$, so that $A_n = \emptyset$. We will show by induction that $u \sim_{A_i \setminus \{a\}} v$ for all $a \in A$. Then, since $A_n \setminus \{a\} = \emptyset$, $u = v$, as required.

For the base case, for any $a \in A$, $u \approx_a da \approx_a v$ so $u \approx_a v$, and so by Lemma 3, $u \approx_{\overline{a}} v$. But $A_0 \setminus \{a\} = \overline{a}$ and so we are done.

For the inductive case, suppose that $u \sim_{A_i \setminus \{a\}} v$ for all $a \in A$. Let $a \in A$. If $a = a_i$ then $A_{i+1} \setminus \{a\} = A_i \setminus \{a\}$, by definition of A_{i+1} and so $u \sim_{A_{i+1} \setminus \{a\}} v$, as required. So suppose $a \neq a_i$. We have that $u \sim_{A_i \setminus \{a\}} v$, so by Lemma 2, observing that $(A_i \setminus \{a\}) \setminus (A_{i+1} \setminus \{a\}) = \{a_i\}$, there is a w such that $u \sim_{A_{i+1} \setminus \{a\}} w \sim_{a_i} v$. But $A_{i+1} \setminus \{a\} \subseteq \overline{a_i}$ so by isolation, $u \approx_{a_i} w$. Also, as above, $u \approx_{a_i} da_i \approx_{a_i} v$, so by transitivity, $w \approx_{a_i} v$. But then, by determinism, $w = v$ and so $u \sim_{A_{i+1} \setminus \{a\}} v$, as required. □

Proof of Theorem 1. Firstly, the direction from left to right is a consequence of Lemma 7. Now, suppose we have a social decision frame $F = \langle W, A, \sim, \approx, \leq \rangle$ that is is connected, isolated, unordered, deterministic, linear, with a value size $\leq 2^{\aleph_0}$ and has a finite number of agents. The strategies of each agent a will be given by

$$D_a = \{[u]_a\} u \in W$$

Since F has a value size $\leq 2^{\aleph_0}$ there is a strong $\leq$-homomorphism $h \colon W \to \mathbb{R}$, i.e. v is such that $u \leq v$ iff $hu \leq hv$.[9] We lift h to a function $h_{\min} \colon \text{pow } W \to \mathbb{R}$ by defining

$$h_{\min}(X) = \min\{h(u)\} u \in X$$

Next we define the utility function $U \colon W(A, D) \to \mathbb{R}$ by

$$U(w) \leq U(w') \quad \text{iff} \quad h_{\min} \bigcap_{a \in A} w_a \leq h_{\min} \bigcap_{a \in A} w'_a$$

This completes our construction of a strategic game frame $G(A, D, U)$. We must now show that $G(A, D, U)$ is isomorphic to F. The isomorphism $f \colon W(A, D) \to W$ is defined by

$$\bigcap_{a \in A} w_a = \{f(w)\}$$

[9] The cardinality of W itself may be much bigger, but the quotient under $\bowtie$ gives us a linear order of size no bigger than $\mathbb{R}$, which can therefore be isomorphically embedded in $\langle \mathbb{R}, \leq \rangle$, so giving a strong homomorphism from W into $\mathbb{R}$.

That f is well-defined follows from Lemmas 8 and 9, which together imply that $\bigcap_{a\in A} w_a$ is a singleton. To see that it is a bijection, we define $g\colon W \to W(A,D)$ by $g(u)_a = [u]_a$ for each $a \in A$, and claim that this is the inverse of f. [10]

Finally, to see that f (equivalently g) is an isomorphism, we need only check the following sequences of equivalences:

$u \approx_a v$	
$[u]_a = [v]_a$	Defn. $[u]_a$
$g(u)_a = g(v)_a$	Defn. g
$g(u) \approx g(v)$	Defn. $\approx_a$ in $G(A,D,U)$.

$u \sim_a v$	
$u \in (v)_a$	Defn. $(u)_a$
$u \in \bigcap_{b\in\overline{a}}[v]_b$	Lemma 4
$u \in [v]_b$ for all $b \neq a$	Defn. $\overline{a}$
$[u]_b = [v]_b$ for all $b \neq a$	Defn. $[v]_b$
$g(u)_b = g(v)_b$ for all $b \neq a$	Defn. g
$g(u) \sim_a g(v)$	Defn. $\sim_a$ in $G(A,D,U)$.

[10] For $fg(u) = u$ note that, by definition, $\{fg(u)\} = \bigcap_{a\in A} g(u)_a = \bigcap_{a\in A}[u]_a = \{u\}$. For $gf(w)_a = w_a$, also by definition $gf(w)_a = [f(w)]_a = [\bigcap_{a\in A} w_a]_a = w_a$.

$w \leq v$	
$U(w) \leq U(v)$	Def. $\leq$ in $G(A, D, U)$.
$h_{\min} \bigcap_{a \in A} w_a \leq h_{\min} \bigcap_{a \in A} v_a$	Defn. U
$h_{\min}\{f(w)\} \leq h_{\min}\{f(v)\}$	Defn. f
$hf(w) \leq hf(v)$	Defn. $h_{\min}$
$f(w) \leq f(v)$	h strong homomorphism

□

6 Norms of rationality in game theory

Having established that strategic games (with either pure or mixed strategies) can be regarded as a special class of social decision frames, we will now apply the norms of decision making from Section 3 to this class and show that they correspond precisely to various concepts from game theory. Our present purpose is merely to confirm that our definitions in the general case make good sense in a known special case. But we are also making the first step in a larger project of using social decision frames to model a range of game-theoretic concepts.

6.1 Dominated strategies

Given a strategic game frame $G(A, D, U)$, agent a's strategy $d \in D_a$ is *dominated* by another strategy $d' \in D_a$ iff

1. d' is sure to be at least as good as d: $w[^a_{d'}] \geq_a w[^a_d]$ for all $w \in W(A, D)$, and
2. d' may be better than d: $w[^a_{d'}] >_a w[^a_d]$ for some $w \in W(A, D)$.

Situations in which an agent plays a dominated strategy are deemed irrational because she would risk nothing by playing differently, and has a possibility of doing better. That this definition correctly applies to standard accounts of game theory is confirmed by the following easy lemma.

Lemma 10 (Domination in Game Frames)**.** *Given a pure strategy game* $P(A, S, \Omega, O, u)$*, a pure strategy* $s_i \in S_i$ *is dominated by some pure strategy* $s_i' \in S_i$ *iff for all pure strategy profiles* $s_1, \ldots, s_n$ *and* $s_1', \ldots, s_n'$ *in which* $s_j = s_j'$ *for all* $j \neq i$*,*

$$U_{a_i}(O(s_1', \ldots, s_n')) \geq U_{a_i}(O(s_1, \ldots, s_n))$$

and for some pure strategy profiles $s_1, \ldots, s_n$ *and* $s_1', \ldots, s_n'$ *in which* $s_j = s_j'$ *for all* $j \neq i$*,*

$$U_{a_i}(O(s_1', \ldots, s_n')) > U_{a_i}(O(s_1, \ldots, s_n))$$

Similarly for mixed strategy games.

Note that game theorists typically distinguish between a weak and strict form of domination; this is the weak version. The distinction is not so important for judging whether a strategy is rational, because strict domination implies weak domination. When we judge a strategy to be irrational because it is dominated, the sense of 'rational' is clearly that of *a priori* rationality. A player can tell that the strategy is dominated based only on her knowledge of her own capacities and preferences and of the capacities (but not necessarily the preferences) of other players. To make the connection between *a priori* rationality and domination, we need a technical lemma:

Lemma 11 (Free Preference in Game Frames)**.** *In any game frame* $G(A, D, U)$*,* $u <_{aF} v$ *iff*

1. $w[^a_{u_a}] \leq_a w[^a_{v_a}]$ *for all* $w \in W(A, D)$*.*
2. $w[^a_{u_a}] <_a w[^a_{v_a}]$ *for some* $w \in W(A, D)$*.*

Proof. It is enough to prove that $u \leq_{aF} v$ iff $w[^a_{u_a}] \leq_a w[^a_{v_a}]$ for all $w \in W$. (The rest follows from the linearity of $\leq_a$ in a game frame and the definition of $u <_{aF} v$ as $u \leq_{aF} v$ and $v \not\leq_{aF} u$.) By definition, $u \leq_{aF} v$ iff $u' \leq_a v'$ for all $u' \approx_a u$ and all $v' \approx_a v$ such that $u' \sim_a v'$. But the following are equivalent:

$$u' \approx_a u \text{ and } v' \approx_a v \text{ and } u' \sim_a v'$$

$$u'_a = u_a \text{ and } v'_a = v_a \text{ and } u'_b = v'_b \text{ for all } b \neq a$$

$$u' = w[^a_{u_a}] \text{ and } v' = w[^a_{v_a}] \text{ for some } w \in W(A, D)$$

Moreover, $w[^a_{u_a}] \approx_a u$ and $w[^a_{v_a}] \approx_a v$ and $w[^a_{u_a}] \sim_a w[^a_{v_a}]$ for all $w \in W(A, D)$. So $u \leq_{aF} v$ iff $w[^a_{u_a}] \leq_a w[^a_{v_a}]$ for all $w \in W(A, D)$. □

Theorem 2 (Dominated Strategies)**.** *$d \in D_a$ is a dominated strategy iff no situation w for which $w_a = d$ is* a priori *rational for a.*

Proof. Suppose $w_a = d$ and $v_a = d'$. Then from the definition of 'dominates' and Lemma 11, d is dominated by d' iff $w <_{aF} v$. Thus d is a dominated strategy iff for each w with $w_a = d$ there is a v such that $w <_{aF} v$, i.e., iff there is no *a priori* rational w with $w_a = d$. □

6.2 Nash equilibrium

The central concept of the theory of strategic games is that of Nash equilibrium, which is defined in terms of the players having played in such a way that they could not have done better, given the moves of the other players. This is an *a posteriori* concept. We imagine the players evaluating their moves after they know what has happened. More precisely, given a strategic game frame $G(A, D, U)$, a situation $w \in W(A, D)$ is a *best response* for agent a iff there is no strategy $d \in D_a$ such that

$$w <_a w[^a_d]$$

It is a *Nash equilibrium* iff it is a best response for all agents.[11]

That this definition correctly applies to standard definitions of 'best response' in game theory is confirmed by the following easy lemma.

Lemma 12 (Best Response in Game Frames)**.** *Given a pure strategy game $P(A, S, \Omega, O, u)$, a pure strategy profile $s_1, \ldots, s_n$ is a best response for player a_i iff she can not have a strictly better outcome by taking another pure strategy, that is, there is no $s \in S_{a_i}$ such that*

$$U_{a_i}(O(s_1, \ldots, s_n)) < U_{a_i}(O(s_1, \ldots, s'_n))$$

where $s'_i = s$ and $s'_j = s_i$ for all $j \neq i$. Likewise, given a mixed strategy game $M(A, S, \Omega, O, u)$, a mixed strategy profile $\delta_1, \ldots, \delta_n$ is a best response for player a_i

[11] The linearity of the preference relation allows us to give an alternative definition but equivalent definition of 'best response' and hence Nash equilibrium: '$w \geq w[^a_d]$ for every strategy $d \in D_a$'.

iff there is no $\delta \in \Delta_i$ such that the expected utility of playing with δ would have produced a better outcome:

$$U_{a_i}(O(\delta_1, \ldots, \delta_n)) < U_{a_i}(O(\delta'_1, \ldots, \delta'_n))$$

where $\delta'_i = \delta$ and $\delta'_j = \delta_i$ for all $j \neq i$.

We can now see that 'best response' corresponds precisely to *a posteriori* rationality.

Theorem 3 (Best Response)**.** *u is a best response for player a iff u is* a posteriori *rational for a.*

Proof. The following are equivalent:

u is a best response for player a.	
There is no strategy $d \in D_a$ such that $u[^a_d] >_a u$.	Defn. 'best response'
There is no $v \sim_a u$ such that $v >_a u$.	Defn. $\sim_a$
iff u is *a posteriori rational* for a.	by Lemma 1

□

Corollary 1 (Nash)**.** *A situation w is a Nash equilibrium iff it is* a posteriori *rational for all agents.*

7 Conclusion

The main contribution of the present paper is the attempt to provide a sharp distinction between descriptive and normative aspects of decision making in both the individual and social settings, to justify certain norms in their own terms, and to confirm their correctness by applying them to game theory. The emphasis on a relational analysis of the fundamental concepts (knowledge,

freedom and preference) avoids the need for an explicit account of action. Our main result is Theorem 1, which gives a useful characterisation of the assumptions of strategic game theory.

This is part of an ongoing project. A hidden agenda is the use of hybrid modal logic to axiomatise the class of strategic game frames, and provide a language for expressing the concepts of *a priori* and *a posteriori* rationality, building on Cui et al. (2009) and Seligman (2010). We also intend to use this framework to explore other concepts in game theory and other games, such as those of imperfect information, incomplete games, sequential games, etc.

Acknowledgements Guo is supported by the National Social Science Foundation of China (09CZX033), the Foundation for Humanities and Social Sciences of the Ministry of Education of China (08JC72040002), the Fundamental Research Funds of Southwest University (SWU0909512) and the Key Project of the Chongqing Key Research Base in Humanities and Social Sciences titled 'A Study on the Extensions of Dynamic Epistemic Logic' (11SKB16).

The research for this paper was conducted during Seligman's Academic Leave from the University of Auckland, New Zealand, in 2010 and a subsequent visit to South West University during May 2012. He is grateful to the University of Auckland for allowing him to reorganise teaching duties, etc, to make this happen, and to South West University for their kind invitation and support.

References

R. Bonanno. A syntactic approach to rationality in games with ordinal payoffs. In *Proceedings of the 8th Conference on Logic and the foundation of Game and Decision Theory (LOFT 2008)*, pages 59–86, Amsterdam, the Netherlands, 2008. Amsterdam University Press.

J. Cui, M. Guo, and X. Tang. Characterizations of itreated admissibility based on pegl. In X. He, J. Horty, and E. Pacuit, editors, *Logic, Rationality and Interaction, Proceedings of the Second International Workshop, LORI 2009, Chongqing, China, October*, pages 76–89, Berlin, 2009. Springer.

J. C. Harsanyi. Games with incomplete information played by 'bayesian' players, parts i, ii, and iii. *Management Science*, 14:159–182, 320–334, and 486–502, 1967/68.

R. Jeffrey. *The Logic of Decision*. New York: McGraw-Hill, New York,USA, 1965.

E. Lorini and F. Schwarzentruber. A modal logic of epistemic games. *Games*, 1:478–526, 2010.

M. Osberne and A. Rubinstein. *A Course in Game theory*. MIT Press, Cambridge, MA, USA, 1994.

M. Pauly. A modal logic for coalitional power in games. *Journal of Logic and Computation*, 12(1):149–166, 2002.

M. Peterson. *An Introduction to Decision Theory*. Cambridge University Press, Cambridge, UK, 2009.

J. Seligman. Hybrid logic for analysing games. Presented at "Door to Logic" workshop, Tsinghua University, Beijing, 24 May 2010.

J. van Benthem. Games in dynamic-epistemic logic. *Bulletin of Economic Research(Proceedings LOFT-4, Torino)*, 53(4):219–248, October 2001.

J. van Benthem. Rational dynamics and epistemic logic in games. *International Game Theory Review*, 9(1):13–45, May 2007.

J. van Benthem and Ştefan Minică. Toward a dynamic logic of questions. In E. P. X. He, J. Horty, editor, *Logic Rationality and Interaction*, pages 27–41, Chongqing, China, August 2009. Springer.

J. van Benthem, S. van Otterloo, and O. Roy. Preference logic, conditionals and solution concepts in games. In H. Lagerlund, S. Lindstrom, and R. Sliwinski, editors, *Modality matters: Twenty-five essays in honour of Krister Segerberg*, pages 61–77. Uppsala University, 2006.

W. van der Hoek and M. Pauly. Modal logic for games and information. In P. Blackburn, J. van Benthem, and F. Wolter, editors, *Handbook of Modal Logic*, pages 1078–1148. Elsevier, Amsterdam, the Netherlands, 2006.

A Uniform Logic of Information Update

Wesley H. Holliday, Tomohiro Hoshi and Thomas F. Icard, III

Logical Dynamics Lab, Center for the Study of Language and Information
Cordura Hall, 210 Panama Street, Stanford, CA 94305
Department of Philosophy, Building 90, Stanford University, Stanford, CA 94305
weshollidayatstanfordedu, thoshiatstanfordedu, icardatstanfordedu

Abstract

Unlike standard modal logics, many dynamic epistemic logics are not closed under uniform substitution. A distinction therefore arises between the logic and its *substitution core*, the set of formulas all of whose substitution instances are valid. The classic example of a non-uniform dynamic epistemic logic is Public Announcement Logic (PAL), and a well-known open problem is to axiomatize the substitution core of PAL. In this paper we solve this problem for PAL over the class of all relational models with infinitely many agents, PAL-$\mathbf{K}_\omega$, as well as standard extensions thereof, e.g., PAL-$\mathbf{T}_\omega$, PAL-$\mathbf{S4}_\omega$, and PAL-$\mathbf{S5}_\omega$. We introduce a new Uniform Public Announcement Logic (UPAL), prove completeness of a deductive system with respect to UPAL semantics, and show that this system axiomatizes the substitution core of PAL.[1]

[1]This paper is a preprint of "A Uniform Logic of Information Dynamics," forthcoming in *Advances in Modal Logic*, Vol. 9 (College Publications).

1 Introduction

One of the striking features of many of the *dynamic epistemic logics* (Plaza 1989, Gerbrandy and Groenevelt 1997, Baltag et al. 1998, van Ditmarsch, H. et al. 2008, van Benthem, J. 2011a) studied in the last twenty years is the failure of closure under *uniform substitution* in these systems. Given a valid principle of information update in such a system, uniformly substituting complex epistemic formulas for atomic sentences in the principle may result in an *invalid* instance. Such failures of closure under uniform substitution turn out to reveal insights into the nature of information change (van Benthem, J. 2004, van Ditmarsch, H. and Kooi 2006, Balbiani et al. 2008, Holliday and Icard, III 2010, van Ditmarsch, H. et al. 2011). They also raise the question: what are the more robust principles of information update that are valid in all instances, that are *schematically* valid? Even for the simplest system of dynamic epistemic logic, Public Announcement Logic (PAL) (Plaza 1989), the answer has been unknown. In van Benthem's "Open Problems in Logical Dynamics" (van Benthem, J. 2006a), Question 1 is whether the set of schematic validities of PAL is axiomatizable.[2]

In this paper, we give an axiomatization of the set of schematic validities—or *substitution core*—of PAL over the class of all relational models with infinitely many agents, PAL-$\mathbf{K}_\omega$, as well as standard extensions thereof, e.g., PAL-$\mathbf{T}_\omega$, PAL-$\mathbf{S4}_\omega$, and PAL-$\mathbf{S5}_\omega$. After reviewing the basics of PAL in §1.1, we introduce the idea of Uniform Public Announcement Logic (UPAL) in §1.2, prove completeness of a UPAL deductive system in §3 with respect to alternative semantics introduced in §2, and show that it axiomatizes the substitution core of PAL in §4. In §5, we demonstrate our techniques with examples, and in §6 we conclude by discussing extensions of these techniques to other logics.

Although much could be said about the conceptual significance of UPAL as a uniform logic of information update, here we only present the formal results. For conceptual discussion of PAL, we refer the reader to the textbooks (van

[2]Dynamic epistemic logics are not the only non-uniform modal logics to have been studied. Other examples include Buss's (Buss 1990) modal logic of "pure provability," Åqvist's (Aqvist 1973) two-dimensional modal logic (see (Segerberg 1973)), Carnap's (Carnap 1946) modal system for logical necessity (see (Ballarin 2005, Schurz 2005)), an epistemic-doxastic logic proposed by Halpern (Halpern 1996), and the full computation tree logic CTL* (see (Reynolds 2001)). Among propositional logics, inquisitive logic (Mascarenhas 2009, Ciardelli 2009) is a non-uniform example. In some of these cases, the schematically valid fragment—or *substitution core*—turns out to be another known system. For example, the substitution core of Carnap's system **C** is **S5** (Schurz 2005), and the substitution core of inquisitive logic is Medvedev Logic (Ciardelli 2009, §3.4).

Ditmarsch, H. et al. 2008, van Benthem, J. 2011a). Our work here supports a theme of other recent work in dynamic epistemic logic: despite its apparent simplicity, PAL and its variants prove to be a rich source for mathematical investigation (see, e.g., van Benthem, J. 2006a;b, Kooi 2007, Holliday and Icard, III 2010, Holliday et al. 2011, Wang 2011, Ma 2011, van Benthem, J. 2011b, Holliday et al. 2012, Wang and Cao 2012).

1.1 Review of PAL

We begin our review of PAL with the language we will use throughout.

Definition 1.1. For a set At of atomic sentences and a set Agt of agent symbols with $|\mathsf{Agt}| = \kappa$, the language $\mathcal{L}^{\kappa}_{\mathsf{PAL}}$ is generated by the following grammar:

$$\varphi \ ::= \top \mid \ p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Diamond_a \varphi \mid \langle\varphi\rangle\varphi,$$

where $p \in \mathsf{At}$ and $a \in \mathsf{Agt}$. We define $\Box_a\varphi$ as $\neg\Diamond_a\neg\varphi$ and $[\varphi]\psi$ as $\neg\langle\varphi\rangle\neg\psi$.

- $\mathsf{Sub}(\varphi)$ is the set of subformulas of φ;
- $\mathsf{At}(\varphi) = \mathsf{At} \cap \mathsf{Sub}(\varphi)$;
- $\mathsf{Agt}(\varphi) = \{a \in \mathsf{Agt} \mid \Diamond_a\psi \in \mathsf{Sub}(\varphi) \text{ for some } \psi \in \mathcal{L}^{\kappa}_{\mathsf{PAL}}\}$;
- $\mathsf{An}(\varphi) = \{\chi \in \mathcal{L}^{\kappa}_{\mathsf{PAL}} \mid \langle\chi\rangle\psi \in \mathsf{Sub}(\varphi) \text{ for some } \psi \in \mathcal{L}^{\kappa}_{\mathsf{PAL}}\}$.

We will be primarily concerned with the language $\mathcal{L}^{\omega}_{\mathsf{PAL}}$ with infinitely many agents, which leads to a more elegant treatment than $\mathcal{L}^{n}_{\mathsf{PAL}}$ for some arbitrary finite n. In §6 we will briefly discuss the single-agent and finite-agent cases.

We will consider two interpretations of $\mathcal{L}^{\kappa}_{\mathsf{PAL}}$, one now and one in §2. The standard interpretation uses the following models and truth definition.

Definition 1.2. Models for PAL are tuples of the form $\mathcal{M} = \langle W, \{R_a\}_{a\in\mathsf{Agt}}, V\rangle$, where W is a non-empty set, R_a is a binary relation on W, and $V\colon \mathsf{At} \to \mathcal{P}(W)$.

Definition 1.3. Given a PAL model $\mathcal{M} = \langle W, \{R_a\}_{a\in\mathsf{Agt}}, V\rangle$ with $w \in W$, $\varphi, \psi \in$

$\mathcal{L}^{\kappa}_{\mathsf{PAL}}$, and $p \in \mathsf{At}$, we define $\mathcal{M}, w \vDash \varphi$ as follows:

$$
\begin{array}{lll}
\mathcal{M}, w \vDash \top; & & \\
\mathcal{M}, w \vDash p & \text{iff} & w \in V(p); \\
\mathcal{M}, w \vDash \neg\varphi & \text{iff} & \mathcal{M}, w \nvDash \varphi; \\
\mathcal{M}, w \vDash \varphi \wedge \psi & \text{iff} & \mathcal{M}, w \vDash \varphi \text{ and } \mathcal{M}, w \vDash \psi; \\
\mathcal{M}, w \vDash \Diamond_a \varphi & \text{iff} & \exists v \in W\colon wR_a v \text{ and } \mathcal{M}, v \vDash \varphi; \\
\mathcal{M}, w \vDash \langle\varphi\rangle\psi & \text{iff} & \mathcal{M}, w \vDash \varphi \text{ and } \mathcal{M}_{|\varphi}, w \vDash \psi,
\end{array}
$$

where $\mathcal{M}_{|\varphi} = \langle W_{|\varphi}, \{R_{a_{|\varphi}}\}_{a\in\mathsf{Agt}}, V_{|\varphi}\rangle$ is the model such that

$$
\begin{array}{l}
W_{|\varphi} = \{v \in W \mid \mathcal{M}, v \vDash \varphi\}; \\
\forall a \in \mathsf{Agt}\colon R_{a_{|\varphi}} = R_a \cap (W_{|\varphi} \times W_{|\varphi}); \\
\forall p \in \mathsf{At}\colon V_{|\varphi}(p) = V(p) \cap W_{|\varphi}.
\end{array}
$$

We use the notation $[\![\varphi]\!]^{\mathcal{M}} = \{v \in W \mid \mathcal{M}, v \vDash \varphi\}$. For a class of models C, $\mathrm{Th}_{\mathcal{L}^{\kappa}_{\mathsf{PAL}}}(\mathsf{C})$ is the set of formulas of $\mathcal{L}^{\kappa}_{\mathsf{PAL}}$ that are valid over C.

For the following statements, we use the standard nomenclature for normal modal logics, e.g., **K**, **T**, **S4**, and **S5** for the unimodal logics and $\mathbf{K}_\kappa$, $\mathbf{T}_\kappa$, $\mathbf{S4}_\kappa$, and $\mathbf{S5}_\kappa$ for their multimodal versions with $|\mathsf{Agt}| = \kappa$ (assume κ countable). Let $\mathrm{Mod}(\mathbf{L}_\kappa)$ be the class of all models of the logic $\mathbf{L}_\kappa$, so $\mathrm{Mod}(\mathbf{K}_\kappa)$ is the class of all models, $\mathrm{Mod}(\mathbf{T}_\kappa)$ is the class of models with reflexive R_a relations, etc. We write L_κ for the Hilbert-style deductive system whose set of theorems is $\mathbf{L}_\kappa$, and for any deductive system S, we write $\vdash_{\mathsf{S}} \varphi$ when φ is a theorem of S.

Theorem 1 (PAL Axiomatization (Plaza 1989))**.** Let PAL-L_κ be the system ex-

tending L_κ with the following rule and axioms:[3]

i.	(replacement)	$\dfrac{\psi \leftrightarrow \chi}{\varphi(\psi/p) \leftrightarrow \varphi(\chi/p)}$
ii.	(atomic reduction)	$\langle\varphi\rangle p \leftrightarrow (\varphi \wedge p)$
iii.	(negation reduction)	$\langle\varphi\rangle\neg\psi \leftrightarrow (\varphi \wedge \neg\langle\varphi\rangle\psi)$
iv.	(conjunction reduction)	$\langle\varphi\rangle(\psi \wedge \chi) \leftrightarrow (\langle\varphi\rangle\psi \wedge \langle\varphi\rangle\chi)$
v.	(diamond reduction)	$\langle\varphi\rangle\Diamond_a\psi \leftrightarrow (\varphi \wedge \Diamond_a\langle\varphi\rangle\psi)$.

For all $\varphi \in \mathcal{L}^\kappa_{\mathsf{PAL}}$,

$$\vdash_{\mathsf{PAL\text{-}K}_\kappa} \varphi \text{ iff } \varphi \in \mathrm{Th}_{\mathcal{L}^\kappa_{\mathsf{PAL}}}(\mathrm{Mod}(\mathbf{K}_\kappa)).$$

The same result holds for $\mathsf{T}_\kappa/\mathbf{T}_\kappa$, $\mathsf{S4}_\kappa/\mathbf{S4}_\kappa$, and $\mathsf{S5}_\kappa/\mathbf{S5}_\kappa$ in place of $\mathsf{K}_\kappa/\mathbf{K}_\kappa$.

Although we have taken diamond operators as primitive for convenience in later sections, typically the PAL axiomatization is stated in terms of box operators by replacing axiom schemas ii - v by the following: $[\varphi]p \leftrightarrow (\varphi \to p)$; $[\varphi]\neg\psi \leftrightarrow (\varphi \to \neg[\varphi]\psi)$; $[\varphi](\psi \wedge \chi) \leftrightarrow ([\varphi]\psi \wedge [\varphi]\chi)$; $[\varphi]\Box_a\psi \leftrightarrow (\varphi \to \Box_a[\varphi]\psi)$.

1.2 Introduction to UPAL

As noted above, one of the striking features of PAL is that it is not closed under uniform substitution. In the terminology of Goldblatt (1992), PAL is not a *uniform* modal logic. For example, the valid atomic reduction axiom has invalid substitution instances, e.g., $\langle p\rangle\Box_a p \leftrightarrow (p \wedge \Box_a p)$. Given this observation, a distinction arises between PAL and its *substitution core*, defined as follows.

[3] If L_κ contains the rule of uniform substitution, then we must either restrict this rule so that in $\mathsf{PAL\text{-}L}_\kappa$ we can only substitute into formulas φ with $\mathsf{An}(\varphi) = \emptyset$, or remove the rule and add for each axiom of L_κ all substitution instances of that axiom with formulas in $\mathcal{L}^\kappa_{\mathsf{PAL}}$. Either way, we take the rules of modus ponens and $\Box_a$-necessitation from L_κ to apply in $\mathsf{PAL\text{-}L}_\kappa$ to all formulas. Finally, for $\varphi, \psi \in \mathcal{L}^\kappa_{\mathsf{PAL}}$ and $p \in \mathsf{At}(\varphi)$, $\varphi(\psi/p)$ is the formula obtained by replacing all occurrences of p in φ by ψ. For alternative axiomatizations of PAL, see Wang 2011, Wang and Cao 2012.

Definition 1.4. A substitution is any $\sigma\colon \mathsf{At} \to \mathcal{L}^{\kappa}_{\mathsf{PAL}}$; and $(\cdot)^{\sigma}\colon \mathcal{L}^{\kappa}_{\mathsf{PAL}} \to \mathcal{L}^{\kappa}_{\mathsf{PAL}}$ is the extension such that $(\varphi)^{\sigma}$ is obtained from φ by replacing each $p \in \mathsf{At}(\varphi)$ by $\sigma(p)$ (Blackburn et al. 2001, Def. 1.18). The *substitution core* of $\mathrm{Th}_{\mathcal{L}^{\kappa}_{\mathsf{PAL}}}(\mathsf{C})$ is the set

$$\{\varphi \in \mathcal{L}^{\kappa}_{\mathsf{PAL}} : (\varphi)^{\sigma} \in \mathrm{Th}_{\mathcal{L}^{\kappa}_{\mathsf{PAL}}}(\mathsf{C}) \text{ for all substitutions } \sigma\}.$$

Formulas in the substitution core of $\mathrm{Th}_{\mathcal{L}^{\kappa}_{\mathsf{PAL}}}(\mathsf{C})$ are *schematically valid* over C.

Examples of formulas that are in $\mathrm{Th}_{\mathcal{L}^{\kappa}_{\mathsf{PAL}}}(\mathrm{Mod}(\mathbf{K}_{\kappa}))$ but are not in the substitution core of $\mathrm{Th}_{\mathcal{L}^{\kappa}_{\mathsf{PAL}}}(\mathrm{Mod}(\mathbf{K}_{\kappa}))$ include the following (for $\kappa \geq 1$):[4]

$[p]p$	$\Box_a p \to [p]\Box_a p$
$[p]\Box_a p$	$\Box_a p \to [p](p \to \Box_a p)$
$[p](p \to \Box_a p)$	$\Box_a(p \to q) \to (\langle q\rangle \Box_a r \to \langle p\rangle \Box_a r)$
$[p \wedge \neg\Box_a p]\neg(p \wedge \neg\Box_a p)$	$(\langle p\rangle \Box_a r \wedge \langle q\rangle \Box_a r) \to \langle p \vee q\rangle \Box_a r.$

We discuss the epistemic significance of failures of uniformity in Holliday et al. 2012. Burgess (2003, 147f) explains the logical significance of uniformity:

> The standard aim of logicians at least from Russell onward has been to characterize the class [of] all formulas *all of whose instantiations are true*. Thus, though Russell was a logical atomist, when he endorsed $p \vee \sim p$ as [a] law of logic, he did not mean to be committing himself only to the view that the disjunction of any *logically atomic* statement with its negation is true, but rather to be committing himself to the view that the disjunction of *any statement whatsoever* with its negation is true . . . . This has remained the standard employment of statement letters ever since, not only among Russell's successors in the classical tradition, but also among the great majority of formal logicians who have thought classical logic to be in need of additions and/or amendments, including C. I. Lewis, the founder of modern modal logic. With such an understanding of the role of statement letters, it is clear that if A is a law of logic, and B is any substitution in A, then B also is a law of logic . . . . Thus it is that the rule of substitution applies

[4]The first two principles in the second column are schematically valid over transitive single-agent models, but not over all single-agent models or over transitive multi-agent models.

> not only in classical logic, but in standard, Lewis-style modal logics (as well as in intuitionistic, temporal, relevance, quantum, and other logics). None of this is meant to deny that there may be circumstances where it is legitimate to adopt some other understanding of the role of statement letters. If one does so, however, it is indispensable to note the conceptual distinction, and highly advisable to make a notational and terminological distinction.

In PAL, an atomic sentence p has the same truth value at any pointed models $\mathcal{M}, w$ and $\mathcal{M}_{|\varphi}, w$, whereas a formula containing a modal operator may have different truth values at $\mathcal{M}, w$ and $\mathcal{M}_{|\varphi}, w$, which is why uniform substitution does not preserve PAL-validity. Hence in PAL an atomic sentence cannot be thought of as a *propositional variable* in the ordinary sense of something that stands in for any proposition. By contrast, if we consider the substitution core of PAL as a logic in its own right, for which semantics will be given in §2, then we can think of the atomic sentences as genuine propositional variables.

The distinction between PAL and its substitution core leads to Question 1 in van Benthem's list of "Open Problems in Logical Dynamics":

Question 1 (van Benthem, J. 2006b;a; 2011a)**.** Is the substitution core of PAL axiomatizable?

To answer this question, we will introduce a new framework of Uniform Public Announcement Logic (UPAL), which we use to prove the following:

Theorem 2 (Axiomatization of the PAL Substitution Core)**.**

Let $\mathsf{UPAL\text{-}L}_\kappa$ be the system extending L_κ with the following rules and axioms:[5]

1.	(uniformity)	$\dfrac{\varphi}{(\varphi)^\sigma}$ for any substitution σ
2.	(necessitation)	$\dfrac{\varphi}{[p]\varphi}$
3.	(extensionality)	$\dfrac{\varphi \leftrightarrow \psi}{\langle\varphi\rangle p \leftrightarrow \langle\psi\rangle p}$
4.	(distribution)	$[p](q \to r) \to ([p]q \to [p]r)$
5.	(p-seriality)	$p \to \langle p\rangle\top$
6.	(truthfulness)	$\langle p\rangle\top \to p$
7.	($\top$-reflexivity)	$p \to \langle\top\rangle p$
8.	(functionality)	$\langle p\rangle q \to [p]q$
9.	(pa-commutativity)	$\langle p\rangle\Diamond_a q \to \Diamond_a\langle p\rangle q$
10.	(ap-commutativity)	$\Diamond_a\langle p\rangle q \to [p]\Diamond_a q$
11.	(composition)	$\langle p\rangle\langle q\rangle r \leftrightarrow \langle\langle p\rangle q\rangle r$.

For all $\varphi \in \mathcal{L}^\omega_{\mathsf{PAL}}$,

$$\vdash_{\mathsf{UPAL\text{-}K}_\omega} \varphi \text{ iff } \varphi \text{ is in the substitution core of } \mathrm{Th}_{\mathcal{L}^\omega_{\mathsf{PAL}}}(\mathrm{Mod}(\mathbf{K}_\omega)).$$

The same result holds for $\mathsf{T}_\omega/\mathbf{T}_\omega$, $\mathsf{S4}_\omega/\mathbf{S4}_\omega$, and $\mathsf{S5}_\omega/\mathbf{S5}_\omega$ in place of $\mathsf{K}_\omega/\mathbf{K}_\omega$, with only minor adjustments to the proof (see note 6).

Theorem 3 (Axiomatization of the PAL Substitution Core cont.)**.**

[5] As in $\mathsf{PAL\text{-}L}_\kappa$, in $\mathsf{UPAL\text{-}L}_\kappa$ we take the rules of modus ponens and $\Box_a$-necessitation from L_κ to apply to all formulas in $\mathcal{L}^\kappa_{\mathsf{PAL}}$.

1. $\vdash_{\mathsf{UPAL\text{-}T}_\omega} \varphi$ iff φ is in the substitution core of $\mathrm{Th}_{\mathcal{L}^\omega_{\mathsf{PAL}}}(\mathrm{Mod}(\mathbf{T}_\omega))$;
2. $\vdash_{\mathsf{UPAL\text{-}S4}_\omega} \varphi$ iff φ is in the substitution core of $\mathrm{Th}_{\mathcal{L}^\omega_{\mathsf{PAL}}}(\mathrm{Mod}(\mathbf{S4}_\omega))$;
3. $\vdash_{\mathsf{UPAL\text{-}S5}_\omega} \varphi$ iff φ is in the substitution core of $\mathrm{Th}_{\mathcal{L}^\omega_{\mathsf{PAL}}}(\mathrm{Mod}(\mathbf{S5}_\omega))$.

Unless the specific base system L_κ matters, we simply write '`UPAL`' and '`PAL`'. It is easy to check that all the axioms of PAL except atomic reduction are derivable in UPAL, and the rule of replacement is an admissible rule in UPAL. Another system with the same theorems as `UPAL`, but presented in a format closer to that of the typical box version of `PAL`, is the following (with $\bot := \neg\top$):

I.	(uniformity)	$\dfrac{\varphi}{(\varphi)^\sigma}$ for any substitution σ
II.	(RE)	$\dfrac{\varphi \leftrightarrow \psi}{[p]\varphi \leftrightarrow [p]\psi}$
III.	([ ]-extensionality)	$\dfrac{\varphi \leftrightarrow \psi}{[\varphi]p \leftrightarrow [\psi]p}$
IV.	(N)	$[p]\top$
V.	($\top$-reflexivity)	$[\top]p \rightarrow p$
VI.	($\bot$-reduction)	$[p]\bot \leftrightarrow \neg p$
VII.	($\neg$-reduction)	$[p]\neg q \leftrightarrow (p \rightarrow \neg[p]q)$
VIII.	($\wedge$-reduction)	$[p](q \wedge r) \leftrightarrow ([p]q \wedge [p]r)$
IX.	($\Box_a$-reduction)	$[p]\Box_a q \leftrightarrow (p \rightarrow \Box_a[p]q)$
X.	([ ]-composition)	$[p][q]r \leftrightarrow [p \wedge [p]q]r$.

We have formulated `UPAL` as in Theorem 2 to make clear the correspondence between axioms and the semantic conditions in Definition 2.3 below, as well as to make clear the specific properties used in the steps of our main proof.

2 Semantics for UPAL

In this section we introduce semantics for Uniform Public Announcement Logic, for which the system of UPAL is shown to be sound and complete in §3.

Definition 2.1. Models for UPAL are tuples $\mathfrak{M} = \langle M, \{\mathcal{R}_a\}_{a\in\mathsf{Agt}}, \{\mathcal{R}_\varphi\}_{\varphi\in\mathcal{L}^\kappa_{\mathsf{PAL}}}, \mathcal{V}\rangle$; M is a non-empty set, $\mathcal{R}_a$ and $\mathcal{R}_\varphi$ are binary relations on M, and $\mathcal{V}\colon \mathsf{At} \to \mathcal{P}(M)$.

Unlike in the PAL truth definition, in the UPAL truth definition we treat $\langle\varphi\rangle$ like any other modal operator.

Definition 2.2. Given a UPAL model $\mathfrak{M} = \langle M, \{\mathcal{R}_a\}_{a\in\mathsf{Agt}}, \{\mathcal{R}_\varphi\}_{\varphi\in\mathcal{L}^\kappa_{\mathsf{PAL}}}, \mathcal{V}\rangle$ with $w \in M$, $\varphi, \psi \in \mathcal{L}^\kappa_{\mathsf{PAL}}$, and $p \in \mathsf{At}$, we define $\mathfrak{M}, w \Vdash \varphi$ as follows:

$$
\begin{array}{lll}
\mathfrak{M}, w \Vdash \top; & & \\
\mathfrak{M}, w \Vdash p & \text{iff} & w \in \mathcal{V}(p); \\
\mathfrak{M}, w \Vdash \neg\varphi & \text{iff} & \mathfrak{M}, w \nVdash \varphi; \\
\mathfrak{M}, w \Vdash \varphi \wedge \psi & \text{iff} & \mathfrak{M}, w \Vdash \varphi \text{ and } \mathfrak{M}, w \Vdash \psi; \\
\mathfrak{M}, w \Vdash \Diamond_a\varphi & \text{iff} & \exists v \in M\colon w\mathcal{R}_a v \text{ and } \mathfrak{M}, v \Vdash \varphi; \\
\mathfrak{M}, w \Vdash \langle\varphi\rangle\psi & \text{iff} & \exists v \in M\colon w\mathcal{R}_\varphi v \text{ and } \mathfrak{M}, v \Vdash \psi.
\end{array}
$$

We use the notation $\|\varphi\|^{\mathfrak{M}} = \{v \in M \mid \mathfrak{M}, v \Vdash \varphi\}$.

Instead of giving the $\langle\varphi\rangle$ operators a special truth clause, we ensure that they behave in a PAL-like way by imposing constraints on the $\mathcal{R}_\varphi$ relations in Definition 2.3 below. Wang and Cao (2012) have independently proposed a semantics for PAL in this style, with respect to which they prove that PAL is complete. The difference comes in the specific constraints for UPAL vs. PAL.

Definition 2.3. A UPAL model $\mathfrak{M} = \langle M, \{\mathcal{R}_a\}_{a\in\mathsf{Agt}}, \{\mathcal{R}_\varphi\}_{\varphi\in\mathcal{L}^\kappa_{\mathsf{PAL}}}, \mathcal{V}\rangle$ is *legal* iff the following conditions hold for all $\psi, \chi \in \mathcal{L}^\kappa_{\mathsf{PAL}}$, $w, v \in M$, and $a \in \mathsf{Agt}$:

(**extensionality**) if $\|\psi\|^{\mathfrak{M}} = \|\chi\|^{\mathfrak{M}}$, then $\mathcal{R}_\psi = \mathcal{R}_\chi$;

(ψ**-seriality**) if $w \in \|\psi\|^{\mathfrak{M}}$, then $\exists v$: $w\mathcal{R}_\psi v$;

(**truthfulness**) if $w\mathcal{R}_\psi v$, then $w \in \|\psi\|^{\mathfrak{M}}$;

($\top$**-reflexivity**) $w\mathcal{R}_\top w$;

(**functionality**) if $w\mathcal{R}_\psi v$, then for all $u \in M$, $w\mathcal{R}_\psi u$ implies $u = v$;

(ψa**-commutativity**) if $w\mathcal{R}_\psi v$ and $v\mathcal{R}_a u$, then $\exists$z: $w\mathcal{R}_a z$ and $z\mathcal{R}_\psi u$;

($a\psi$**-commutativity**) if $w\mathcal{R}_a v$, $v\mathcal{R}_\psi u$ and $w \in \|\psi\|^{\mathfrak{M}}$,
then $\exists$z: $w\mathcal{R}_\psi z$ and $z\mathcal{R}_a u$;

(**composition**) $\mathcal{R}_{\langle\psi\rangle\chi} = \mathcal{R}_\psi \circ \mathcal{R}_\chi$.

In §4, we will also refer to weaker versions of the first and third conditions:

(**extensionality for** φ) if $\psi, \chi \in \mathsf{An}(\varphi) \cup \{\top\}$ and $\|\psi\|^{\mathfrak{M}} = \|\chi\|^{\mathfrak{M}}$,
then $\mathcal{R}_\psi = \mathcal{R}_\chi$;

(**truthfulness for** φ) if $\psi \in \mathsf{An}(\varphi) \cup \{\top\}$ and $w\mathcal{R}_\psi v$, then $w \in \|\psi\|^{\mathfrak{M}}$.

It is easy to see that each of the axioms of UPAL in Theorem 2 corresponds to the condition of the same name written in boldface in Definition 2.3.

3 Completeness of UPAL

In this section, we take our first step toward proving Theorem 2 by proving:

Theorem 4 (Soundness and Completeness)**.** The system of UPAL-K_ω given in

Theorem 2 is sound and complete for the class of legal UPAL models.

Soundness is straightforward. To prove completeness, we use the standard canonical model argument.

Definition 3.1. The canonical model $\mathfrak{M}^c = \langle M^c, \{\mathcal{R}^c_a\}_{a\in\mathsf{Agt}}, \{\mathcal{R}^c_\varphi\}_{\varphi\in\mathcal{L}^\kappa_{\mathsf{PAL}}}, \mathcal{V}^c\rangle$ is defined as follows:

1. $M^c = \{\Gamma \mid \Gamma$ is a maximally UPAL-$\mathtt{K}_\omega$-consistent set$\}$;
2. $\Gamma\mathcal{R}^c_a\Delta$ iff $\psi \in \Delta$ implies $\Diamond_a\psi \in \Gamma$;
3. $\Gamma\mathcal{R}^c_\varphi\Delta$ iff $\psi \in \Delta$ implies $\langle\varphi\rangle\psi \in \Gamma$;
4. $\mathcal{V}^c(p) = \{\Gamma \in M^c \mid p \in \Gamma\}$.

The following fact, easily shown, will be used in the proof of Lemma 2.

Fact 1. For all $\Gamma \in M^c$, $\varphi \in \mathcal{L}^\kappa_{\mathsf{PAL}}$, if $\langle\varphi\rangle\top \in \Gamma$, then $\{\psi \mid \langle\varphi\rangle\psi \in \Gamma\} \in M^c$.

The proof of the truth lemma is standard (Blackburn et al. 2001, §4.2).

Lemma 1 (Truth)**.** For all $\Gamma \in M^c$ and $\varphi \in \mathcal{L}^\kappa_{\mathsf{PAL}}$,

$$\mathfrak{M}^c, \Gamma \Vdash \varphi \text{ iff } \varphi \in \Gamma.$$

To complete the proof of Theorem 4, we need only check the following.

Lemma 2 (Legality)**.** $\mathfrak{M}^c$ is a legal model.

Proof. Suppose $\|\varphi\|^{\mathfrak{M}^c} = \|\psi\|^{\mathfrak{M}^c}$, so by Lemma 1 and the properties of maximally consistent sets, $\varphi \leftrightarrow \psi \in \Gamma$ for all $\Gamma \in M^c$. Hence $\vdash_{\text{UPAL-}\mathtt{K}_\omega} \varphi \leftrightarrow \psi$, for if $\neg(\varphi \leftrightarrow \psi)$ is UPAL-$\mathtt{K}_\omega$-consistent, then $\neg(\varphi \leftrightarrow \psi) \in \Delta$ for some $\Delta \in M^c$, contrary to what was just shown. It follows that for any $\alpha \in \mathcal{L}^\kappa_{\mathsf{PAL}}$, $\vdash_{\text{UPAL-}\mathtt{K}_\omega} \langle\varphi\rangle\alpha \leftrightarrow \langle\psi\rangle\alpha$, given the extensionality and uniformity rules of UPAL-$\mathtt{K}_\omega$. Hence if $\Gamma_1\mathcal{R}^c_\varphi\Gamma_2$, then for all $\alpha \in \Gamma_2$, $\langle\varphi\rangle\alpha \in \Gamma_1$ and $\langle\psi\rangle\alpha \in \Gamma_1$ by the consistency of Γ_1, which means $\Gamma_1\mathcal{R}^c_\psi\Gamma_2$. The other direction is the same, whence $\mathcal{R}^c_\varphi = \mathcal{R}^c_\psi$. $\mathfrak{M}^c$ satisfies **extensionality**.

Suppose $\Gamma_1\mathcal{R}^c_{\langle\varphi\rangle\psi}\Gamma_2$, so for all $\alpha \in \Gamma_2$, $\langle\langle\varphi\rangle\psi\rangle\alpha \in \Gamma_1$. Hence $\langle\varphi\rangle\langle\psi\rangle\alpha \in \Gamma_1$ given the composition axiom and uniformity rule of UPAL-$\mathtt{K}_\omega$, so $\langle\varphi\rangle\top \in \Gamma_1$ by normal modal reasoning with the distribution axiom. It follows by Fact 1 and Definition (*c*).3 that there is some Σ_1 such that $\Gamma_1\mathcal{R}_\varphi\Sigma_1$ and $\langle\psi\rangle\alpha \in \Sigma_1$, and by

similar reasoning that there is some Σ_2 such that $\Sigma_1 \mathcal{R}_\psi \Sigma_2$ and $\alpha \in \Sigma_2$. Hence $\Gamma_2 \subseteq \Sigma_2$, so $\Gamma_2 = \Sigma_2$ given that Γ_2 is maximal. Therefore, $\mathcal{R}^c_{\langle\varphi\rangle\psi} \subseteq \mathcal{R}^c_\varphi \circ \mathcal{R}^c_\psi$. The argument in the other direction is similar. $\mathfrak{M}^c$ satisfies **composition**.

We leave the other legality conditions to the reader. □

4 Bridging UPAL and PAL

In this section, we show that UPAL axiomatizes the substitution core of PAL. It is easy to check that all of the axioms of UPAL are PAL schematic validities, and all of the rules of UPAL preserve schematic validity, so UPAL derives only PAL schematic validities. To prove that UPAL derives all PAL schematic validities, we show that if φ is not derivable from UPAL, so by Theorem 4 there is a legal UPAL model falsifying φ, then there is a substitution τ and a PAL model falsifying $(\varphi)^\tau$, in which case φ is not schematically valid over PAL models.

Proposition 1. For any formula $\varphi \in \mathcal{L}^\omega_{\mathsf{PAL}}$, if there is a legal UPAL model $\mathfrak{M} = \langle M, \{\mathcal{R}_a\}_{a\in\mathsf{Agt}}, \{\mathcal{R}_\psi\}_{\psi\in\mathcal{L}^\omega_{\mathsf{PAL}}}, \mathcal{V}\rangle$ with $w_0 \in M$ such that $\mathfrak{M}, w_0 \nVdash \varphi$, then there is a PAL model $\mathcal{N} = \langle N_0, \{S_a\}_{a\in\mathsf{Agt}}, U\rangle$ with $w_0 \in N_0$ and a substitution τ such that $\mathcal{N}, w_0 \nvDash (\varphi)^\tau$.

Our first step in proving Proposition 1 is to show that we can reduce φ to a certain simple form, which will help us in constructing the substitution τ.

Definition 4.1. The set of *simple* formulas is generated by the grammar

$$\varphi ::= \top \mid p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Diamond_a\varphi \mid \langle\varphi\rangle p,$$

where $p \in \mathsf{At}$ and $a \in \mathsf{Agt}$.

Proposition 2. For every $\varphi \in \mathcal{L}^\kappa_{\mathsf{PAL}}$, there is a simple formula $\varphi' \in \mathcal{L}^\kappa_{\mathsf{PAL}}$ that is equivalent to φ over legal UPAL models (and all PAL models).

Proof. The proof is similar to the standard PAL reduction argument (van Ditmarsch, H. et al. 2008, §7.4), only we do not perform atomic reduction steps, and we use the composition axiom of UPAL to eliminate consecutive occurrences of dynamic operators. □

By Proposition 2, given that $\mathfrak{M}$ is legal, we can assume that φ is simple. Before constructing $\mathcal{N}$ and τ, we show that our initial model $\mathfrak{M}$ can be transformed into an intermediate model $\mathfrak{N}$ that satisfies a property (part 2 of Lemma 3) that we will take advantage of in our proofs below. We will return to the role of this property in relating UPAL to PAL in Example 1 and §6.

For what follows, we need some new notation. First, let

$$\mathcal{R}_{\mathsf{Agt}} = \bigcup_{a \in \mathsf{Agt}} \mathcal{R}_a;$$

$\mathcal{R}^*$ is the reflexive transitive closure of $\mathcal{R}$; and $\mathcal{R}(w) = \{v \in M \mid w\mathcal{R}v\}$.

Lemma 3. For any legal model $\mathfrak{M} = \langle M, \{\mathcal{R}_a\}_{a\in\mathsf{Agt}}, \{\mathcal{R}_\varphi\}_{\varphi\in\mathcal{L}^\omega_{\mathsf{PAL}}}, \mathcal{V}\rangle$ with $w_0 \in M$ such that $\mathfrak{M}, w_0 \Vdash \varphi$, there is a model $\mathfrak{N} = \langle N, \{\mathcal{S}_a\}_{a\in\mathsf{Agt}}, \{\mathcal{S}_\varphi\}_{\varphi\in\mathcal{L}^\omega_{\mathsf{PAL}}}, \mathcal{U}\rangle$ with $w_0 \in N$ such that

1. $\mathfrak{N}, w_0 \Vdash \varphi$;
2. if $\alpha, \beta \in \mathsf{An}(\varphi) \cup \{\top\}$ and $\|\alpha\|^{\mathfrak{N}} \neq \|\beta\|^{\mathfrak{N}}$, then
$$\|\alpha\|^{\mathfrak{N}} \cap \mathcal{S}^*_{\mathsf{Agt}}(w_0) \neq \|\beta\|^{\mathfrak{N}} \cap \mathcal{S}^*_{\mathsf{Agt}}(w_0).$$
3. $\mathfrak{N}$ satisfies $\top$**-reflexivity**, **functionality**, **extensionality for** φ and **truthfulness for** φ.

Proof. Consider some $\alpha, \beta \in \mathsf{An}(\varphi) \cup \{\top\}$ such that $\|\alpha\|^{\mathfrak{M}} \neq \|\beta\|^{\mathfrak{M}}$. Hence there is some $v \in M$ such that $\mathfrak{M}, v \nVdash \alpha \leftrightarrow \beta$. Let $\mathfrak{M}'$ be exactly like $\mathfrak{M}$ except that for some $x \notin \mathsf{Agt}(\varphi)$, $w_0\mathcal{R}'_x v$.[6] Then it is easy to show that for all $\psi \in \mathsf{Sub}(\varphi)$ and $u \in M$,

$$\mathfrak{M}', u \Vdash \psi \text{ iff } \mathfrak{M}, u \Vdash \psi.$$

Hence $\mathfrak{M}', w_0 \Vdash \varphi$ and $\mathfrak{M}', v \nVdash \alpha \leftrightarrow \beta$. Then given $w_0\mathcal{R}'_x v$, we have

$$\|\alpha\|^{\mathfrak{M}'} \cap \mathcal{R}'^*_{\mathsf{Agt}}(w_0) \neq \|\beta\|^{\mathfrak{M}'} \cap \mathcal{R}'^*_{\mathsf{Agt}}(w_0).$$

Finally, one can check that $\mathfrak{M}'$ satisfies $\top$**-reflexivity**, **functionality**, **extensionality for** φ and **truthfulness for** φ by the construction. By repeating this procedure, starting now with $\mathfrak{M}'$, for each of the finitely many α and β as described above, one obtains a model $\mathfrak{N}$ as described in Lemma 3 □

[6] As noted after Theorem 2, we can modify our proof for other models classes. For example, for the class of models with equivalence relations, in this step we can define $\mathcal{R}'_x$ to be the smallest equivalence relation extending $\mathcal{R}_x$ such that $w_0\mathcal{R}'_x v$. Note that since $\alpha, \beta \in \mathsf{An}(\varphi)\cup\{\top\}$ and $x \notin \mathsf{Agt}(\varphi)$, no matter how we define $\mathcal{R}'_x$, the following claim in the text still holds.

Obtaining $\mathfrak{N}$ from $\mathfrak{M}$ as in Lemma 3, we now define our PAL model $\mathcal{N} = \langle N_0, \{S_a\}_{a\in\mathsf{Agt}}, U\rangle$. Let $N_0 = \mathcal{S}^*_{\mathsf{Agt}}(w_0)$; for some $z \notin \mathsf{Agt}(\varphi)$, let S_z be the universal relation on N_0; and for each $a \in \mathsf{Agt}$ with $a \neq z$, let S_a be the restriction of $\mathcal{S}_a$ to N_0. We will define the valuation U after constructing the substitution τ. The following facts will be used in the proof of Lemma 5.

Fact 2.

1. For all $a \in \mathsf{Agt}$ and $w \in N_0$, $S_a(w) = \mathcal{S}_a(w)$.
2. if $\|\alpha\|^{\mathfrak{N}} \cap N_0 = \|\beta\|^{\mathfrak{N}} \cap N_0$, then for all $u \in N_0$,

$$\mathfrak{N}, u \Vdash \langle\alpha\rangle\chi \text{ iff } \mathfrak{N}, u \Vdash \langle\beta\rangle\chi.$$

Proof. Part 1 is obvious. For part 2, if $\|\alpha\|^{\mathfrak{N}} \cap N_0 = \|\beta\|^{\mathfrak{N}} \cap N_0$, then $\|\alpha\|^{\mathfrak{N}} = \|\beta\|^{\mathfrak{N}}$ by Lemma 3.2, so $\mathcal{S}_\alpha = \mathcal{S}_\beta$ by Lemma 3.3 (**extensionality for** φ). □

Remark 1. There is another way of transforming the given UPAL model $\mathfrak{M} = \langle M, \{\mathcal{R}_a\}_{a\in\mathsf{Agt}}, \{\mathcal{R}_\varphi\}_{\varphi\in\mathcal{L}^\omega_{\mathsf{PAL}}}, \mathcal{V}\rangle$ into a PAL model $\mathcal{N}$ sufficient for our purposes. First, let $\mathfrak{N} = \langle N, \{\mathcal{S}_a\}_{a\in\mathsf{Agt}}, \{\mathcal{S}_\varphi\}_{\varphi\in\mathcal{L}^\omega_{\mathsf{PAL}}}, \mathcal{U}\rangle$ be exactly like $\mathfrak{M}$ except that for some $z \notin \mathsf{Agt}(\varphi)$, $\mathcal{S}_z$ is the universal relation on N, and observe that $\mathfrak{N}$ satisfies the conditions of Lemma 3. Second, take $\mathcal{N} = \langle N_0, \{S_a\}_{a\in\mathsf{Agt}}, U\rangle$ such that $N_0 = N$, $S_a = \mathcal{S}_a$, and U is defined as below, and observe that Fact 2 holds. Then the proof can proceed as below. The difference is that this approach takes the domain of the PAL model to be the entire domain of the UPAL model $\mathfrak{N}$, with S_z as the universal relation on this entire domain, whereas our approach takes the domain of the PAL model to be just that of the "epistemic submodel" *generated by* w_0 *in* $\mathfrak{N}$, $\mathcal{S}^*_{\mathsf{Agt}}(w_0)$, with S_z as the universal relation on this set. We prefer the latter approach because it allows us to work with smaller PAL models when we carry out the construction with concrete examples as in §5.

To construct $\tau(p)$ for $p \in \mathsf{At}(\varphi)$, let $B_1, \ldots, B_m$ be the sequence of all B_i such that $\langle B_i\rangle p \in \mathsf{Sub}(\varphi)$, and let $B_0 := \top$. For $0 \leq i, j \leq m$, if $\|B_i\|^{\mathfrak{N}} \cap N_0 = \|B_j\|^{\mathfrak{N}} \cap N_0$, delete one of B_i or B_j from the list (but never B_0), until there is no such pair. Call the resulting sequence $A_0, \ldots, A_n$, and define

$$s(i) = \{j \mid 0 \leq j \leq n \text{ and } \|A_j\|^{\mathfrak{N}} \cap N_0 \subsetneq \|A_i\|^{\mathfrak{N}} \cap N_0\}.$$

Extend the language with new variables $p_0, \ldots, p_n$ and $a_0, \ldots, a_n$, and define $\tau(p) = \gamma_0 \wedge \cdots \wedge \gamma_n$ such that

$$\gamma_i := (\Box_z a_i \wedge \bigwedge_{j\in s(i)} \neg\Box_z a_j) \to p_i.$$

Having extended the language for each $p \in \mathsf{At}(\varphi)$, define the valuation U for N_0 such that for each $p \in \mathsf{At}(\varphi)$, $U(p) = \mathcal{U}(p) \cap N_0$, and for the new variables:

(a) $U(p_i) = \{w \in N_0 \mid \exists u\colon w\mathcal{S}_{A_i}u \text{ and } u \in \mathcal{U}(p)\}$;

(b) $U(a_i) = \|A_i\|^{\mathfrak{N}} \cap N_0$.

Hence:

(**a**) $[\![p_i]\!]^{\mathcal{N}} = \{w \in N_0 \mid \exists u\colon w\mathcal{S}_{A_i}u \text{ and } u \in \mathcal{U}(p)\}$;

(**b**) $[\![a_i]\!]^{\mathcal{N}} = \|A_i\|^{\mathfrak{N}} \cap N_0$.

Note that it follows from (**a**) and the UPAL truth definition that

(**c**) $[\![p_i]\!]^{\mathcal{N}} = \|\langle A_i\rangle p\|^{\mathfrak{N}} \cap N_0$.

Using these facts, we will show that $\mathfrak{N}, w_0 \nVdash \varphi$ implies $\mathcal{N}, w_0 \nvDash \tau(\varphi)$.

Lemma 4. For all $0 \leq i \leq n$,

$$[\![\tau(p)]\!]^{\mathcal{N}_{|a_i}} = [\![p_i]\!]^{\mathcal{N}}.$$

Proof. We first show that for $0 \leq i, j \leq n$, $i \neq j$:

(i) $[\![\gamma_i]\!]^{\mathcal{N}_{|a_i}} = [\![p_i]\!]^{\mathcal{N}_{|a_i}}$;

(ii) $[\![\gamma_j]\!]^{\mathcal{N}_{|a_i}} = [\![a_i]\!]^{\mathcal{N}_{|a_i}} (= N_{0|a_i})$.

For (i), we claim that

$$[\![\Box_z a_i \wedge \bigwedge_{k \in s(i)} \neg\Box_z a_k]\!]^{\mathcal{N}_{|a_i}} = N_{0|a_i}.$$

Since a_i is atomic, $[\![\Box_z a_i]\!]^{\mathcal{N}_{|a_i}} = N_{0|a_i}$. By definition of the s function and (**b**), for all $k \in s(i)$, $[\![a_k]\!]^{\mathcal{N}} \subsetneq [\![a_i]\!]^{\mathcal{N}}$, so $[\![\neg\Box_z a_k]\!]^{\mathcal{N}_{|a_i}} = N_{0|a_i}$. Hence the claimed equation holds, so $[\![\gamma_i]\!]^{\mathcal{N}_{|a_i}} = [\![p_i]\!]^{\mathcal{N}_{|a_i}}$ given the structure of γ_i.

For (ii), we claim that for $j \neq i$,

$$[\![\Box_z a_j \wedge \bigwedge_{k \in s(j)} \neg\Box_z a_k]\!]^{\mathcal{N}_{|a_i}} = \emptyset.$$

By construction of the sequence $A_0, \ldots, A_n$ for p and (**b**), $[\![a_j]\!]^{\mathcal{N}} \neq [\![a_i]\!]^{\mathcal{N}}$. Hence if not $[\![a_i]\!]^{\mathcal{N}} \subsetneq [\![a_j]\!]^{\mathcal{N}}$, then $[\![a_i]\!]^{\mathcal{N}} \not\subseteq [\![a_j]\!]^{\mathcal{N}}$, so $[\![\Box_z a_j]\!]^{\mathcal{N}_{|a_i}} = \emptyset$ because S_z is the universal relation on N_0. If $[\![a_i]\!]^{\mathcal{N}} \subsetneq [\![a_j]\!]^{\mathcal{N}}$, then by (**b**) and the definition of s, $i \in s(j)$; since a_i is atomic, $[\![\neg\Box_z a_i]\!]^{\mathcal{N}_{|a_i}} = \emptyset$. In either case the claimed equation holds, so $[\![\gamma_j]\!]^{\mathcal{N}_{|a_i}} = N_{0|a_i}$ given the structure of γ_j.

Given the construction of τ, (i) and (ii) imply:

$$[\![\tau(p)]\!]^{\mathcal{N}_{|a_i}} = [\![\gamma_i]\!]^{\mathcal{N}_{|a_i}} \cap \bigcap_{j \neq i} [\![\gamma_j]\!]^{\mathcal{N}_{|a_i}} = [\![p_i]\!]^{\mathcal{N}_{|a_i}} \cap [\![a_i]\!]^{\mathcal{N}_{|a_i}} = [\![p_i]\!]^{\mathcal{N}},$$

where the last equality holds because $[\![p_i]\!]^{\mathcal{N}} \subseteq [\![a_i]\!]^{\mathcal{N}}$, which follows from (**a**), (**b**), and the fact that $\mathfrak{N}$ satisfies **truthfulness for** φ. □

We now establish the connection between the UPAL model $\mathfrak{N}$ on the one hand and the PAL model $\mathcal{N}$ and substitution τ on the other.

Lemma 5. For all simple subformulas χ of φ,

$$[\![(\chi)^\tau]\!]^{\mathcal{N}} = \|\chi\|^{\mathfrak{N}} \cap N_0.$$

Proof. By induction on χ. For the base case, we must show $[\![(p)^\tau]\!]^{\mathcal{N}} = \|p\|^{\mathfrak{N}} \cap N_0$ for $p \in \mathsf{At}(\varphi)$. By construction of the sequence $A_0, \ldots, A_n$ for p, $A_0 = \top$, so $\|A_0\|^{\mathfrak{N}} \cap N_0 = N_0$. Then by (**b**), $[\![a_0]\!]^{\mathcal{N}} = N_0$, and hence

$$\begin{aligned}
[\![(p)^\tau]\!]^{\mathcal{N}} &= [\![(p)^\tau]\!]^{\mathcal{N}_{|a_0}} \\
&= [\![p_0]\!]^{\mathcal{N}} && \text{by Lemma 4} \\
&= \{w \in N_0 \mid \exists u : w \mathcal{S}_{A_0} u \text{ and } u \in \mathcal{U}(p)\} && \text{by (\textbf{a})} \\
&= \{w \in N_0 \mid w \in \mathcal{U}(p)\} && \text{by } \top\textbf{-reflexivity} \\
& && \text{and } \textbf{functionality} \\
&= \|p\|^{\mathfrak{N}} \cap N_0.
\end{aligned}$$

The boolean cases are straightforward. Next, we must show $[\![(\Box_a \varphi)^\tau]\!]^{\mathcal{N}} =$

$||\Box_a\varphi||^{\mathfrak{N}} \cap N_0$. For the inductive hypothesis, $[\![(\varphi)^\tau]\!]^{\mathcal{N}} = ||\varphi||^{\mathfrak{N}} \cap N_0$, so

$$\begin{aligned}
[\![(\Box_a\varphi)^\tau]\!]^{\mathcal{N}} &= [\![\Box_a(\varphi)^\tau]\!]^{\mathcal{N}} \\
&= \{w \in N_0 \mid S_a(w) \subseteq [\![(\varphi)^\tau]\!]^{\mathcal{N}}\} \\
&= \{w \in N_0 \mid S_a(w) \subseteq ||\varphi||^{\mathfrak{N}} \cap N_0\} \\
&= \{w \in N_0 \mid S_a(w) \subseteq ||\varphi||^{\mathfrak{N}}\} && \text{given } S_a \subseteq N_0 \times N_0 \\
&= \{w \in N_0 \mid \mathcal{S}_a(w) \subseteq ||\varphi||^{\mathfrak{N}}\} && \text{by Fact 2.1} \\
&= ||\Box_a\varphi||^{\mathfrak{N}} \cap N_0.
\end{aligned}$$

Finally, we must show $[\![(\langle B_i\rangle p)^\tau]\!]^{\mathcal{N}} = ||\langle B_i\rangle p||^{\mathfrak{N}} \cap N_0$. For the inductive hypothesis, $[\![(B_i)^\tau]\!]^{\mathcal{N}} = ||B_i||^{\mathfrak{N}} \cap N_0$. By construction of the sequence $A_0, \ldots, A_n$ for $p \in \mathsf{At}(\varphi)$, there is some A_j such that

$$(\star) \quad ||B_i||^{\mathfrak{N}} \cap N_0 = ||A_j||^{\mathfrak{N}} \cap N_0.$$

Therefore,

$$\begin{aligned}
[\![(B_i)^\tau]\!]^{\mathcal{N}} &= ||A_j||^{\mathfrak{N}} \cap N_0 \\
&= [\![a_j]\!]^{\mathcal{N}} && \text{by } (\mathbf{b}),
\end{aligned}$$

and hence

$$\begin{aligned}
[\![(\langle B_i\rangle p)^\tau]\!]^{\mathcal{N}} &= [\![\langle(B_i)^\tau\rangle(p)^\tau]\!]^{\mathcal{N}} \\
&= [\![\langle a_j\rangle(p)^\tau]\!]^{\mathcal{N}} \\
&= [\![(p)^\tau]\!]^{\mathcal{N}_{|a_j}} \\
&= [\![p_j]\!]^{\mathcal{N}} && \text{by Lemma 4} \\
&= ||\langle A_j\rangle p||^{\mathfrak{N}} \cap N_0 && \text{by } (\mathbf{c}) \\
&= ||\langle B_i\rangle p||^{\mathfrak{N}} \cap N_0 && \text{given } (\star) \text{ and Fact 2.2.}
\end{aligned}$$

The proof by induction is complete. □

With the following fact, we complete the proof of Proposition 1.

Fact 3. $\mathcal{N}, w_0 \not\models (\varphi)^\tau$.

Proof. Immediate from Lemma 5 given $\mathfrak{N}, w_0 \not\Vdash \varphi$. □

5 Examples

In this section, we work out two examples illustrating how the techniques of §4 allow us to find, for any formula φ that is valid but not schematically valid in PAL, a PAL model that falsifies a substitution instance of φ. The proof in §4 shows that all we need to do is find a legal UPAL model falsifying φ. However, since legal UPAL models are generally large, we would like to instead find a small UPAL model falsifying φ, from which we can read off a PAL model that falsifies a substitution instance of φ. In fact, we can always do so provided that the model satisfies a weaker condition than legality. For a given $\varphi \in \mathcal{L}^\kappa_{\mathsf{PAL}}$, we say that a UPAL model $\mathfrak{M}$ is *φ-legal* iff it satisfies all of the legality conditions of Definition 2.3 when we replace **ψ-seriality** with:

(ψ-**seriality for** φ) if $\psi \in \mathsf{An}(\varphi) \cup \{\top\}$ and $w \in \|\psi\|^{\mathfrak{M}}$,

then $\exists v$: $w \mathcal{R}_\psi v$.

Hence in a φ-legal model, we can let all of the infinitely many $\mathcal{R}_\psi$ relations irrelevant to φ be empty, which makes constructing φ-legal models easier. With this new notion, we can state a simple method for finding a PAL model that falsifies a substitution instance of the non-schematically valid φ:

Step 1. Transform φ into an equivalent simple formula φ'.

Step 2. Find a φ'-legal pointed UPAL model $\mathfrak{M}, w_0$ such that $\mathfrak{M}, w_0 \not\Vdash \varphi'$.

Step 3. Obtain $\mathcal{N}$ and τ from $\mathfrak{M}, w_0$ as in §4 so that $\mathcal{N}, w_0 \not\models (\varphi')^\tau$.

Since $\varphi \leftrightarrow \varphi'$ is schematically valid in PAL, we have $\mathcal{N}, w \not\models (\varphi)^\tau$, as desired. The key to this method is that the construction in §4 also establishes the following variant of Proposition 1:

Proposition 3. For any *simple* $\varphi \in \mathcal{L}^\omega_{\mathsf{PAL}}$, if there is a *$\varphi$-legal* UPAL model $\mathfrak{M} = \langle M, \{\mathcal{R}_a\}_{a \in \mathsf{Agt}}, \{\mathcal{R}_\psi\}_{\psi \in \mathcal{L}^\omega_{\mathsf{PAL}}}, \mathcal{V}\rangle$ with $w_0 \in M$ such that $\mathfrak{M}, w_0 \not\Vdash \varphi$, then there is a

PAL model $\mathcal{N} = \langle N_0, \{S_a\}_{a\in\mathsf{Agt}}, U\rangle$ with $w_0 \in N_0$ and a substitution τ such that $\mathcal{N}, w_0 \not\models (\varphi)^\tau$.

This proposition holds because if φ is already simple, then the only properties of $\mathfrak{M}$ used in the proof of Fact 3 are **$\top$-reflexivity**, **functionality**, **extensionality for** φ and **truthfulness for** φ, which are part of φ-legality.

Finally, if φ does not contain any occurrence of a dynamic operator in the scope of any other, then we can simply skip Step 1 and do Steps 2 and 3 for φ itself. One can check that the construction in §4 works not only with a simple formula, but more generally with any formula with the scope restriction.

Example 1. Consider the PAL-valid formula $\varphi := [p]p$, which is already simple. Let us try to falsify φ in a φ-legal UPAL model. The obvious first try is $\mathfrak{M}$ in Fig. 1, which is indeed a φ-legal UPAL model, in which all $\mathcal{R}_a$ relations are empty. (We simplify the diagrams by omitting all reflexive $\mathcal{R}_\top$ loops.) However, $\mathfrak{M}$ has an un-PAL-like property: although $\|\top\|^{\mathfrak{M}} \cap \mathcal{R}^*_{\mathsf{Agt}}(w_0) = \|p\|^{\mathfrak{M}} \cap \mathcal{R}^*_{\mathsf{Agt}}(w_0)$, we have $w_0\mathcal{R}_\top w_0$ but not $w_0\mathcal{R}_p w_0$. (See §6 for why this is un-PAL-like.) To eliminate this property, we modify $\mathfrak{M}$ to $\mathfrak{N} = \langle N, \{\mathcal{S}_a\}_{a\in\mathsf{Agt}}, \{\mathcal{S}_\psi\}_{\psi\in\mathcal{L}^\omega_{\mathsf{PAL}}}, \mathcal{U}\rangle$ in Fig. 1 as in Lemma 3.[7] Next, following the procedure in §4, we obtain the PAL model $\mathcal{N} = \langle N_0, \{S_a\}_{a\in\mathsf{Agt}}, U\rangle$ in Fig. 1 and the substitution τ given below.

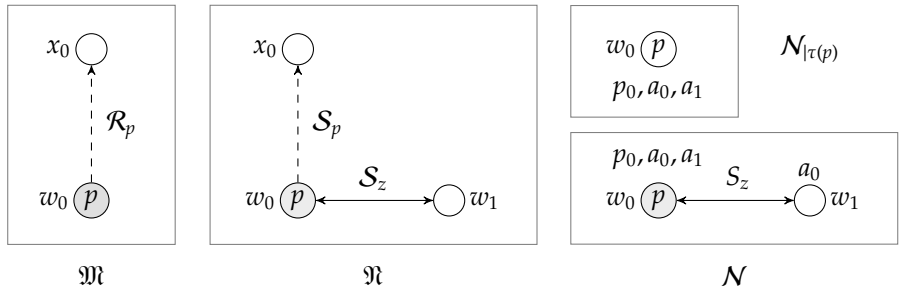


Figure 1: UPAL and PAL Models for Example 1

Where $A_0 := \top$, $A_1 := p$, and a_0, a_1, p_0, and p_1 are the new atoms, we define the valuation U in $\mathcal{N}$ such that:

[7] In fact, the construction of Lemma 3 would connect w_0 to x_0 by $\mathcal{R}_z$, but note that we can always connect w_0 to a new point falsifying $\alpha \leftrightarrow \beta$ (in this case, $\top \leftrightarrow p$) instead.

$$U(a_0) = \|A_0\|^{\mathfrak{N}} \cap N_0 = \{w_0, w_1\};$$
$$U(a_1) = \|A_1\|^{\mathfrak{N}} \cap N_0 = \{w_0\};$$
$$U(p_0) = \{w \in N_0 \mid \exists u\colon w\mathcal{S}_{A_0}u \text{ and } u \in \mathcal{U}(p)\} = \{w_0\};$$
$$U(p_1) = \{w \in N_0 \mid \exists u\colon w\mathcal{S}_{A_1}u \text{ and } u \in \mathcal{U}(p)\} = \emptyset.$$

Defining the function s such that

$$s(i) = \{j \mid 0 \leq j \leq n \text{ and } \|A_j\|^{\mathfrak{N}} \cap N_0 \subsetneq \|A_i\|^{\mathfrak{N}} \cap N_0\},$$

we have $s(0) = \{1\}$ and $s(1) = \emptyset$. Defining $\tau(p) = \gamma_0 \wedge \cdots \wedge \gamma_n$ such that

$$\gamma_i := (\Box_z a_i \wedge \bigwedge_{j \in s(i)} \neg\Box_z a_j) \to p_i,$$

we have

$$\tau(p) = ((\Box_z a_0 \wedge \neg\Box_z a_1) \to p_0) \wedge (\Box_z a_1 \to p_1).$$

Observe:

$$[\![(\Box_z a_0 \wedge \neg\Box_z a_1) \to p_0]\!]^{\mathcal{N}} = \{w_0\};$$
$$[\![\Box_z a_1 \to p_1]\!]^{\mathcal{N}} = \{w_0, w_1\};$$
$$[\![\tau(p)]\!]^{\mathcal{N}} = \{w_0\}.$$

Hence $\mathcal{N}_{|\tau(p)}$ is the model displayed in the upper-right in Fig. 1. Observe:

$$[\![(\Box_z a_0 \wedge \neg\Box_z a_1) \to p_0]\!]^{\mathcal{N}_{|\tau(p)}} = \{w_0\};$$
$$[\![\Box_z a_1 \to p_1]\!]^{\mathcal{N}_{|\tau(p)}} = \emptyset;$$
$$[\![\tau(p)]\!]^{\mathcal{N}_{|\tau(p)}} = \emptyset.$$

Hence $\mathcal{N}, w_0 \not\models ([p]p)^{\tau}$, so our starting formula φ is not schematically valid over PAL models.

Example 2. Consider the PAL-valid formula $\varphi := [p \wedge \neg\Box_b p]\neg(p \wedge \neg\Box_b p)$.[8] Let us try to falsify φ in a φ-legal UPAL model. The obvious first try is the model $\mathfrak{N}$ in

[8]Here we could transform $\varphi := [p \wedge \neg\Box_b p]\neg(p \wedge \neg\Box_b p)$ into the simple

$$\varphi' := (p \wedge \neg\Box_b p) \to \neg([p \wedge \neg\Box_b p]p \wedge ((p \wedge \neg\Box_b p) \to \neg((p \wedge \neg\Box_b p) \to \Box_b[p \wedge \neg\Box_b p]p))),$$

but as noted before Example 1, if φ does not contain any occurrence of a dynamic operator in the scope of any other, then we can skip Step 1 and do Steps 2 and 3 for φ itself.

Fig. 2. However, $\mathfrak{A}$ is not φ-legal, since it violates ψz**-commutativity** for $\psi := p \wedge \neg\Box_b p$. By modifying $\mathfrak{A}$ to $\mathfrak{N} = \langle N, \{\mathcal{S}_a\}_{a\in\mathsf{Agt}}, \{\mathcal{S}_\psi\}_{\psi\in\mathcal{L}^{\omega}_{\mathsf{PAL}}}, \mathcal{U}\rangle$ in Fig. 2, we obtain a φ-legal UPAL model with $\mathfrak{N}, w_0 \nVdash \varphi$. (In this case, the transformation of Lemma 3 is uncecessary, since the condition of Lemma 3.2 is already satisfied by $\mathfrak{N}$.) Following the procedure of §4, we obtain the PAL model $\mathcal{N} = \langle N_0, \{S_a\}_{a\in\mathsf{Agt}}, U\rangle$ in Fig. 3 and the substitution τ given below.

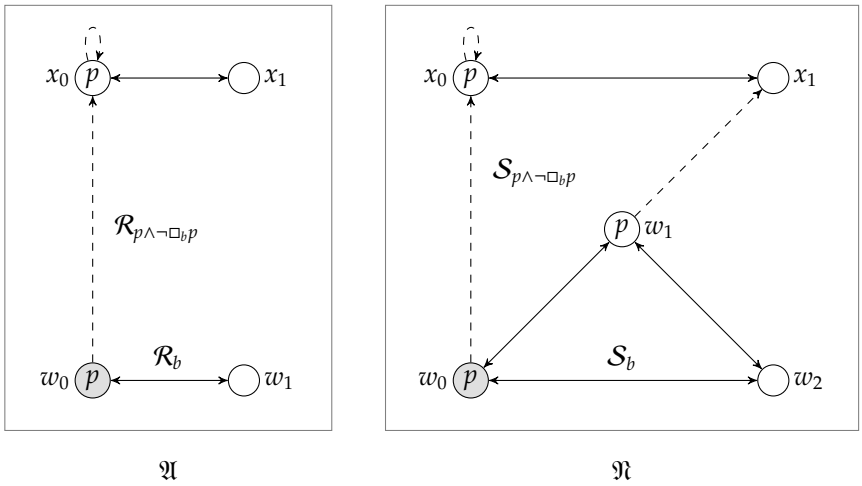


Figure 2: UPAL Models for Example 2

Where $A_0 := \top$, $A_1 := p \wedge \neg\Box_b p$, and a_0, a_1, p_0, and p_1 are the new atoms, we define the valuation U in $\mathcal{N}$ such that:

$$U(a_0) = \|A_0\|^{\mathfrak{N}} \cap N_0 = \{w_0, w_1, w_2\};$$
$$U(a_1) = \|A_1\|^{\mathfrak{N}} \cap N_0 = \{w_0, w_1\};$$
$$U(p_0) = \{w \in N_0 \mid \exists u\colon w\mathcal{S}_{A_0}u \text{ and } u \in \mathcal{U}(p)\} = \{w_0, w_1\};$$
$$U(p_1) = \{w \in N_0 \mid \exists u\colon w\mathcal{S}_{A_1}u \text{ and } u \in \mathcal{U}(p)\} = \{w_0\}.$$

Defining the function s as before, we have $s(0) = \{1\}$ and $s(1) = \emptyset$. Since this is the same s as in Example 1, the substitution is also the same:

$$\tau(p) = ((\Box_z a_0 \wedge \neg\Box_z a_1) \to p_0) \wedge (\Box_z a_1 \to p_1).$$

Note that since the construction of $\mathcal{N}$ from $\mathfrak{N}$ is such that $S_z = S_b$, we can simply take $\Box_z$ to be $\Box_b$ in $\tau(p)$, so that $\mathsf{Agt}((\varphi)^\tau) = \mathsf{Agt}(\varphi) = \{b\}$.

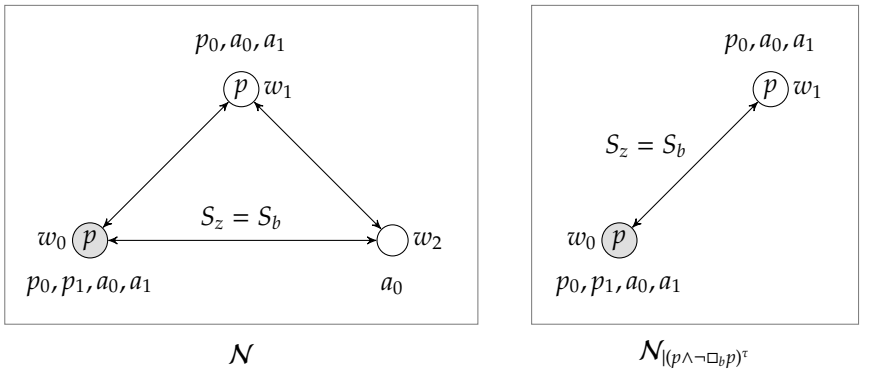


Figure 3: PAL Models for Example 2

Observe:

$$[\![(\Box_z a_0 \wedge \neg\Box_z a_1) \to p_0]\!]^{\mathcal{N}} = \{w_0, w_1\};$$
$$[\![\Box_z a_1 \to p_1]\!]^{\mathcal{N}} = \{w_0, w_1, w_2\};$$
$$[\![\tau(p)]\!]^{\mathcal{N}} = \{w_0, w_1\};$$
$$[\![\tau(p) \wedge \neg\Box_b \tau(p)]\!]^{\mathcal{N}} = \{w_0, w_1\}.$$

Hence $\mathcal{N}_{|(p\wedge\neg\Box_b p)^\tau}$ is the model displayed on the right in Fig. 3. Observe:

$$[\![(\Box_z a_0 \wedge \neg\Box_z a_1) \to p_0]\!]^{\mathcal{N}_{|(p\wedge\neg\Box_b p)^\tau}} = \{w_0, w_1\};$$
$$[\![\Box_z a_1 \to p_1]\!]^{\mathcal{N}_{|(p\wedge\neg\Box_b p)^\tau}} = \{w_0\};$$
$$[\![\tau(p)]\!]^{\mathcal{N}_{|(p\wedge\neg\Box_b p)^\tau}} = \{w_0\};$$
$$[\![\tau(p) \wedge \neg\Box_b \tau(p)]\!]^{\mathcal{N}_{|(p\wedge\neg\Box_b p)^\tau}} = \{w_0\}.$$

Hence $\mathcal{N}, w_0 \not\models ([p\wedge\neg\Box_b p]\neg(p\wedge\neg\Box_b p))^\tau$, so our starting formula φ is not schematically valid over PAL models.

We invite the reader to work out other examples using UPAL, starting from the other valid but not schematically valid PAL principles mentioned in §1.2.

6 Discussion

In this paper, we have shown that UPAL axiomatizes the substitution core of PAL with infinitely many agents. In this final section, we briefly discuss the axiomatization question for the single-agent and finite-agent cases. For a given language and class of models, the key question is how close we can come to expressing that two formulas are co-extensional in the epistemic submodel generated by the current point. For example, this condition is expressed by the formula $\Box_a^+(\varphi \leftrightarrow \psi)$ (where $\Box_a^+\alpha := \alpha \wedge \Box_a\alpha$) in *single*-agent PAL over *transitive* models. In this case, we get a new schematic validity in PAL:

(inner extensionality) $\quad \Box_a^+(\varphi \leftrightarrow \psi) \rightarrow (\langle\varphi\rangle\alpha \leftrightarrow \langle\psi\rangle\alpha)$.

The corresponding legality condition for UPAL models is:

(**inner extensionality**) if $\|\varphi\|^{\mathfrak{M}} \cap \mathcal{R}_a(w) = \|\psi\|^{\mathfrak{M}} \cap \mathcal{R}_a(w)$,
then $w\mathcal{R}_\varphi v$ iff $w\mathcal{R}_\psi v$,

which does not follow from any of the other legality conditions.

For multiple agents, we cannot in general express the co-extensionality of two formulas in the epistemic submodel generated by the current point; however, if we allow our models to be *non-serial*, then we do get related schematic validities for the single and finite-agent cases that are not derivable in UPAL-K_n (where the antecedent can be written using $\Box_a$ operators and $\bot$):[9]

(FPE) "all $\mathcal{R}_{\mathsf{Agt}}$-paths from the current point are of length $\leq n$" $\rightarrow$
$(E^n(\varphi \leftrightarrow \psi) \rightarrow (\langle\varphi\rangle\alpha \leftrightarrow \langle\psi\rangle\alpha))$,

[9] The (FPE) axioms are also schematically valid over serial models, because the antecedent is always false, but then they are also derivable using the seriality axiom $\Diamond_a\top$.

where

$$E^0\alpha := \alpha \wedge \bigwedge_{a \in \mathsf{Agt}} \Box_a \alpha \text{ and } E^n\alpha := \alpha \wedge E^0 E^{n-1}\alpha.$$

The corresponding legality condition for UPAL is:

(**FPE**) if $\mathcal{R}^*_{\mathsf{Agt}}(w)$ is path-finite and $\|\varphi\|^{\mathfrak{M}} \cap \mathcal{R}^*_{\mathsf{Agt}}(w) = \|\psi\|^{\mathfrak{M}} \cap \mathcal{R}^*_{\mathsf{Agt}}(w)$, then $w\mathcal{R}_\varphi v$ iff $w\mathcal{R}_\psi v$, where

$\mathcal{R}^*_{\mathsf{Agt}}(w)$ is path-finite just in case every $\mathcal{R}_{\mathsf{Agt}}$-path from w ends in a dead-end point in finitely many steps. This shows why the axiomatization of the substitution core of PAL-$\mathbf{K}_\omega$ is more elegant than that of PAL-$\mathbf{K}_n$: with infinitely many agents we cannot express the "everybody knows" modality E, so we do not need to add to UPAL the infinitely many FPE axioms.

Finally, if we consider PAL with the standard *common knowledge* operator C, then we can express co-extensionality in the generated epistemic submodel using the formula $C(\varphi \leftrightarrow \psi)$, in which case we get the new schematic validity

(common extensionality) $C(\varphi \leftrightarrow \psi) \rightarrow (\langle\varphi\rangle\alpha \leftrightarrow \langle\psi\rangle\alpha)$.

The corresponding legality condition in UPAL is:

(**common extensionality**) if $\|\varphi\|^{\mathfrak{M}} \cap \mathcal{R}^*_{\mathsf{Agt}}(w) = \|\psi\|^{\mathfrak{M}} \cap \mathcal{R}^*_{\mathsf{Agt}}(w)$, then $w\mathcal{R}_\varphi v$ iff $w\mathcal{R}_\psi v$.

We leave it to future work to give analyses for the above languages analogous to the analysis we have given here for $\mathcal{L}^\omega_{\mathsf{PAL}}$. A natural next step is to axiomatize the substitution core of the system of PAL-RC (van Benthem, J. et al. 2006) with relativized common knowledge. Relativized common knowledge $C(\varphi, \psi)$ is interpreted in UPAL models exactly as in PAL models. We conjecture that UPAL together with the relativized common knowledge reduction axiom $\langle p\rangle C(q, r) \leftrightarrow C(\langle p\rangle q, \langle p\rangle r)$, the common extensionality axiom above, and the appropriate base

logic (see van Benthem, J. et al. 2006) axiomatizes the substitution core of PAL-RC with finitely or infinitely many agents over any of the model classes we have discussed. Indeed, it can be shown using arguments similar to those of §4 that the set of formulas in the language $\mathcal{L}^{\kappa}_{\text{PAL-RC}}$ that are valid over legal UPAL models with **common extensionality** is exactly the substitution core of PAL-RC. Hence it only remains to prove that the extended system just described—call it UPAL-RC—is sound and complete for this model class. Such a proof requires a finite canonical model construction to deal with common knowledge, and we cannot go into the details here.

Another natural step is to attempt to apply the strategies of this paper to axiomatize the substitution cores of other dynamics epistemic logics, including the full system of DEL (van Benthem, J. 2011a, Ch. 4). One may imagine a general program of "uniformizing" dynamic epistemic logics, of which UPAL is only the beginning.

Acknowledgements We thank Johan van Benthem for stimulating our interest in the topic of this paper and the AiML referees for very helpful comments.

References

van Benthem, J. What One May Come to Know. *Analysis*, 64(2):95–105, 2004.

van Benthem, J. Open Problems in Logical Dynamics. In D. Gabbay, S. Goncharov, and M. Zakharyashev, editors, *Mathematical Problems from Applied Logic I*, pages 137–192. Springer, 2006a.

van Benthem, J. One is a Lonely Number: Logic and Communication. In Z. Chatzidakis, P. Koepke, and W. Pohlers, editors, *Logic Colloquium '02*, pages 96–129. ASL & A.K. Peters, 2006b.

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

van Benthem, J. Two Logical Faces of Belief Revision. manuscript, 2011b.

van Benthem, J., J. van Eijck, and B. Kooi. Logics of communication and change. *Information and Computation*, 204(11):1620–1662, 2006.

van Ditmarsch, H. and B. Kooi. The Secret of My Success. *Synthese*, 151: 201–232, 2006.

van Ditmarsch, H., W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*. Springer, 2008.

van Ditmarsch, H., W. van der Hoek, and P. Iliev. Everything is Knowable – How to Get to Know *Whether* a Proposition is True. *Theoria*, 78(2):93–114, 2011.

L. Aqvist. Modal Logic with Subjunctive Conditionals and Dispositional Predicates. *Journal of Philosophical Logic*, 2:1–76, 1973.

P. Balbiani, A. Baltag, H. van Ditmarsch, A. Herzig, T. Hoshi, and T. de Lima. 'Knowable' as 'known after an announcement'. *The Review of Symbolic Logic*, 1:305–334, 2008.

R. Ballarin. Validity and Necessity. *Journal of Philosophical Logic*, 34:275–303, 2005.

A. Baltag, L. Moss, and S. Solecki. The Logic of Public Announcements, Common Knowledge and Private Suspicions. In I. Gilboa, editor, *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 98)*, pages 43–56. Morgan Kaufmann, 1998.

P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Cambridge University Press, 2001.

J. P. Burgess. Which Modal Models are the Right Ones (for Logical Necessity)? *Theoria*, 18:145–158, 2003.

S. R. Buss. The Modal Logic of Pure Provability. *Notre Dame Journal of Formal Logic*, 31(2):225–231, 1990.

R. Carnap. Modalities and Quantification. *The Journal of Symbolic Logic*, 11(2): 33–64, 1946.

I. A. Ciardelli. Inquisitive semantics and intermediate logics. Master's thesis, University of Amsterdam, 2009. ILLC Master of Logic Thesis Series MoL-2009-11.

J. Gerbrandy and W. Groenevelt. Reasoning about Information Change. *Journal of Logic, Language and Information*, 6(2):147–169, 1997.

R. Goldblatt. *Logics of Time and Computation*. CSLI Press, 1992.

J. Y. Halpern. Should Knowledge Entail Belief? *Journal of Philosophical Logic*, 25:483–494, 1996.

W. H. Holliday and T. F. Icard, III. Moorean Phenomena in Epistemic Logic. In L. Beklemishev, V. Goranko, and V. Shehtman, editors, *Advances in Modal Logic*, volume 8 of , pages 178–199. College Publications, 2010.

W. H. Holliday, T. Hoshi, and T. F. Icard, III. Schematic Validity in Dynamic Epistemic Logic: Decidability. In H. van Ditmarsch, J. Lang, and S. Ju, editors, *Proceedings of the Third International Workshop on Logic, Rationality and Interaction (LORI-III)*, volume 6953 of *Lecture Notes in Artificial Intelligence*, pages 87–96. Springer, 2011.

W. H. Holliday, T. Hoshi, and T. F. Icard, III. Information Dynamics and Uniform Substitution. manuscript, 2012.

B. Kooi. Expressivity and completeness for public update logics via reduction axioms. *Journal of Applied Non-Classical Logics*, 17(2):231–253, 2007.

M. Ma. Mathematics of Public Announcements. In H. van Ditmarsch, J. Lang, and S. Ju, editors, *Proceedings of the Third International Workshop on Logic, Rationality and Interaction (LORI-III)*, volume 6953 of *Lecture Notes in Artificial Intelligence*, pages 193–205. Springer, 2011.

S. Mascarenhas. Inquisitive semantics and logic. Master's thesis, University of Amsterdam, 2009. ILLC Master of Logic Thesis Series MoL-2009-18.

J. Plaza. Logics of public communications. In M. Emrich, M. Pfeifer, M. Hadzikadic, and Z. Ras, editors, *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems*, pages 201–216. Oak Ridge National Laboratory, 1989.

M. Reynolds. An Axiomatization of Full Computation Tree Logic. *The Journal of Symbolic Logic*, 66(3):1011–1057, 2001.

G. Schurz. Logic, matter of form, and closure under substitution. In M. Bilkova and L. Behounek, editors, *The Logica Yearbook*, pages 33–46. Filosofia, 2005.

K. Segerberg. Two-dimensional modal logic. *Journal of Philosophical Logic*, 2 (1):77–96, 1973.

Y. Wang. On Axiomatizations of PAL. In H. van Ditmarsch, J. Lang, and S. Ju, editors, *Proceedings of the Third International Workshop on Logic, Rationality and Interaction (LORI-III)*, volume 6953 of *Lecture Notes in Artificial Intelligence*, pages 314–327. Springer, 2011.

Y. Wang and Q. Cao. On Axiomatizations of Public Announcement Logic. manuscript, 2012.

Characterizing Definability of Second-Order Generalized Quantifiers

Juha Kontinen and Jakub Szymanik

Department of Mathematics and Statistics, University of Helsinki,
Institute of Artificial Intelligence, University of Groningen
juha.kontinen@helsinki.fi, jakub.szymanik@gmail.com

Abstract

We study definability of second-order generalized quantifiers. We show that the question whether a second-order generalized quantifier Q_1 is definable in terms of another quantifier Q_2, the base logic being monadic second-order logic, reduces to the question if a quantifier $Q_1^\star$ is definable in $\mathrm{FO}(Q_2^\star, <, +, \times)$ for certain first-order quantifiers $Q_1^\star$ and $Q_2^\star$. We use our characterization to show new definability and non-definability results for second-order generalized quantifiers. We also show that the monadic second-order majority quantifier Most[1] is not definable in second-order logic.[1]

1 Introduction

The notion of generalized quantifier goes back to Mostowski (1957) and Lindström (1966). Generalized quantifiers were first mainly studied in the frame-

[1]This is an extended version of Kontinen and Szymanik (2011)

work of model theory. The study of generalized quantifiers extended to the context of finite model theory via applications to descriptive complexity theory. We refer to Väänänen (1999) and Ebbinghaus and Flum (1999) for surveys of first-order generalized quantifiers in finite model theory. Generalized quantifiers have been also extensively studied in the formal semantics of natural language (see Peters and Westerståhl 2006, for a survey).

The study of second-order generalized quantifiers is a relatively new and unexplored area in finite model theory. On the other hand, second-order logic (SO) and its many fragments have been studied extensively starting from Fagin's characterization of NP in terms of existential second-order logic (Fagin 1974). Second-order generalized quantifiers were first studied in the context of finite structures by Burtschick and Vollmer (1998). Shortly after, Andersson (2002) studied the expressive power of families of second-order generalized quantifiers determined by the syntactic types of quantifiers. Kontinen (2005, 2010, 2006) studied definability questions of second-order generalized quantifiers. In the case of first-order quantifiers, definability of a quantifier Q in a logic $\mathcal{L}$ means that the class of structures, used to interpret Q, is axiomatizable in $\mathcal{L}$. In the second-order case, the analogous concept of definability was formulated in Kontinen (2005; 2010). In this article, we give a computationally motivated characterization for the notion of definability of second-order generalized quantifiers.

Burtschick and Vollmer (1998) noticed that second-order generalized quantifiers can be used to logically characterize complexity classes defined in terms of so-called *Leaf Languages*. The leaf languages approach in computational complexity theory, introduced by Bovet, Crescenzi, and Silvestri (1992), is a unifying approach to define complexity classes. The central idea behind this approach is to generalize the conditions under which, e.g., a Turing machine or an automaton accepts its input. Many complexity classes can be defined in this context in terms of suitable leaf languages. On the other hand, a complexity class defined in terms of a leaf language B can be under certain conditions characterized logically by a logic of the form:

$$Q_B\,\mathrm{FO},$$

where Q_B is a second-order generalized quantifier corresponding to the language B. In the context of leaf languages, polynomial time non-deterministic Turing machines can be sometimes replaced by non-deterministic finite automata (so-called finite leaf automata) without a significant decrease in complexity (Peichl and Vollmer 2001). Galota and Vollmer (2001) showed that com-

plexity classes defined by finite leaf automata can be logically characterized in terms of monadic second-order generalized quantifiers. This result nicely extends the well known Büchi and Elgot (1958), Büchi (1962), Trakhtenbrot (1961) characterization of regular languages in terms of monadic second-order logic (MSO).

The definability theory of second-order generalized quantifiers has some similarities and differences compared to that of first-order generalized quantifiers. For example, it was observed by Kontinen (2005) that the binary second-order existential quantifier cannot be defined in terms of any monadic second-order generalized quantifiers. This result is in contrast with the fact (a corollary of a result of Andersson (2002)) that all classes of finite first-order structures are already definable in terms of monadic second-order generalized quantifiers.

In this paper we prove a general result characterizing the question when a quantifier Q is definable in MSO(Q', +), where + denotes the built-in addition relation. We assume the built-in addition in order to unleash the expressive power embodied by MSO. Recall that, while MSO corresponds to regular languages over strings, MSO(+) corresponds to the linear fragment of the polynomial hierarchy (LINH) on strings (More and Olive 1997). Some of our results can be generalized to the case where the base logic is full second-order logic instead of MSO(+).

Our characterization is based on the idea connecting oracle separation results with lower bound results for small constant depth circuits (see, e.g., Furst et al. 1984, Yao 1985, Håstad 1987, Torán 1988, Vollmer 1998). We show that a second-order generalized quantifier Q_1 is definable in the logic MSO(Q_2, +) iff for certain first-order encodings $Q_i^\star$ of Q_i, $Q_1^\star$ is definable in FO($Q_2^\star$, +, ×). It is worth noting that the latter condition implies that $Q_1^\star$ is AC^0 (Turing) reducible to $Q_2^\star$. We use our characterization to show new definability and non-definability results for second-order generalized quantifiers. In particular, we show that the monadic second-order majority quantifier Most^1 is not definable in second-order logic.

This answers the question left open in Kontinen and Szymanik (2008) (see also Szymanik 2009), where second-order generalized quantifiers were used to model collective quantification in natural language, for example:

1. Most girls gathered.
2. All soldiers surrounded the Alamo

The common strategy in formalizing collective quantification has been to de-

fine the meanings of collective determiners, quantifying over collections, using certain type-shifting operations. These operations, i.e., lifts, define the collective interpretations of determiners systematically from the standard meanings of quantifiers (see. e.g., van der Does 1993, Winter 2001). In Kontinen and Szymanik (2008) we show that all these lifts are definable in second-order logic. In this paper we prove that some collective quantifiers (second-order generalized quantifiers) are not definable in second-order logic. Therefore, there is no second-order definable lift expressing their collective meaning. This is clearly a restriction of the type-shifting approach. One possible alternative would be to use second-order generalized quantifiers in the study of collective semantics, as we already proposed in Kontinen and Szymanik (2008). However, as it follows from this paper the computational complexity of such approach is excessive and hence it may not be a plausible model of collective quantification in natural language (see Szymanik and Zajenkowski 2010, Szymanik 2010, for a discussion of computational restrictions in natural language semantics). Hence, it may be wise to turn in the direction of another well-known way of studying collective quantification in natural language, the many-sorted (algebraic) tradition (see Lønning 2011). Another linguistic interpretation of our results might be that computational complexity restricts the expressive power of everyday language (see Mostowski and Szymanik 2012). Namely, even though natural language can in principle realize collective quantifiers non-definable in second-order logic, its everyday fragment does not contain such constructions due to their high complexity.

2 Preliminaries

In this article all structures are assumed to be finite. The universe of a structure $\mathfrak{A}$ is denoted by A. Without loss of generality, we may assume that A is always of the form $\{0, \ldots, m\}$ for some $m \in \mathbb{N}$. For a logic $\mathcal{L}$, the set of τ-formulas of $\mathcal{L}$ is denoted by $\mathcal{L}[\tau]$. If φ is a τ-sentence, then the class of τ-models of φ is denoted by $\mathrm{Mod}(\varphi)$. A class K of τ-models is said to be axiomatizable in a logic $\mathcal{L}$, if $K = \mathrm{Mod}(\varphi)$ for some sentence $\varphi \in \mathcal{L}[\tau]$. For logics $\mathcal{L}$ and $\mathcal{L}'$, we write $\mathcal{L} \leq \mathcal{L}'$, if for every τ and every sentence $\varphi \in \mathcal{L}[\tau]$ there is a sentence $\psi \in \mathcal{L}'[\tau]$ such that $\mathrm{Mod}(\varphi) = \mathrm{Mod}(\psi)$. The set of natural numbers is denoted by $\mathbb{N}$ and $\mathbb{N}^*$ denotes the set $\mathbb{N} \setminus \{0\}$.

Sometimes we assume that our structures (and logics) are equipped with auxil-

iary built-in relations. In addition to the built-in ordering $<$, which is interpreted naturally, we also use the ternary relations $+$ and $\times$. The relations $+$ and $\times$ are defined as

$$\begin{aligned} +(i,j,k) &\quad\Leftrightarrow\quad i+j=k, \\ \times(i,j,k) &\quad\Leftrightarrow\quad i\times j=k. \end{aligned}$$

The relation BIT is a further important relation which is defined by: $\mathrm{BIT}(a,j)$ holds iff the bit of order 2^j is 1 in the binary representation $\mathrm{bin}(a)$ of a. The presence of built-in relations is signalled, e.g., by the notation $\mathrm{FO}(<)$. It is well known that $\mathrm{FO}(<,+,\times) \equiv \mathrm{FO}(<,\mathrm{BIT})$ (see Immerman 1999).

Note that $<$ is easily definable in $\mathrm{FO}(+)$ and hence, in the presence of $+$, we sometimes do not mention $<$ explicitly.

We assume that the reader is familiar with the basics of computational complexity theory. Below, we recall certain results from descriptive complexity theory. It is instructive to note that many of the logics considered in this article correspond to interesting complexity classes. We mention first the logic $\mathrm{FO}(<,+,\times)$ which corresponds exactly to the so-called logarithmic hierarchy (LH). This class is the logarithmic analogue of the polynomial hierarchy (PH), corresponding to SO (Stockmeyer 1976), defined in terms of alternating Turing machines (ATM) running in polynomial time with $O(1)$ alternations. In between LH and PH we have the linear hierarchy (LINH) corresponding to the logic MSO(+) over strings (More and Olive 1997).

In this article also majority quantifiers are discussed and studied. It is well-known that majority quantifiers can be used to logically characterize counting computations. The following counting hierarchies are relevant for this article: the logarithmic counting hierarchy (LCH), the linear counting hierarchy (LINCH), and the (polynomial) counting hierarchy (CH) all of which can be defined, with analogous resource bounds as LH, LINH, and PH, in terms of so-called Threshold Turing machines (Parberry and Schnitger 1988). On the logical side, majority quantifiers (defined in Section 2.1) can be used to provide logical counterparts for these classes: $\mathrm{FO}(\mathrm{M},+,\times) \equiv \mathrm{LCH}$ (Barrington et al. 1990), $\mathrm{FO}(\mathrm{Most}^1,<) \equiv \mathrm{LINCH}$ (over strings) (Kontinen and Niemistö 2011), and $\mathrm{FO}(\mathrm{Most}^k)_{k\in\mathbb{N}^*} \equiv \mathrm{CH}$ (Kontinen 2009). Furthermore, in circuit complexity, it is known that LH corresponds exactly to DLOGTIME-uniform AC^0 and LCH to DLOGTIME-uniform TC^0 (Barrington et al. 1990). Also, DLOGTIME-uniform $\mathrm{AC}^0[p]$ (AC^0 with unbounded fan-in MOD_p gates) corresponds on the logical side to $\mathrm{FO}(\mathrm{D}_k,+,\times)$ (Barrington et al. 1990).

2.1 Generalized quantifiers

In this section we briefly recall some basics of generalized quantifiers.

Let $\tau = \{P_1, \ldots, P_r\}$ be a relational vocabulary, where P_i is l_i-ary for $1 \leq i \leq r$, and Q a class of τ-structures closed under isomorphisms. The class Q gives rise to a generalized quantifier which we also denote by Q. The tuple $s = (l_1, \ldots, l_r)$ is the *type* of the quantifier Q.

Definition 2.1. The extension FO(Q) of first-order logic by a quantifier Q is defined as follows:

1. The formula formation rules of FO are extended by the rule: if for $1 \leq i \leq r$, $\varphi_i(\overline{x}_i)$ is a formula and $\overline{x}_i$ is an l_i-tuple of pairwise distinct variables then $Q\overline{x}_1, \ldots, \overline{x}_r\,(\varphi_1(\overline{x}_1), \ldots, \varphi_r(\overline{x}_r))$ is a formula.
2. The satisfaction relation of FO is extended by the rule:

$$\mathfrak{A} \models Q\overline{x}_1, \ldots, \overline{x}_r\,(\varphi_1(\overline{x}_1), \ldots, \varphi_r(\overline{x}_r)) \text{ iff } (A, \varphi_1^{\mathfrak{A}}, \ldots, \varphi_r^{\mathfrak{A}}) \in Q,$$

where $\varphi_i^{\mathfrak{A}} = \{\overline{a} \in A^{l_i} \mid \mathfrak{A} \models \varphi_i(\overline{a})\}$.

We say that a quantifier Q is definable in a logic $\mathcal{L}$ if the class Q is axiomatizable in $\mathcal{L}$. Note that Q is trivially definable in FO(Q). If $\mathcal{L}$ has the substitution property and is closed under FO-operations, then definability of Q in $\mathcal{L}$ implies that FO(Q) $\leq \mathcal{L}$. So, among such logics, FO(Q) is the minimal logic in which Q is definable.

Example 1. The following quantifiers will be discussed in the following sections. Suppose $S \subseteq \mathbb{N}$ and $k \in \mathbb{N}$.

$$\begin{aligned} \exists &= \{(A, P) \mid P \subseteq A \text{ and } P \neq \emptyset\} \\ \mathrm{M} &= \{(A, P) \mid P \subseteq A \text{ and } |P| > |A|/2\} \\ Q_S &= \{(A, P) \mid P \subseteq A \text{ and } |P| \in S\} \\ \mathrm{I} &= \{(A, P_1, P_2) \mid P_i \subseteq A \text{ and } |P_1| = |P_2|\} \end{aligned}$$

If S is of the form $\{kn \mid n \in \mathbb{N}\}$ for some $k \in \mathbb{N}$, we denote Q_S by D_k.

We will also refer to the *vectorizations* of the quantifiers D_k and M later. The nth vectorization of D_k is the following quantifier

$$\mathrm{D}_k^n = \{(A, P) \mid P \subseteq A^n \text{ and } |P| = 0 \bmod k\},$$

and the nth vectorization of M is

$$\mathrm{M}^n = \{(A, P) \mid P \subseteq A^n \text{ and } |P| > |A^n|/2\}.$$

Let us then turn to second-order generalized quantifiers. Let $t = (s_1, \ldots, s_w)$, where $s_i = (l^i_1, \ldots, l^i_{r_i})$ is a tuple of positive integers for $1 \leq i \leq w$. A second-order structure of type t is a structure of the form $(A, P_1, \ldots, P_w)$, where $P_i \subseteq \mathcal{P}(A^{l^i_1}) \times \cdots \times \mathcal{P}(A^{l^i_{r_i}})$.

Definition 2.2. A second-order generalized quantifier Q of type t is a class of structures of type t such that Q is closed under isomorphisms.

A quantifier Q is *monadic* if $l^i_j = 1$ for all $1 \leq j \leq r_i$ and $1 \leq i \leq w$. Let us look at some examples of second-order generalized quantifiers.

Example 2. Suppose $S \subseteq \mathbb{N}$ and $k \in \mathbb{N}$.

$$\begin{aligned}
\exists^2_k &= \{(A, P) \mid P \subseteq \mathcal{P}(A^k) \text{ and } P \neq \emptyset\} \\
\text{Even} &= \{(A, P) \mid P \subseteq \mathcal{P}(A) \text{ and } |P| \text{ is even}\} \\
\text{Even}' &= \{(A, P) \mid P \subseteq \mathcal{P}(A) \text{ and } \forall X \in P\ (|X| \text{ is even})\} \\
\text{Most}^k &= \{(A, P) \mid P \subseteq \mathcal{P}(A^k) \text{ and } |P| > 2^{|A|^k - 1}\} \\
\mathrm{I}^2 &= \{(A, P_1, P_2) \mid P_i \subseteq \mathcal{P}(A) \text{ and } |P_1| = |P_2|\} \\
Q_S &= \{(A, P) \mid P \subseteq \mathcal{P}(A) \text{ and } |P| \in S\}
\end{aligned}$$

Analogously to the first-order case, if S is of the form $\{kn \mid n \in \mathbb{N}\}$ for some $k \in \mathbb{N}$, we denote Q_S by $\mathcal{D}_k$.

The first example is the familiar k-ary second-order existential quantifier. The quantifier Even says that a formula holds for an even number of subsets of the universe. On the other hand, the quantifier Even′ says that all the subsets satisfying a formula have an even cardinality. The quantifier Most^k is the k-ary second-order version of M expressing that a formula holds for more than half of the k-ary relations.

Definition 2.3. The extension FO(Q) of FO by a quantifier Q is defined as follows:

1. The formula formation rules of FO are extended by the rule: if for $1 \leq i \leq w$, $\varphi_i(\overline{X}_i)$ is a formula and $\overline{X}_i = (X_{1,i}, \ldots, X_{r_i,i})$ is a tuple of pairwise distinct

predicate variables such that the arity of $X_{j,i}$ is l^i_j for $1 \leq j \leq r_i$, then

$$Q\overline{X}_1, \ldots, \overline{X}_w\,(\varphi_1(\overline{X}_1), \ldots, \varphi_w(\overline{X}_w))$$

is a formula.

2. Satisfaction relation of FO is extended by the rule:

$$\mathfrak{A} \models Q\overline{X}_1, \ldots, \overline{X}_w\,(\varphi_1, \ldots, \varphi_w) \text{ iff } (A, \varphi_1^{\mathfrak{A}}, \ldots, \varphi_w^{\mathfrak{A}}) \in Q,$$

where $\varphi_i^{\mathfrak{A}} = \{\overline{R} \in \mathcal{P}(A^{l^i_1}) \times \cdots \times \mathcal{P}(A^{l^i_{r_i}}) \mid \mathfrak{A} \models \varphi_i(\overline{R})\}$.

2.2 Definability

Recall that a first-order generalized quantifier Q is definable in a logic $\mathcal{L}$ if the class Q is axiomatizable in $\mathcal{L}$. This condition can be reformulated as follows assuming $\mathcal{L}$ has the substitution property:

Theorem 1. *A first-order quantifier Q is definable in a logic $\mathcal{L}$ if and only if $\mathcal{L} \equiv \mathcal{L}(Q)$.*

How do we formalize definability for second-order quantifiers? Intuitively, e.g., the monadic second-order existential quantifier $\exists^2_1$ is definable in a logic $\mathcal{L}$ if there is a uniform way to express

$$\exists^2_1 X \psi(X)$$

for any formula $\psi(X)$ in the logic $\mathcal{L}$. Over a model $\mathfrak{A}$, $\psi(X)$ defines a collection of subsets

$$\{C \subseteq A \mid \mathfrak{A} \models \psi(C)\},$$

so the problem is to find a way to express the non-emptyness of this collection in a way which does not depend on the particular formula $\psi(X)$. This was formalized in Kontinen (2010) using second-order relations.

Definition 2.4. Let $\mathcal{L}$ be a logic, $t = (s_1, \ldots, s_w)$ a second-order type, and let $\mathcal{G}_1, \ldots, \mathcal{G}_w$ be first-order quantifier symbols of types $s_1, \ldots, s_w$.

1. The logic $\mathcal{L}(\mathcal{G}_1, \ldots, \mathcal{G}_w)$ is obtained by extending the syntax of $\mathcal{L}$ in terms of the quantifiers $\mathcal{G}_1, \ldots, \mathcal{G}_w$.

2. The models of $\mathcal{L}(\mathcal{G}_1, \ldots, \mathcal{G}_w)$ are of the form $\mathcal{A} = (\mathfrak{A}, G_1, \ldots, G_w)$, where $\mathfrak{A}$ is a first-order model and

$$G_i \subseteq \mathcal{P}(A^{l_1^i}) \times \cdots \times \mathcal{P}(A^{l_{r_i}^i}).$$

3. The quantifiers $\mathcal{G}_i$ are interpreted using the relations G_i:

$$\mathcal{A} \models \mathcal{G}_i \bar{x}_1, \ldots, \bar{x}_{r_i}(\varphi_1(\bar{x}_1), \ldots, \varphi_{r_i}(\bar{x}_{r_i}))$$

iff $(\varphi_1^{\mathcal{A}}, \ldots, \varphi_{r_i}^{\mathcal{A}}) \in G_i$.

Note that if $\varphi \in \mathcal{L}(\mathcal{G}_1, \ldots, \mathcal{G}_w)$ is a sentence of vocabulary $\tau = \emptyset$. Then

$$\mathrm{Mod}(\varphi) = \{(A, G_1, \ldots, G_w) \mid (A, G_1, \ldots, G_w) \models \varphi\}$$

corresponds to a second-order generalized quantifier of type t. This observation can be used to formalize definability of second-order generalized quantifiers. Below, we assume that $\mathcal{L}$ is closed under substitution.

Definition 2.5. Let Q be a quantifier of type t. The quantifier Q is definable in a logic $\mathcal{L}$ if there is $\varphi \in \mathcal{L}(\mathcal{G}_1, \ldots, \mathcal{G}_w)$ of vocabulary $\sigma = \emptyset$ such that for any t-structure $(A, G_1, \ldots, G_w)$,

$$(A, G_1, \ldots, G_w) \models \varphi \Leftrightarrow (A, G_1, \ldots, G_w) \in Q.$$

The following was shown in Kontinen (2010):

Theorem 2. *If Q is definable in $\mathcal{L}$ then $\mathcal{L} \equiv \mathcal{L}(Q)$.*

The converse of Theorem 2 does not hold:

Theorem 3 (Kontinen (2010))**.** *There is a quantifier Q of type* ((1)) *which is not definable in* FO *and satisfies* FO $\equiv$ FO(Q)*.*

Definability questions of second-order quantifiers has been studied in Kontinen (2010; 2006; 2009). We recall the following results.

Theorem 4 (Kontinen (2006))**.** *Let t be type and $\mathcal{B}_t$ the collection of all second-order quantifiers of types less than t. Then there is a quantifier Q of type t such that Q is not definable in* SO($\mathcal{B}_t$)*.*

Theorem 4 is proved with respect to a natural ordering of the types of second-order generalized quantifiers. Theorem 4 is existential in nature and does not give us a concrete non-definable quantifier. It was observed in Kontinen (2005) that it is not so difficult to find concrete quantifiers which cannot be defined using any monadic quantifiers. Denote by $\mathbb{Q}$ the collection of all monadic second-order generalized quantifiers.

Theorem 5 (Kontinen (2005))**.** *The quantifier* $\exists_2^2$ *is not definable in* FO($\mathbb{Q}$)*.*

It is worth noting that the logic FO($\mathbb{Q}$) is capable of defining all classes of first-order structures (cf. Theorem 6.2 in Andersson (2002)). Finally, we recall the following result about second-order majority quantifiers:

Theorem 6 (Kontinen (2009))**.** *The quantifier* $\exists_k^2$ *is definable in* FO(Most^k)*.*

It interesting to note that definability of Most^1 in the logic SO would imply that PH $\equiv$ CH in computational complexity. This observation was discussed in Kontinen and Szymanik (2008). In this paper we show that the quantifier Most^1 is not definable in SO, but, analogously to Theorem 3, this non-definability result does not imply that PH $\subsetneq$ CH.

3 Characterizing definability

The computational analogue of a first-order generalized quantifier is the notion of an oracle (see Immerman 1999). Let Q be a quantifier of vocabulary τ and $\mathcal{L}$ a logic. The idea is that in $\mathcal{L}(Q)$ we can query "without a cost" if a definable τ-structure $\mathfrak{A}$ is a member of the class Q. Recall that a second-order generalized quantifier $\mathcal{Q}$ of type ((1)) is definable, e.g., in SO if there is a sentence $\varphi \in \mathrm{SO}(\mathcal{G})$ such that for all second-order structures (A, G):

$$(A, G) \models \varphi \Leftrightarrow (A, G) \in \mathcal{Q}. \tag{1}$$

It is not immediately clear how to view this notion in computational terms. The set G corresponds to a local first-order quantifier and, if we treat G as an oracle, then in (1) we are infact trying to define a property oracles. One way to proceed is to formalize definability of a quantifier $\mathcal{Q}$ in terms of oracle Turing machines that treat (a suitable initial segment) the oracle as part of the input. However, in this article we do not follow that idea as there is a more familiar route to

take. An important observation here is that the set G can be of exponential size compared to the domain A. This observation can be used to show that SO-definability of $\mathcal{Q}$ reduces to logarithmic time definability.

Our result is a logical version of the results connecting oracle separation results with lower bound results for small constant depth circuits (Furst et al. 1984, Yao 1985, Håstad 1987, Torán 1988, Vollmer 1998, see, e.g.,). For example, in Torán (1988), Torán studied oracle separations in the counting hierarchy and noticed that there is essentially no difference between an oracle Turing machine writing an oracle query on its query tape and a logarithmic time Turing machine writing an address on its random access tape. He used this analogy to show that an oracle separation result for classes in the polynomial counting hierarchy implies a real separation for the corresponding classes in the logarithmic counting hierarchy LINCH (equivalently in DLOGTIME-uniform TC^0). We use a logical version of this idea: we show that SO and the relation G in (1) can be replaced by FO and a unary relation P by passing from A to a domain of cardinality $2^{|A|}$.

In this section we mainly restrict attention to monadic second-order generalized quantifiers. We interpret definability of second-order quantifiers in MSO(+) in the natural way: for example, a second-order quantifier $\mathcal{Q}$ of type ((1)) is definable in MSO(+) if there is $\varphi \in \mathrm{MSO}(\mathcal{G}, +)$ such that for all structures $(A, +, G)$: $(A, +, G) \models \varphi \Leftrightarrow (A, G) \in \mathcal{Q}$. In particular, Theorem 2 can be proved analogously in this setting.

Next we define a first-order encoding of a second-order structure of type t, for a monadic t. We use the fact that there is a one-to-one correspondence between integers $m \in B = \{0, \ldots, 2^n - 1\}$ and subsets of $A = \{0, \ldots, n-1\}$ seen as length-n binary numbers. Therefore, relations of A can be encoded in terms of tuples of elements of B and, further, sets of relations of A by relations of B.

Definition 3.1. Let $t = (s_1, \ldots, s_w)$ be a type where $s_i = (1, \ldots, 1)$ is of length r_i for $1 \leq i \leq w$. Let $\mathfrak{A} = (A, G_1, \ldots, G_w)$ be a t-structure where $A = \{0, \ldots, n-1\}$ and $G_i \subseteq \mathcal{P}(A) \times \cdots \times \mathcal{P}(A)$. Denote by $\hat{\mathfrak{A}} = (B, P_1, \ldots, P_w)$ the following first-order structure of vocabulary $\tau = \{P_1, \ldots, P_w\}$, where P_i is a r_i-ary predicate, and

1. $B = \{0, \ldots, 2^n - 1\}$,
2. $P_i = \{(j_1, \ldots, j_{r_i}) \in B^{r_i} \mid (J_1, \ldots, J_{r_i}) \in G_i\}$, where, for $1 \leq k \leq r_i$, the length-n binary representation of j_k is given by $s_0 \cdots s_{n-1}$, and $s_l = 1 \Leftrightarrow l \in J_k$.

For a quantifier $\mathcal{Q}$ of type t, we denote by $\mathcal{Q}^\star$ the first-order quantifier of

vocabulary τ defined by

$$Q^\star := \{\hat{\mathfrak{A}} : \mathfrak{A} \in Q\}.$$

It is easy to see that the quantifier $Q^\star$ has only structures in cardinalities of the form 2^n and that $|G_i| = |P_i|$ for $1 \leq i \leq w$. Note also that the quantifier $Q^\star$ encoding Q may depend on the ordering of the domain B and hence does not strictly speaking correspond to a Lindström quantifier of vocabulary τ but a τ-quantifier with build-in arithmetic relations (quantifiers defined and studied in Hella et al. (2010)). On the other hand, for the numerical quantifiers Q discussed in the rest of this section, the first-order encodings $Q^\star$ are obviously order invariant and hence correspond to Lindström quantifiers of vocabulary τ. We are now ready for the main result of this article.

Theorem 7. *Let Q_1 and Q_2 be monadic quantifiers. Then Q_1 is definable in* $\mathrm{MSO}(Q_2, +)$ *if and only if $Q_1^\star$ is definable in* $\mathrm{FO}(Q_2^\star, +, \times)$.

Proof. To simplify notation, we assume that the type of Q_1 and Q_2 is $((1,1))$ and $((1),(1))$, respectively.

Let us first assume that Q_1 is definable in the logic $\mathrm{MSO}(Q_2, +)$. Then there is a sentence $\varphi \in \mathrm{MSO}(Q_2, \mathcal{G}, +)$ such that for all structures $(A, +, G)$

$$(A, +, G) \models \varphi \Leftrightarrow (A, G) \in Q_1 .$$

We shall next show that there is a sentence $\varphi^* \in \mathrm{FO}(Q_2^\star +, \times)[\{P\}]$, where P is binary, such that for all structures $\mathfrak{A} = (A, G)$:

$$(A, +, G) \models \varphi \Leftrightarrow (B, P, <, +, \times) \models \varphi^*, \tag{2}$$

where $(B, P) = \hat{\mathfrak{A}}$ (see Definition 3.1). We define φ^* via the following translation:

$$\begin{aligned}
x_i = x_j &\rightsquigarrow x_i = x_j \\
x_i + x_j = x_k &\rightsquigarrow x_i + x_j = x_k \\
Y_i(x_j) &\rightsquigarrow \mathrm{BIT}(y_i, n - (x_j + 1)) \\
\mathcal{G}x_i, x_j(\psi_1(x_i), \psi_2(x_j)) &\rightsquigarrow \exists z_1 \exists z_2 \Big(P(z_1, z_2) \wedge \bigwedge_{1 \le i \le 2} \forall (w < n)(\psi_i^*(w) \\
&\qquad \leftrightarrow \mathrm{BIT}(z_i, n - (w + 1))) \Big) \\
\psi \wedge \theta &\rightsquigarrow \psi^* \wedge \theta^* \\
\neg\psi &\rightsquigarrow \neg\psi^* \\
\exists x_i \psi &\rightsquigarrow \exists x_i (x_i < n \wedge \psi^*(x_i)) \\
\exists Y_i \psi &\rightsquigarrow \exists y_i \psi^* \\
Q_2\, Y_i, Y_j(\psi(Y_i), \theta(Y_j)) &\rightsquigarrow Q_2^\star\, y_i, y_j(\psi^*(y_i), \theta^*(y_j))
\end{aligned}$$

It is now straigthforward to show that for all formulas $\psi \in \mathrm{MSO}(Q_2, \mathcal{G}, +)$, structures (A, G), and assignments s

$$(A, +, G) \models_s \psi \Leftrightarrow (B, P, <, +, \times) \models_{s^*} \psi^*,$$

where the assignment s^* is defined such that $s^*(x_i) = s(x_i)$ for all first-order variables x_i, and, if $s(Y_i) = D \subseteq \{0, \dots, n-1\}$, then $s^*(y_i)$ is the unique $d < 2^n$ whose binary representation is given by $s_0 \cdots s_{n-1}$ where $s_j = 1 \iff j \in D$.

In the formula translation, we use the predicate BIT, which is FO(+, ×)-definable, to recover the set D from the integer d. By the above translation, the sentence

$$\exists n(|B| = 2^n \wedge \varphi^*)$$

of the logic $\mathrm{FO}(Q_2^\star, +, \times)$ now defines the quantifier $Q_1^\star$.

Let us then show the converse implication. Assume that $\varphi \in \mathrm{FO}(Q_2^\star, +, \times)$ defines the quantifier $Q_1^\star$. The idea is now to translate $\varphi \in \mathrm{FO}(Q_2^\star, +, \times)$ to $\varphi' \in \mathrm{MSO}(Q_2, \mathcal{G}, +)$ such that for all $\mathfrak{A} = (A, G)$:

$$(A, +, G) \models \varphi' \Leftrightarrow (B, P, <, +, \times) \models \varphi. \tag{3}$$

Analogously to the first translation, we encode integers in the domain $B = \{0, \dots, 2^n - 1\}$ in terms of subsets $X \subseteq \{0, \dots, n-1\}$. We use the following formulas $X = Y$, $X < Y$, $X + Y = Z$, and $X \times Y = Z$ expressing arithmetic

operations on binary numbers. The first three formulas are FO(+)-expressible, and the fourth is expressible in the logic FO(M, +, ×) ≤ MSO(+) (Hesse et al. 2002). The translation $\varphi \rightsquigarrow \varphi'$ is now defined as follows.

$$\begin{aligned}
P(x_i, x_j) &\rightsquigarrow \mathcal{G}z_1, z_2(X_i(z_1), X_j(z_2)) \\
x_i = x_j &\rightsquigarrow X_i = X_j \\
x_i < x_j &\rightsquigarrow X_i < X_j \\
x_i + x_j = x_k &\rightsquigarrow X_i + X_j = X_k \\
x_i \times x_j = x_k &\rightsquigarrow X_i \times X_j = X_k \\
\psi \wedge \varphi &\rightsquigarrow \psi' \wedge \varphi' \\
\neg\psi &\rightsquigarrow \neg\psi' \\
\exists x_i \psi(x_i) &\rightsquigarrow \exists X_i \psi'(X_i) \\
Q_2^\star x_i, x_j(\psi(x_i), \theta(x_j)) &\rightsquigarrow \mathcal{Q}_2\, X_i, X_j(\psi'(X_i), \theta'(X_j))
\end{aligned}$$

It is straightforward to show that this translation works as intended. In particular, it follows that the sentence $\varphi' \in \mathrm{MSO}(\mathcal{Q}_2, \mathcal{G}, +)$ now defines the quantifier $\mathcal{Q}_1$. □

Let us then discuss some corollaries of Theorem 7. We need the following definition.

Definition 3.2. Let $t = (s_1, \ldots, s_w)$ and τ be as in Definition 3.1. Let $\mathcal{Q}$ be a quantifier of type t. The quantifier $\mathcal{Q}$ is *numerical* if there is a relation $T \subseteq \mathbb{N}^w$ such that for all t-structures $(A, P_1, \ldots, P_w)$

$$(A, P_1, \ldots, P_w) \in \mathcal{Q} \Leftrightarrow (|P_1|, \ldots, |P_w|) \in T.$$

We denote $\mathcal{Q}$ by $\mathcal{Q}_T$ and by Q_T the first-order numerical quantifier (defined analogously) of vocabulary τ.

It is easy to see that, for a numerical $\mathcal{Q}_T$, the quantifier $\mathcal{Q}_T^\star$ (see Definition 3.1) is just the restriction of the corresponding first-order quantifier Q_T to the cardinalities 2^n:

$$\mathcal{Q}_T^\star = \{(A, P_1, \ldots, P_w) \in Q_T : |A| = 2^n \text{ for some } n \in \mathbb{N}\}.$$

This observation allows us to show the following:

Theorem 8. *Let $\mathcal{Q}_T$ be a numerical quantifier and $k \in \mathbb{N}$. Then*

1. $\mathcal{Q}_T$ *is definable in* MSO(+) *iff* Q_T *is definable in* FO(+, ×).
2. $\mathcal{Q}_T$ *is definable in* MSO($\mathcal{D}_k$, +) *iff* Q_T *is definable in* FO(D_k, +, ×).
3. $\mathcal{Q}_T$ *is definable in* MSO(Most^1, +) *iff* Q_T *is definable in* FO(M, +, ×).

Proof. The proof is based on the fact that each of the logics FO(<, +, ×), FO(D_k, +, ×), and FO(M, +, ×) is closed under logical reductions. Suppose that $\mathcal{Q}_T$ is of type $t = (s_1, \ldots, s_w)$ and let τ denote the vocabulary of the corresponding first-order quantifier Q_T (see Definition 3.1).

Let us consider claim 2. By Theorem 7 it suffices to show that the following are equivalent:

(*a*) $\mathcal{Q}_T^\star$ is definable in FO($\mathcal{D}_k^\star$, +, ×)
(*b*) Q_T is definable in FO(D_k, +, ×)

Recall that the quantifiers $\mathcal{Q}_T^\star$ and $\mathcal{D}_k^\star$ are the restrictions of the quantifiers Q_T and D_k to cardinalities of the form 2^n, respectively. Let us first note that (*a*) is equivalent with

(*c*) $\mathcal{Q}_T^\star$ is definable in FO(D_k, +, ×).

First of all, since $\mathcal{D}_k^\star$ is easily definable in FO(D_k, +, ×) using the FO(+, ×)-expressible predicate $x = 2^y$, it follows that (*a*) ⇒ (*c*). Assume then that (c) holds and let $\varphi \in \mathrm{FO}(\mathrm{D}_k, +, \times)$ define $\mathcal{Q}_T^\star$. Define a sentence ψ as follows:

$$\psi := \exists n(|A| = 2^n \wedge \varphi(\mathrm{D}_k / \mathcal{D}_k^\star)).$$

Since the quantifier $\mathcal{Q}_T^\star$ contains structures only in cardinalities of the form 2^n it is easy to see that $\psi \in \mathrm{FO}(\mathcal{D}_k^\star, +, \times)$ also defines $\mathcal{Q}_T^\star$.

It now suffices to show that (*b*) and (*c*) are equivalent. Note first that (*b*) ⇒ (*c*) can be easily proved using the predicate $x = 2^y$. We will show (*c*) ⇒ (*b*). Here we use the fact that the logic FO(D_k, +, ×) is closed under logical reductions. We will define Q_T (over all cardinalities) with the help of the quantifier $\mathcal{Q}_T^\star$. Let $\mathfrak{A}$ be a structure. If $|A| = 2^n$ for some $n \in \mathbb{N}$, then $\mathfrak{A} \in Q_T$ can be expressed in terms of the quantifier $\mathcal{Q}_T^\star$. Note that even if $|A|$ is not a power of two, it holds that the least m such that $|A| \leq 2^m$ satisfies $2^m \leq 2|A|$.

We will now sketch how the quantifier Q_T can be defined in terms of $\mathcal{Q}_T^\star$. Assume $\varphi \in \mathrm{FO}(\mathrm{D}_k, +, \times)$ is a sentence defining $\mathcal{Q}_T^\star$. Let $\mathfrak{A} = (A, P_1, \ldots, P_w)$ be a τ-structure, where $A = \{0, \ldots, n-1\}$. We use the following facts:

1. There is a FO($<, +, \times$)-definable query I that maps $\mathfrak{A}$ to the structure $I(\mathfrak{A})$ which is isomorphic to

$$(\{0, \ldots, 2^m - 1\}, P_1, \ldots, P_w, <, +, \times),$$

   where 2^m is the least power of two satisfying $n \leq 2^m$.
2. There is a sentence $\psi \in \mathrm{FO}(\mathrm{D}_k, +, \times)$ such that for all $\mathfrak{A}$:

$$\mathfrak{A} \models \psi \Leftrightarrow I(\mathfrak{A}) \models \varphi.$$

Since Q_T is numerical, the sentence ψ now defines Q_T. The query I is easily definable in FO($<, +, \times$); the domain of $I(\mathfrak{A})$ is defined as $\{(i, j) \in A^2 \mid in + j < 2^m\}$ (see Immerman 1999, for more on first-order queries). The sentence ψ is constructed inductively (see e.g., Immerman 1999, Section 3.2) using, in particular, the fact that the second vectorization D_k^2 of D_k can be expressed in $\mathrm{FO}(\mathrm{D}_k, +, \times)$.

The claims 1 are 3 are proved analogously. For claim 3 we use the facts that $(\mathrm{Most}^1)^\star$ is the restriction of M to the cardinalities 2^n and that the second vectorization M^2 of M is definable in FO(M, $+, \times$) (see Barrington et al. 1990).

□

The following lemma can be now used.

Lemma 1. *Let $S \subseteq \mathbb{N}$, p a prime, and $q > 1$ relatively prime to p. Then*

1. *Q_S is definable in* FO($+, \times$) *iff S either finite or cofinite.*
2. *D_q is not definable in* $\mathrm{FO}(\mathrm{D}_p, +, \times)$.

Proof. The first claim follows from non-definability of the language PARITY in FO($+, \times$) (Furst et al. 1984, Ajtai 1983) (see Theorem 4.3 in Barrington et al. (2005)). The second claim goes back to Smolensky (1987). □

By combining Theorem 8 and Lemma 1 we can show the following.

Corollary 1. *Let $S \subseteq \mathbb{N}$, p a prime, and $q > 1$ relatively prime to p. Then*

1. *$\boldsymbol{Q}_S$ is definable in* MSO(+) *iff S is either finite or cofinite.*
2. *$\mathcal{D}_q$ is not definable in* $\mathrm{MSO}(\mathcal{D}_p, +)$.

Another corollary of Theorem 8 is that the quantifier Most^1 is not definable in the logic MSO(+).

Corollary 2. *The quantifier* Most^1 *is not definable in* MSO(+).

Proof. For a contradiction, let us assume that Most^1 is definable in MSO(+). By the results of Kontinen (2009), the quantifier I^2 can then also be defined in MSO(+). Now, since I^2 is numerical, Theorem 8 implies that the quantifier I is definable in FO(+, ×). This is a contradiction since

$$\text{FO}(\text{I}, +, \times) \equiv \text{FO}(\text{M}, +, \times) > \text{FO}(+, \times).$$

□

It is possible to replace MSO(+) by SO in Theorem 7. The idea is that, if Q_1 is definable in $\text{SO}(Q_2)$, then in the defining formula, for some k, only relations of arity at most k are quantified. We will not pursued this generalization in full generality but only consider the special case of the quantifier Most^1.

Theorem 9. *The quantifier* Most^1 *is not definable in* SO.

Proof. It suffices to show that Most^1 is not definable in $\text{FO}(\exists^2_k)$ for any k. For a contradiction, assume that Most^1 is definable in $\text{FO}(\exists^2_k)$. We will now proceed as follows: First an analogous translation as in Theorem 7 is used to show that definability of Most^1 in $\text{FO}(\exists^2_k)$ implies that a certain padded version of the class M is definable in FO(+, ×) over cardinalities 2^{n^k}. This class corresponds to a variant L of the binary language MAJ

$$\text{MAJ} = \{w \in \{0,1\}^+ \mid |w|_1 > |w|_0\},$$

when ordered $\{P\}$-structures are viewed as binary strings. Definability of L in FO(+, ×) would allow us to construct constant depth quasipolynomial size $(2^{\log(n)^{O(1)}})$ AND/OR circuits for MAJ contradicting the result of Yao (1985) and Håstad (1987).

We will now discuss the proof in more detail. Note first that, in order to translate the quantifier $\exists^2_k$ to the logic FO(+, ×), we need to redefine the structure $\hat{\mathfrak{A}}$ (see Definition 3.1) to have a domain of the form $\{0, \ldots, 2^{n^k} - 1\}$ instead of $\{0, \ldots, 2^n - 1\}$. In other respects the definition of $\hat{\mathfrak{A}}$ is not altered. We can

now use the fact that there is a one-to-one correspondence between integers $m \in \{0, \ldots, 2^{n^k} - 1\}$ and k-ary relations R of $\{0, \ldots, n-1\}$. In other words, by using the lexicographic ordering on k-tuples, a relation R can be encoded by a binary string of length n^k corresponding to the binary representation of a unique integer $m < 2^{n^k}$. It is straightforward to adjust the translation in the proof of Theorem 7 to this setting. The only difference is that the unary relations Y_i are now k-ary. The translation is modified as follows to translate the k-ary atomic formulas $Y_i(x_1, \ldots, x_k)$:

$$Y_i(x_1, \ldots, x_k) \rightsquigarrow \exists z(\mathrm{BIT}(y_i, n^k - (z+1)) \wedge z = n^{k-1}x_1 + \cdots + nx_{k-1} + x_k).$$

We assumed that the quantifier Most^1 is definable in $\mathrm{FO}(\exists^2_k)$ which now implies that the following class $(\mathrm{Most}^1)^\star$

$$(\mathrm{Most}^1)^\star = \{(B, P, <, +, \times) \mid (B, P) = \hat{\mathfrak{A}} \text{ and } \mathfrak{A} \in \mathrm{Most}^1\}$$

can be defined in the logic $\mathrm{FO}(+, \times)$. Note that $(B, P, <, +, \times) \in (\mathrm{Most}^1)^\star$ iff $B = \{0, \ldots, 2^{n^k} - 1\}$, $P \subseteq \{0, \ldots, 2^n - 1\}$, and $|P| > 2^{n-1}$. By viewing the structures of $(\mathrm{Most}^1)^\star$ as binary words, it follows that the binary language L

$$L = \{w_1 w_2 \in \{0,1\}^* : \ |w_1| = 2^n,\ w_1 \in \mathrm{MAJ},\ |w_2| = 2^{n^k} - 2^n,\ w_2 \in 0^*\},$$

can be defined in the logic $\mathrm{FO}(+, \times)$.

Since $\mathrm{FO}(+, \times)$ corresponds to DLOGTIME-uniform AC^0, we get that there is a uniform family $(C_n)_{n \in \mathbb{N}}$ of constant depth polynomial size AND/OR circuits accepting L. These circuits can be now used to construct a family $(C'_{2^n})_{n \in \mathbb{N}}$ of constant depth quasipolynomial size AND/OR circuits for MAJ in input lengths 2^n: the circuit C'_{2^n} for length 2^n binary words is acquired from the circuit $C_{2^{n^k}}$ by turning the input gates with index i, for $2^n < i \leq 2^{n^k}$, to constant 0 gates. It is easy to see that the size of C'_m is $2^{O(\log(m)^k)}$. This is a contradiction with the results of Yao (1985) and Håstad (1987) showing that such a quasipolynomial size family $(C'_{2^n})_{n \in \mathbb{N}}$ cannot exist. □

Conclusion

We have shown that definability of second-order generalized quantifiers can be reduced to definability of first-order generalized quantifiers. We have indicated

a couple of corollaries to our characterization but surely there is more to be done here, e.g., with replacing the base logic MSO(+) by SO as in Theorem 9. In particular, Theorem 9 solves the open problem proposed in Kontinen and Szymanik (2008), where we studied the collective meanings of natural language quantifiers. It suggests, as we argued in Kontinen and Szymanik (2008), that the type-shifting strategy (see Winter 2001) to define the meanings of natural language quantification might be too restricted in its computational power. It is likely that second-order logic is not enough to capture natural language semantics. Another interpretation would be that everyday language does not realize hard collective quantifiers (for sure they are marginal at best) due to their complexity.

Acknowledgements The first author was supported by grant 127661 of the Academy of Finland. The second author was supported by Vici grant NWO-277-80-001. Both authors would like to thank Heribert Vollmer for valuable comments.

References

M. Ajtai. Σ^1_1-formulae on finite structures. *Ann. Pure Appl. Logic*, 24(1):1–48, 1983. ISSN 0168-0072.

A. Andersson. On second-order generalized quantifiers and finite structures. *Ann. Pure Appl. Logic*, 115(1-3):1–32, 2002.

D. A. M. Barrington, N. Immerman, and H. Straubing. On uniformity within NC^1. *J. Comput. System Sci.*, 41(3):274–306, 1990.

D. A. M. Barrington, N. Immerman, C. Lautemann, N. Schweikardt, and D. Thérien. First-order expressibility of languages with neutral letters or: The Crane Beach conjecture. *J. Comput. System Sci.*, 70(2):101–127, 2005.

D. P. Bovet, P. Crescenzi, and R. Silvestri. A uniform approach to define complexity classes. *Theor. Comput. Sci.*, 104(2):263–283, 1992.

J. R. Büchi. On a decision method in restricted second-order arithmetic. In *Proceedings Logic, Methodology and Philosophy of Sciences 1960*, Stanford, CA, 1962. Stanford University Press.

J. R. Büchi and C. C. Elgot. Decision problems of weak second order arithmetics and finite automata, Part I. *Notices of the American Mathematical Society*, 5:834, 1958.

H.-J. Burtschick and H. Vollmer. Lindström quantifiers and leaf language definability. *Int. J. Found. Comput. Sci.*, 9(3):277–294, 1998.

J. van der Does. Sums and quantifiers. *Linguistics and Philosophy*, 16(5):509–550, 1993.

H.-D. Ebbinghaus and J. Flum. *Finite model theory, 2nd edition*. Perspectives in Mathematical Logic. Springer-Verlag, 1999.

R. Fagin. Generalized first-order spectra and polynomial-time recognizable sets. In *Complexity of computation (Proc. SIAM-AMS Sympos. Appl. Math., New York, 1973)*, pages 43–73. SIAM–AMS Proc., Vol. VII. Amer. Math. Soc., Providence, R.I., 1974.

M. L. Furst, J. B. Saxe, and M. Sipser. Parity, circuits, and the polynomial-time hierarchy. *Math. Systems Theory*, 17(1):13–27, 1984.

M. Galota and H. Vollmer. A generalization of the Büchi-Elgot-Trakhtenbrot-theorem. In *Computer Science Logic*, Lecture Notes in Computer Science, pages 355–368, Berlin Heidelberg, 2001. Springer Verlag.

J. T. Håstad. *Computational limitations of small-depth circuits*. MIT Press, Cambridge, MA, 1987.

L. Hella, J. Kontinen, and K. Luosto. Regular representations of uniform TC^0. Institut Mittag-Leffler Preprint Series (The Program of Fall 2009), 2010.

W. Hesse, E. Allender, and D. A. M. Barrington. Uniform constant-depth threshold circuits for division and iterated multiplication. *J. Comput. System Sci.*, 65(4):695–716, 2002. Special issue on complexity, 2001 (Chicago, IL).

N. Immerman. *Descriptive complexity*. Graduate Texts in Computer Science. Springer-Verlag, New York, 1999.

J. Kontinen. *Definability of second order generalized quantifiers*. PhD thesis, University of Helsinki, 2005.

J. Kontinen. The hierarchy theorem for second order generalized quantifiers. *J. Symbolic Logic*, 71(1):188–202, 2006.

J. Kontinen. A logical characterization of the counting hierarchy. *ACM Trans. Comput. Log.*, 10(1), 2009.

J. Kontinen. Definability of second order generalized quantifiers. *Arch. Math. Logic*, 49(3):379–398, 2010.

J. Kontinen and H. Niemistö. Extensions of MSO and the monadic counting hierarchy. *Information and Computation*, 209(1):1–19, 2011.

J. Kontinen and J. Szymanik. A remark on collective quantification. *Journal of Logic, Language and Information*, 17(2):131–140, 2008.

J. Kontinen and J. Szymanik. Characterizing definability of second-order generalized quantifiers. In L. D. Beklemishev and R. de Queiroz, editors, *WoLLIC*, volume 6642 of *Lecture Notes in Computer Science*, pages 187–200. Springer, 2011.

P. Lindström. First order predicate logic with generalized quantifiers. *Theoria*, 32:186–195, 1966.

J. T. Lønning. Plurals and collectives. In J. van Benthem and A. ter Meulen, editors, *Handbook of Logic and Language*, pages 989–1035. Elsevier, second edition, 2011.

M. More and F. Olive. Rudimentary languages and second-order logic. *Math. Logic Quart.*, 43(3):419–426, 1997.

A. Mostowski. On a generalization of quantifiers. *Fund. Math.*, 44:12–36, 1957. ISSN 0016-2736.

M. Mostowski and J. Szymanik. Semantic bounds for everyday language. *Semiotica*, 188(1-4):363–372, 2012.

I. Parberry and G. Schnitger. Parallel computation with threshold functions. *J. Comput. System Sci.*, 36(3):278–302, 1988.

T. Peichl and H. Vollmer. Finite automata with generalized acceptance criteria. *Discrete Mathematics and Theoretical Computer Science*, 4:179–192, 2001.

S. Peters and D. Westerståhl. *Quantifiers in Language and Logic*. Clarendon Press, Oxford, 2006.

R. Smolensky. Algebraic methods in the theory of lower bounds for boolean circuit complexity. In *STOC*, pages 77–82, 1987.

L. J. Stockmeyer. The polynomial-time hierarchy. *Theor. Comput. Sci.*, 3(1):1–22 (1977), 1976.

J. Szymanik. *Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language*. PhD thesis, Universiteit van Amsterdam, 2009.

J. Szymanik. Computational complexity of polyadic lifts of generalized quantifiers in natural language. *Linguistics and Philosophy*, 33:215–250, 2010.

J. Szymanik and M. Zajenkowski. Comprehension of simple quantifiers. Empirical evaluation of a computational model. *Cognitive Science: A Multidisciplinary Journal*, 34(3):521–532, 2010.

J. Torán. *Structural properties of the counting hierarchies*. PhD thesis, Facultat d'Informatica de Barcelona, Barcelona, Spain, 1988.

B. A. Trakhtenbrot. Finite automata and logic of monadic predicates. *Doklady Akademii Nauk SSSR*, 140:326–329, 1961. In Russian.

J. Väänänen. Generalized quantifiers, an introduction. In *Generalized quantifiers and computation (Aix-en-Provence, 1997)*, volume 1754 of *Lecture Notes in Comput. Sci.*, pages 1–17. Springer, Berlin, 1999.

H. Vollmer. Relating polynomial time to constant depth. *Theor. Comput. Sci.*, 207(1):159–170, 1998.

Y. Winter. *Flexibility principles in Boolean semantics*. The MIT Press, London, 2001.

A. C.-C. Yao. Separating the polynomial-time hierarchy by oracles. *Foundations of Computer Science, Annual IEEE Symposium on*, 0:1–10, 1985.

Incorporating Action Models into the Situation Calculus

Yongmei Liu and Hector J. Levesque

Sun Yat-sen University, China; University of Toronto, Canada
ymliu@mail.sysu.edu.cn, hector@cs.toronto.edu

Abstract

While both situation calculus and dynamic epistemic logics (DELs) are concerned with reasoning about actions and their effects, historically, the emphasis of situation calculus was on physical actions in the single-agent case, in contrast, DELs focused on epistemic actions in the multi-agent case. In recent years, the cross-fertilization between the two areas has begun to attract attention. In this paper, we incorporate the idea of action models from DELs into the situation calculus to develop a general multi-agent extension of it. We analyze properties of beliefs in this extension, and give examples to illustrate the modeling of multi-agent scenarios in the situation calculus.

1 Introduction

While both situation calculus (Reiter 2001) and dynamic epistemic logics (DELs) (van Ditmarsch et al. 2007) are concerned with reasoning about actions and their effects, historically, the emphasis of situation calculus was on physical actions in the single-agent case, in contrast, DELs focused on epistemic actions in the

multi-agent case. In recent years, cross-fertilization between the two areas has begun to attract attention. In particular, van Benthem (2011) proposed the idea that situation calculus and modal logic meet and merge. van Ditmarsch et al. (2011) embedded a propositional fragment of situation calculus into a DEL. Kelly and Pearce (2008) incorporated ideas from DELs to handle regression for common knowledge in the situation calculus. Baral (2010) proposed to combine results from reasoning about actions and DELs.

In a multi-agent setting, the agents in the domain may have different perspectives of the actions. Baltag et al. (1998) introduced a construct called an action model to represent these differences of perspectives. An action model consists of a set of actions, a precondition for each action, and a binary relation on the set of actions for each agent, which represents the agent's ability to distinguish between the actions. Moreover, they defined an operation by which an action model may be used to update a Kripke world to obtain a successor world modeling the effects of the action execution. They proposed a logic, called action model logic, to reason about action models and their effects on agents' epistemic state. van Benthem et al. (2006) generalized the concept of action model to that of update model where each action is also associated with a postcondition. So action models can model events which bring about epistemic change, but update models can model events which can not only change agents' epistemic state but also the world state.

The situation calculus was first introduced by (McCarthy and Hayes 1969) and historically, one of its major concerns was how to solve the frame problem, that is, how to represent the effects of a world-changing action without explicitly specifying which conditions are not affected by the action. Reiter (1991) gave a solution to the frame problem under some conditions in the form of successor state axioms. This solution to the frame problem has proven useful as the foundation for the high-level robot programming language Golog (Levesque et al. 1997). Scherl and Levesque (1993; 2003) extended Reiter's solution to cover epistemic actions in the single-agent case. Later, Shapiro et al. (1998) extended their work to the multi-agent case, but they only considered public actions whose occurrence is common knowledge. In the last decade, Lakemeyer and Levesque (2004; 2005) proposed a logic called $\mathcal{ES}$, which is a fragment of the situation calculus with knowledge. Recently, Belle and Lakemeyer (2010) gave a multi-agent extension of $\mathcal{ES}$, but as Shapiro et al. (1998), they only considered public actions. So up to now, although there have been extensions of the situation calculus into the multi-agent case, they are not able to account for arbitrary multi-agent scenarios.

In this paper, we incorporate action models into the situation calculus to develop a general multi-agent extension of it. We analyze properties of beliefs in this extension, and give examples to illustrate the modeling of multi-agent scenarios in the situation calculus.

2 Preliminaries

In this section, we introduce the situation calculus, and action model logic.

2.1 Situation calculus and Golog

The situation calculus (Reiter 2001) is a many-sorted first-order language suitable for describing dynamic worlds. There are three disjoint sorts: *action* for actions, *situation* for situations, and *object* for everything else. A situation calculus language $\mathcal{L}_{sc}$ has the following components: a constant S_0 denoting the initial situation; a binary function $do(a, s)$ denoting the successor situation to s resulting from performing action a; a binary predicate $s \sqsubset s'$ meaning that situation s is a proper subhistory of situation s'; a binary predicate $Poss(a, s)$ meaning that action a is possible in situation s; a binary predicate $Poss(a, s)$ meaning that action a is possible in situation s; action functions; a finite number of relational and functional fluents, *i.e.*, predicates and functions taking a situation term as their last argument; and a finite number of situation-independent predicates and functions.

The situation calculus has been extended to accommodate sensing and knowledge. Assume that in addition to ordinary actions that change the world, there are sensing actions which do not change the world but tell the agent information about the world. A special binary function $SR(a, s)$ is used to characterize what the sensing action tells the agent about the world. Knowledge is modeled in the possible-world style by introducing a special fluent $K(s', s)$, meaning that situation s' is accessible from situation s. Note that the order of the arguments is reversed from the usual convention in modal logic. Then knowing φ at situation s is represented as follows:

$$\mathbf{Knows}(\varphi(now), s) \stackrel{def}{=} \forall s'.K(s', s) \supset \varphi(s'),$$

where *now* is used as a placeholder for a situation argument. For example,

Knows$(\exists s^*.now = do(open, s^*), s)$ means knowing that the *open* action has just been executed. When "*now*" only appears as situation arguments to fluents, it is often omitted.

Scherl and Levesque (1993) proposed the following successor state axiom for the K fluent:

$$K(s', do(a, s)) \equiv \exists s^*.K(s^*, s) \wedge s' = do(a, s^*) \wedge SR(a, s^*) = SR(a, s).$$

Intuitively, situation s' is accessible after action a is done in situation s iff it is the result of doing a in some s^* which is accessible from s and agrees with s on the sensing result.

Based on the situation calculus, a logic programming language Golog (Levesque et al. 1997) has been designed for high-level robotic control. It draws considerably from dynamic logic, and has the following programming constructs: primitive actions: α, test actions: $\varphi?$, sequence: $(\delta_1; \delta_2)$, nondeterministic choice of actions: $(\delta_1 \mid \delta_2)$, nondeterministic choice of action arguments: $(\pi x)\delta(x)$, nondeterministic iteration: δ^*, and procedures. The formal semantics of Golog is specified by an abbreviation $Do(\delta, s, s')$, which intuitively means executing δ brings us from situation s to s'. It is inductively defined on δ, and we give some example definitions in the following:

- $Do(\alpha, s, s') \stackrel{def}{=} Poss(\alpha, s) \wedge s' = do(\alpha, s)$;
- $Do(\varphi?, s, s') \stackrel{def}{=} \varphi[s] \wedge s = s'$, where φ is a situation-suppressed formula, and $\varphi[s]$ denotes the formula obtained from φ by taking s as the situation arguments of all fluents.
- $Do(\delta_1; \delta_2, s, s') \stackrel{def}{=} (\exists s'').Do(\delta_1, s, s'') \wedge Do(\delta_2, s'', s')$;
- $Do((\pi x)\delta(x), s, s') \stackrel{def}{=} (\exists x)Do(\delta(x), s, s')$.

2.2 Action model logic (AML)

We fix a finite set of agents $\mathcal{A}$ and a countable set of propositional atoms $\mathcal{P}$. We first define Kripke models.

Definition 2.1. A Kripke model M is a triple (S, R, V) where

- S is a set of states;

- For each agent i, R_i is a binary relation on S;
- For each $s \in S$, $V(s)$ is a subset of the atoms.

A pointed Kripke model is a pair (M, s_0) where M is a Kripke model and s_0 is a state of M.

Definition 2.2. An action model over a language $\mathcal{L}$ is a triple $(A, \rightarrow, pre)$ where

- A is a set of action points;
- For each agent i, $\rightarrow_i$ is a binary relation on A;
- For each action point a, $pre(a) \in \mathcal{L}$ is its precondition.

A pointed action model is a pair (N, a_0) where N is an action model and a_0 is an action point of N.

Definition 2.3. Let $M = (S, R, V)$ be a Kripke model, and $s_0 \in S$. Let $N = (A, \rightarrow, pre)$ be an action model, and $a_0 \in A$. The product of (M, s_0) and (N, a_0), denoted by $(M, s_0) \otimes (N, a_0)$, is a pointed Kripke model (M', s_0') where $M' = (S', R', V')$, and

- $S' = \{(s, a) \mid s \in S, a \in A, \text{ and } M, s \models pre(a)\}$
- $s_0' = (s_0, a_0)$
- $(s, a) R_i (s', a')$ iff $s R_i s'$ and $a \rightarrow_i a'$
- For each $(s, a) \in S'$, $V'((s, a)) = V(s)$.

Definition 2.4. The language $\mathcal{L}_{am}$ of action model logic is defined by

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \psi) \mid B_i\varphi \mid C_{\mathcal{E}}\varphi \mid [N, a_0]\varphi$$

where $p \in \mathcal{P}$, $i \in \mathcal{A}$, $\mathcal{E} \subseteq \mathcal{A}$, $\varphi, \psi \in \mathcal{L}_{am}$, and (N, a_0) is a pointed action model with a finite domain and such that for all action point a, $pre(a) \in \mathcal{L}_{am}$.

Definition 2.5. Let $M = (S, R, V)$ be a Kripke model and s_0 a state of M. The interpretation of formulas is as follows:

1. $M, s_0 \models p$ iff $p \in V(s_0)$;
2. $M, s_0 \models \neg\varphi$ iff $M, s_0 \not\models \varphi$;
3. $M, s_0 \models \varphi \wedge \psi$ iff $M, s_0 \models \varphi$ and $M, s_0 \models \psi$;

4. $M, s_0 \models B_i\varphi$ iff for all s such that $s_0 R_i s$, $M, s \models \varphi$;
5. $M, s_0 \models C_{\mathcal{E}}\varphi$ iff for all s s.t. $s_0 R_{\mathcal{E}} s$, $M, s \models \varphi$, where $R_{\mathcal{E}}$ is the reflexive transitive closure of the union of R_i for $i \in \mathcal{E}$;
6. $M, s_0 \models [N, a_0]\varphi$ iff if $M, s_0 \models pre(a_0)$, then $(M, s_0) \otimes (N, a_0) \models \varphi$.

A formula φ is valid if it is true in any pointed Kripke model.

Example 1. (van Ditmarsch et al. 2007) Two stockbrokers Ann and Bob are having a little break in a Wall Street bar, sitting at a table. A messenger comes in and delivers a letter to Ann. On the envelope is written "urgently requested data on United Agents". Let atom p mean that "United Agents is doing well". Consider the following scenarios:

1. Bob sees that Ann reads the letter. From Bob's point of view, Ann could learn p or she could learn $\neg p$, and he cannot distinguish between these two actions. But Ann can certainly distinguish between the two actions. Thus we get the following action model: $read = (A, \rightarrow, pre)$, where $A = \{0, 1\}$, $pre(0) = \neg p$, $pre(1) = p$, $\rightarrow_a$ is the identity relation, and $\rightarrow_b$ is the total relation.
2. Bob leaves the table; Ann may have read the letter while Bob is away. From Bob's point of view, there are 3 possibilities: Ann learns p, Ann learns $\neg p$, and Ann learns nothing, and he cannot distinguish between these actions. Thus the action model is: $mayread = (A, \rightarrow, pre)$, where $A = \{0, 1, t\}$, $pre(0) = \neg p$, $pre(1) = p$, $pre(t) = true$, $\rightarrow_a$ is the identity relation, and $\rightarrow_b$ is the total relation.

3 A multi-agent extension of the situation calculus

We assume that in addition to ordinary actions which change the world, there are observation actions of the form $observe_{\varphi(now)}$, which does not change the world but observes a condition $\varphi(s)$ holds in the current situation, where $\varphi(s)$ is a formula with a single free situation variable s. For simplicity of notation, we write $\varphi(now)$ instead of $observe_{\varphi(now)}$, and when "now" only appears as situation arguments to fluents, we often omit it. We have $Poss(\varphi(now), s) \equiv \varphi(s)$. There is a special observation action $true$, denoted by nil.

Instead of Scherl and Levesque's K fluent, we now use a fluent $B(i, s', s)$, which means that agent i considers situation s' accessible from situation s. We introduce a special fluent $A(i, a', a, s)$, meaning that in situation s, agent i considers action a' as a possible alternative of action a.

There is a special type of actions, which does not change the world, but changes the A fluent. We call these actions context actions. For each context action c, we have $Poss(c, s) \equiv \forall i \forall a.A(i, a, c, s) \equiv a = c$. That is, c is possible in s iff in s, each agent i considers c as the only alternative to itself. We introduce a special situation-independent predicate $D(c, a)$, which means that a is a possible action for context action c. For each context action c, we use $[\![c]\!]$ to abbreviate for the program $c; (\pi a).D(c, a)?; a$.

We propose the following successor state axiom for the B fluent:

$$B(i, s', do(a, s)) \equiv \exists s^* \exists a^*.B(i, s^*, s) \wedge A(i, a^*, a, s) \wedge \\ [Poss(a, s) \supset Poss(a^*, s^*)] \wedge s' = do(a^*, s^*).$$

Intuitively, for agent i, situation s' is accessible after action a is performed in situation s iff it is the result of doing some alternative a^* of a in some s^* accessible from s, and executability of a in s implies that of a^* in s^*.

In the multi-agent case, a domain of application is specified by a basic action theory of the form:

$$\mathcal{D} = \Sigma \cup \mathcal{D}_{ss} \cup \mathcal{D}_{ap} \cup \mathcal{D}_{una} \cup \mathcal{D}_{S_0}, \text{ where}$$

1. Σ are the foundational axioms:
   (F1) $do(a_1, s_1) = do(a_2, s_2) \supset a_1 = a_2 \wedge s_1 = s_2$
   (F2) $(\neg s \sqsubset S_0) \wedge (s \sqsubset do(a, s') \equiv s \sqsubseteq s')$
   (F3) $\forall P.\forall s[Init(s) \supset P(s)] \wedge \forall a, s[P(s) \supset P(do(a, s))] \supset (\forall s)P(s)$, where $Init(s) \stackrel{def}{=} \neg(\exists a, s')s = do(a, s')$.
   (F4) $B(i, s, s') \supset [Init(s) \equiv Init(s')]$.

   Intuitively, $Init(s)$ means s is the initial situation. A model of the above axioms consists of a forest of isomorphic trees rooted at the initial situations, which can be B-related to only initial situations.
2. $\mathcal{D}_{ss}$ is a set of successor state axioms (SSAs) for fluents. The SSAs for ordinary fluents must satisfy the no-side-effect conditions, *i.e.*, they are not affected by observation or context actions. In this paper, we will

present an SSA for the A fluent by presenting an axiom of the form $A(i, a', a, do(C(\vec{x}), s)) \equiv \Theta$ for each action function $C(\vec{x})$ which may change the A fluent. Usually, such a $C(\vec{x})$ is a context action, and we will also present an axiom of the form $D(C(\vec{x}), a) \equiv \Delta$, which specifies the possible actions for the context action.

3. $\mathcal{D}_{ap}$ is a set of action precondition axioms.
4. $\mathcal{D}_{una}$ is the set of unique names axioms for actions.
5. $\mathcal{D}_{S_0}$ is a set of sentences about S_0.

We now axiomatize the letter example:

Example 2.

1. Bob sees that Ann reads the letter. The axioms are:
   $D(read, a) \equiv a = p \vee a = \neg p$,
   $A(i, a', a, do(read, s)) \equiv D(read, a) \wedge D(read, a') \wedge (i = ann \supset a = a')$.
   So *read* is a context action with two possible actions: p and $\neg p$. After doing *read*, Agent i considers a' as an alternative of a iff both a and a' are possible actions for *read* and if i is Ann, a and a' should be identical.
2. Bob thinks Ann may have read the letter. The axioms are:
   $D(mayread, a) \equiv a = p \vee a = \neg p \vee a = nil$,
   $A(i, a', a, do(mayread, s)) \equiv D(mayread, a) \wedge D(mayread, a') \wedge$
   $(i = ann \supset a = a')$.

We now introduce some notation which will be used in the rest of the paper. Let $\varphi(s)$ be a formula with a single situation variable s.

- Agent i believes φ:

$$\mathbf{Bel}(i, \varphi(now), s) \stackrel{def}{=} \forall s'.B(i, s', s) \supset \varphi(s').$$

- Agent i truly believes φ:

$$\mathbf{TBel}(i, \varphi(now), s) \stackrel{def}{=} \varphi(s) \wedge \mathbf{Bel}(i, \varphi(now), s).$$

- Agent i believes whether φ holds:

$$\mathbf{BW}(i, \varphi(now), s) \stackrel{def}{=} \mathbf{Bel}(i, \varphi(now), s) \vee \mathbf{Bel}(i, \neg\varphi(now), s).$$

- Let $\mathcal{E}$ be a subset of the agents. We let $C(\mathcal{E}, s', s)$ denote the reflexive transitive closure of $\exists i \in \mathcal{E}.B(i, s', s)$, which can be defined with a second-order formula:

$$C(\mathcal{E}, s', s) \stackrel{def}{=} \forall P.\forall u\, P(u,u) \wedge \forall i{\in}\mathcal{E}, u, v, w[P(u,v) \wedge B(i,v,w) \supset P(u,w)] \supset P(s', s).$$

- The agents commonly believes φ:

$$\mathbf{CKnows}(\varphi(now), s) \stackrel{def}{=} \forall s'.C(\mathcal{A}, s', s) \supset \varphi(s'),$$

where $\mathcal{A}$ is the set of all agents.

4 Properties of beliefs

In this section, we analyze properties of beliefs in our formalism. We begin with the main property of beliefs. We use $\Psi_0(a, s)$ to denote the following formula:

$$\forall i.\mathbf{Bel}(i, \exists s^* \exists a^*.A(i, a^*, a, s) \wedge now = do(a^*, s^*) \wedge Poss(a^*, s^*) \wedge \mathcal{D}_{ss}[a^*, s^*], do(a, s)),$$

where $\mathcal{D}_{ss}[a^*, s^*]$ denotes the instantiation of the SSAs for ordinary fluents wrt a^* and s^*. This says that in the situation resulting from doing action a, each agent i believes that some alternative of a was possible and has happened. We use $\Psi_{n+1}(a, s)$ to denote the following formula:

$$\forall i.\mathbf{Bel}(i, \exists s^* \exists a^*.A(i, a^*, a, s) \wedge now = do(a^*, s^*) \wedge Poss(a^*, s^*) \wedge \mathcal{D}_{ss}[a^*, s^*] \wedge \Psi_n(a^*, s^*), do(a, s)).$$

Thus $\Psi_1(a, s)$ says that in the situation resulting from doing action a, each agent i believes that some alternative a^* of a was possible, has happened, and in the resulting situation, each agent believes that some alternative of a^* was possible and has happened.

By the SSA for the B fluent, it is straightforward to prove:

Theorem 1. $\mathcal{D} \models \forall a \forall s.Poss(a, s) \supset \Psi_n(a, s)$ *for all* n.

Proof. We prove by induction on n. Basis: $n = 0$. This directly follows from the SSA for the B fluent. Induction step: Assume that $\mathcal{D} \models \forall a \forall s.Poss(a,s) \supset \Psi_n(a,s)$. By the SSA for the B fluent, we have

$$\forall a \forall s.Poss(a,s) \supset \forall i.\mathbf{Bel}(i, \exists s^* \exists a^*.A(i,a^*,a,s) \wedge now = do(a^*,s^*) \wedge Poss(a^*,s^*) \wedge \mathcal{D}_{ss}[a^*,s^*], do(a,s)).$$

By the induction hypothesis, we have

$$\forall a \forall s.Poss(a,s) \supset \forall i.\mathbf{Bel}(i, \exists s^* \exists a^*.A(i,a^*,a,s) \wedge now = do(a^*,s^*) \wedge Poss(a^*,s^*) \wedge \mathcal{D}_{ss}[a^*,s^*] \wedge \Psi_n(a^*,s^*), do(a,s)),$$

which is $\forall a \forall s.Poss(a,s) \supset \Psi_{n+1}(a,s)$. □

It is also easy to prove the following propositions. By an objective formula, we mean one with only ordinary fluents.

Proposition 1. *Let φ be objective. Suppose that agent i is an* observer *of action φ in situation σ, i.e., $\mathcal{D} \models A(i,a,\varphi,\sigma) \equiv a = \varphi$. Then $\mathcal{D} \models \varphi(\sigma) \supset \mathbf{Bel}(i,\varphi,do(\varphi,\sigma))$.*

Proposition 2. *Let φ be objective. Suppose that agent i is a* partial observer *of action φ in situation σ, i.e., $\mathcal{D} \models A(i,a,\varphi,\sigma) \equiv a = \varphi \vee a = \neg\varphi$. Then $\mathcal{D} \models \varphi(\sigma) \wedge \neg\mathbf{BW}(i,\varphi,\sigma) \supset \neg\mathbf{BW}(i,\varphi,do(\varphi,\sigma))$.*

Proposition 3. *Let φ be objective. Suppose that agent i is* oblivious *of action α in situation σ, i.e., $\mathcal{D} \models A(i,a,\alpha,\sigma) \equiv a = nil$. Then $\mathcal{D} \models Poss(\alpha,\sigma) \supset [\mathbf{Bel}(i,\varphi,\sigma) \equiv \mathbf{Bel}(i,\varphi,do(\alpha,\sigma))]$.*

In the following, we show how we model some special types of actions and prove the desired properties. We first consider public sensing and reading actions, and give their axioms as follows:

- $sense_\varphi(i,\vec{x})$ means agent i senses the truth value of $\varphi(\vec{x})$
- $read_f(i,\vec{x})$ means agent i reads the value of $f(\vec{x})$

1. $D(sense_\varphi(i,\vec{x}),a) \equiv a = \varphi(\vec{x}) \vee a = \neg\varphi(\vec{x})$
2. $A(j,a',a,do(sense_\varphi(i,\vec{x}),s)) \equiv D(sense_\varphi(i,\vec{x}),a) \wedge D(sense_\varphi(i,\vec{x}),a') \wedge (j = i \supset a = a')$
3. $D(read_f(i,\vec{x}),a) \equiv \exists y.a = [f(\vec{x}) = y]$

4. $A(j,a',a,do(read_f(i,\vec{x}),s)) \equiv$
$$D(read_f(i,\vec{x}),a) \wedge D(read_f(i,\vec{x}),a') \wedge (j = i \supset a = a')$$

It is easy to prove:

Proposition 4. *$\mathcal{D}$ entails the following:*

1. $Do([\![sense_\varphi(i,\vec{x})]\!],s,s_1) \supset [B(j,s',s_1) \equiv$
$$\exists s^*.B(j,s^*,s) \wedge Do([\![sense_\varphi(i,\vec{x})]\!],s^*,s') \wedge (j = i \supset \varphi(\vec{x},s) \equiv \varphi(\vec{x},s^*))]$$
2. $Do([\![read_f(i,\vec{x})]\!],s,s_1) \supset [B(j,s',s_1) \equiv$
$$\exists s^*.B(j,s^*,s) \wedge Do([\![read_f(i,\vec{x})]\!],s^*,s') \wedge (j = i \supset f(\vec{x},s) = f(\vec{x},s^*))]$$

This is the same as Shapiro *et al.* 's extension of Scherl and Levesque's SSA for the K fluent to public sensing and reading actions in the multi-agent case. So our account of beliefs and actions subsumes theirs. As an easy corollary, we get

Proposition 5. *Let φ be objective. Then $\mathcal{D}$ entails the following:*

1. $Do([\![sense_\varphi(i,\vec{x})]\!],s,s_1) \supset \mathbf{BW}(i,\varphi(\vec{x}),s_1) \wedge$
$$(j \neq i \supset \mathbf{Bel}(j,\mathbf{BW}(i,\varphi(\vec{x})),s_1))$$
2. $Do([\![read_f(i,\vec{x})]\!],s,s_1) \supset \exists y\mathbf{Bel}(i,f(\vec{x}) = y,s_1) \wedge$
$$(j \neq i \supset \mathbf{Bel}(j,\exists y\mathbf{Bel}(i,f(\vec{x}) = y),s_1))$$

Bacchus et al. (1999) considered noisy sensors: when an agent reads the value of $f(\vec{x})$, she may get a value y such that $|f(\vec{x},s) - y| \leq b$ for some bound b. This can be easily described as follows:

1. $D(nread_f(i,\vec{x}),a) \equiv \exists y.a = [f(\vec{x}) = y]$
2. $A(j,a',a,do(nread_f(i,\vec{x}),s)) \equiv D(nread_f(i,\vec{x}),a) \wedge D(nread_f(i,\vec{x}),a') \wedge$
$$\{j = i \supset (\exists y,y').a = [f(\vec{x}) = y] \wedge a' = [f(\vec{x}) = y'] \wedge |y - y'| \leq b\}$$

As desired, we have

Proposition 6. $\mathcal{D} \models Do([\![nread_f(i,\vec{x})]\!],s,s') \supset \exists y.\mathbf{Bel}(i,|f(\vec{x}) - y| \leq b,s')$.

Similar to noisy sensors, we may have unintended actions: an agent wants to push button m, but she may push button n such that $|m - n| \leq b$. Other agents can observe that she pushes a button, but have no idea which button she pushes.

1. $D(npush(i,m),a) \equiv \exists n.a = push(n)$
2. $A(j,a',a,do(npush(i,m),s)) \equiv D(npush(i,m),a) \wedge D(npush(i,m),a') \wedge$
   $\{j = i \supset (\exists n,n').a = push(n) \wedge a' = push(n') \wedge |m-n| \leq b \wedge |m-n'| \leq b\}$

Proposition 7. $\mathcal{D} \models Do([\![npush(i,m)]\!],s,s') \supset \mathbf{Bel}(i,\exists n.|n-m| \leq b \wedge on(n),s')$.

Finally, we have the following description for the action of publicly truthfully announcing $\varphi(\vec{x})$:

1. $D(pub_\varphi(\vec{x}),a) \equiv a = \varphi(\vec{x})$
2. $A(i,a',a,do(pub_\varphi(\vec{x}),s)) \equiv a = \varphi(\vec{x}) \wedge a' = \varphi(\vec{x})$

Proposition 8. *Let φ be objective.*
Then $\mathcal{D} \models Do([\![pub_\varphi(\vec{x})]\!],s,s') \supset \mathbf{CKnows}(\varphi(\vec{x}),s')$.

5 Extended examples

In this section, we present two extended examples of modeling multi-agent scenarios in the situation calculus. In the first example, the role of each agent is not common knowledge. The second one involves both physical and sensing actions.

Example 3. Ann senses the truth value of p. Bob and Carol are observing Ann. But Ann doesn't know the role of Bob or Carol. Bob and Carol do not know the role of each other. We introduce two actions:

1. $context(x,y,z)$, where x, y, and z represent the role of *Ann*, *Bob*, and *Carol*, respectively, and each takes the values of 0, 1, and 2: 0 means the agent is an observer, 1 means the agent is a partial observer, and 2 means that the agent is oblivious.
2. *uncertain*, which represents the uncertainty among the different contexts.

The axioms are as follows:

1. $A(i,a',a,do(uncertain,s)) \equiv \exists x,y,z,x',y',z'.$
   $a = context(x,y,z) \wedge a' = context(x',y',z') \wedge$
   $[i = ann \supset x = x'] \wedge$
   $[i = bob \supset y = y' \wedge (x = 0 \wedge y = 1 \supset x' = 0)] \wedge$
   $[i = carol \supset z = z' \wedge (x = 0 \wedge z = 1 \supset x' = 0)]$

2. $D(context(x,y,z),a) \equiv a = p \vee a = \neg p \vee a = nil$
3. $A(i,a',a,do(context(x,y,z),s)) \equiv$
   $D(context(x,y,z),a) \wedge D(context(x,y,z),a') \wedge$
   $[role(context(x,y,z),i) = 0 \supset a = a'] \wedge$
   $[role(context(x,y,z),i) = 1 \supset a = a' \vee a \neq nil \wedge a' \neq nil] \wedge$
   $[role(context(x,y,z),i) = 2 \supset a' = nil]$

The reason we have $[i = ann \supset x = x']$ is that Ann knows her own role. The reason we have $[i = bob \supset y = y' \wedge (x = 0 \wedge y = 1 \supset x' = 0)]$ is that Bob knows his own role and Bob is observing Ann. So the actual context is 011. But to Ann, context 022 is possible; to Bob, context 012 is possible; and to Carol, context 021 is possible.

Assume $\mathcal{D}_{S_0}$ contains $\mathbf{CKnows}(\forall i \neg \mathbf{BW}(i,p), S_0)$.
Let $S_1 = do(uncertain; context(0,1,1); p, S_0)$. Then $\mathcal{D}$ entails the following:

1. $\mathbf{BW}(ann,p,S_1)$;
2. $\neg\mathbf{BW}(bob,p,S_1)$;
3. $\mathbf{Bel}(bob,\mathbf{BW}(ann,p),S_1)$;
4. $\neg\mathbf{Bel}(ann,\mathbf{Bel}(bob,\mathbf{BW}(ann,p)),S_1)$;
5. $\neg\mathbf{Bel}(carol,\mathbf{Bel}(bob,\mathbf{BW}(ann,p)),S_1)$.

Example 4. We use a simplified and adapted version of Levesque's Squirrel World. Squirrels and acorns live in a one-dimensional world unbounded on both sides. Each acorn and squirrel is located at some point, and each point can contain any number of squirrels and acorns. Acorns are completely passive. Squirrels can do the following actions:

1. $left(i)$: Squirrel i moves left a unit;
2. $right(i)$: Squirrel i moves right a unit;
3. $pick(i)$: Squirrel i picks up an acorn, which is possible when he is not holding an acorn and there is at least one acorn at his location;
4. $drop(i)$: Squirrel i drops the acorn he is holding;
5. $learn(i,n)$: Squirrel i learns that there are n acorns at his location. We use $smell(i)$ to denote $(\pi n)learn(i,n)$.

A squirrel can observe the action of another squirrel within a distance of 4, but if the action is a sensing action, the result is not observable. Initially, there are

two acorns at each point. There are three squirrels: Nutty, Edgy, and Wally. Initially, they are all at point 0, holding no acorns, and have no knowledge of the number of acorns at each point, and the above is common knowledge. There are 3 ordinary fluents:

1. $hold(i, s)$: Squirrel i is holding an acorn in situation s;
2. $loc(i, p, s)$: Squirrel i is at location p in situation s;
3. $acorn(p, n, s)$: There are n acorns at location p in s.

For illustration, we only present some axioms of $\mathcal{D}$:

1. $Poss(pick(i), s) \equiv \neg hold(i, s) \wedge \exists p(loc(i, p, s) \wedge acorn(p, n, s) \wedge n > 0)$
2. $loc(i, p, do(a, s)) \equiv \Phi_{loc}(i, p, a, s)$, which is:
   $a = left(i) \wedge loc(i, p + 1, s) \vee a = right(i) \wedge loc(i, p - 1, s) \vee$
   $loc(i, p, s) \wedge a \neq left(i) \wedge a \neq right(i)$
3. $A(i, a', a, do(a^*, s)) \equiv \exists j, p, p'[agt(a) = j \wedge \Phi_{loc}(i, p, a^*, s) \wedge \Phi_{loc}(j, p', a^*, s) \wedge$
   $(|p - p'| > 4 \supset a' = nil) \wedge$
   $(|p - p'| \leq 4 \supset a = a' \vee j \neq i \wedge \exists n, n'(a = learn(j, n) \wedge a' = learn(j, n')))]$
4. $\mathbf{CKnows}(\forall i.loc(i, 0) \wedge \neg hold(i) \wedge \forall p \forall n \neg \mathbf{Bel}(i, \neg acorn(p, n)), S_0)$
5. $\forall p.acorn(p, 2, S_0)$
6. $A(i, a', a, S_0) \equiv a = a' \vee \exists j, n, n'.j \neq i \wedge a = learn(j, n) \wedge a' = learn(j, n')$

Let $\varphi(s, s')$ be a formula. We let $\mathbf{Bel}(i, \varphi(now, prev), s)$ denote
$\forall s'.B(i, s', s) \supset \exists s^* \exists a^*.s' = do(a^*, s^*) \wedge \varphi(s', s^*)$.

Then $\mathcal{D}$ entails the following:

1. $Do(\delta_1, S_0, s) \supset \mathbf{TBel}(N, acorn(0, 1), s) \wedge$
   $\mathbf{CKnows}(hold(N) \wedge \exists n \mathbf{TBel}(N, acorn(0, n)), s)$,
   where $\delta_1 = smell(N); pick(N)$.
2. $Do(\delta_1; \delta_2, S_0, s) \supset \mathbf{CKnows}(\exists n(acorn(1, n, prev) \wedge acorn(1, n + 1, now)), s)$,
   where $\delta_2 = right(N); drop(N)$. This says that the squirrels commonly knows that there is one more acorn at point 1 now than previously.
3. $Do(\delta_1; \delta_2; \delta_3, S_0, s) \supset \mathbf{CKnows}(loc(W, -2) \wedge loc(N, 1) \wedge loc(E, 3), s)$,
   where $\delta_3 = left(W)^2; right(E)^3$.
4. $Do(\delta_1; \delta_2; \delta_3; \delta_4, S_0, s) \supset \mathbf{TBel}(N, hold(W) \wedge loc(W, -3) \wedge loc(E, 2), s) \wedge$
   $\mathbf{Bel}(E, \neg hold(W) \wedge loc(W, -2), s) \wedge \mathbf{Bel}(W, loc(E, 3), s)$,
   where $\delta_4 = smell(W); pick(W); left(W); left(E)$.
   Note that now Edgy and Wally have incorrect beliefs about each other.

6 Conclusions

In this paper, by incorporating the idea of action models from DELs, we have developed a general multi-agent extension of the situation calculus. We analyzed properties of multi-agent beliefs in the situation calculus, and showed that we can provide a uniform treatment of special types of actions, such as public sensing and reading actions, noisy sensors and unintended actions, and public announcements. Since DELs are propositional, an advantage of our work is the gain of more expressiveness and compactness in representation. We gave two extended examples to illustrate modeling of multi-agent scenarios in the situation calculus.

There are a number of topics for future research. First of all, as mentioned in the introduction, van Benthem et al. (2006) generalized the concept of action model to that of update model which can be used to model both epistemic and physical actions. They proposed a logic, called logic of communication and change (LCC), to reason about update models. It would be interesting to explore if we can embed action model logic and further LCC into the situation calculus. Secondly, as shown in the Squirrel World example, because of unreliable sources of information, at certain points, agents may have incorrect beliefs about the world and other agents. When incorrect beliefs lead to inconsistent beliefs, belief revision is necessary for the agents to keep functioning in the world. The DEL community has done extensive work on multi-agent belief revision, and a good reference is (Baltag and Smets 2008). The general idea is this: The semantic model is a plausibility model, where for each agent, there is a plausibility order on the set of states or actions. An agent believes φ if φ holds in the most plausible states. When we update a plausibility model by an action plausibility model, give priority to the action plausibility order. In the future, we would like to incorporate this line of work into the situation calculus. Thirdly, while the focus of the current paper is on the representation side, in the future, we would like to investigate reasoning in multi-agent situation calculus. Finally, we would like to explore multi-agent high-level program execution and develop interesting applications of it.

Acknowledgements We thank Johan van Benthem, Hans van Ditmarsch, Alexandru Baltag and Sonja Smets for helpful discussions on the topic of this paper.

References

F. Bacchus, J. Halpern, and H. Levesque. Reasoning about noisy sensors in the situation calculus. *Artificial Intelligence*, 111:171–208, 1999.

A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. In *Texts in Logic and Games, Vol 3*. 2008.

A. Baltag, L. S. Moss, and S. Solecki. The logic of public announcements, common knowlegde, and private suspicions. In *Proc. of the Conference on Theoretical Aspects of Rationality and Knowledge (TARK-98)*, 1998.

C. Baral. Reasoing about actions and change: from single agent actions to multi-agent actions. In *Proc. of the International Conference on Principles of Knowledge Representation and Reasoning (KR-10)*, 2010.

V. Belle and G. Lakemeyer. Reasoning about imperfect information games in the epistemic situation calculus. In *Proc. of the AAAI Conference on Artificial Intelligence (AAAI-10)*, 2010.

R. F. Kelly and A. R. Pearce. Complex epistemic modalities in the situation calculus. In *Proc. of the International Conference on Principles of Knowledge Representation and Reasoning (KR-08)*, 2008.

G. Lakemeyer and H. J. Levesque. Situations, si! Situation terms, no! In *Proc. of the International Conference on Principles of Knowledge Representation and Reasoning (KR-04)*, 2004.

G. Lakemeyer and H. J. Levesque. Semantics for a useful fragment of the situation calculus. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI-05)*, 2005.

H. J. Levesque, R. Reiter, Y. Lespérance, F. Lin, and R. B. Scherl. GOLOG: A logic programming language for dynamic domains. *J. Logic Programming*, 31 (1-3), 1997.

J. McCarthy and P. J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In *Machine Intelligence 4*, pages 463–502. 1969.

R. Reiter. The frame problem in the situation calculus: A simple solution (sometimes) and a completeness result for goal regression. In *Artificial Intelligence and Mathematical Theory of Computation: Papers in Honor of John McCarthy*, pages 359–380. 1991.

R. Reiter. *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*. MIT Press, 2001.

R. B. Scherl and H. J. Levesque. The frame problem and knowledge-producing actions. In *Proc. of the AAAI Conference on Artificial Intelligence (AAAI-93)*, 1993.

R. B. Scherl and H. J. Levesque. Knowledge, action, and the frame problem. *Artificial Intelligence*, 144(1-2):1–39, 2003.

S. Shapiro, Y. Lespérance, and H. J. Levesque. Specifying communicative multi-agent systems. In *Agents and Multi-Agent Systems – Formalisms, Methodologies, and Applications*, volume 1441 of *Lecture Notes in Computer Science*, pages 1–14. Springer, 1998.

J. van Benthem. McCarthy variations in a modal key. *Artificial Intelligence*, 175 (1):428–439, 2011.

J. van Benthem, J. van Eijck, and B. Kooi. Logics of communication and change. *Information and Computation*, 204(11):1620–1662, 2006.

H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*. Springer, 2007.

H. van Ditmarsch, A. Herzig, and T. de Lima. From situation calculus to dynamic epistemic logic. *J. of Logic and Computation*, 21(2):179–204, 2011.

Proofs nets and the categorial flow of information

Michael Moortgat and Richard Moot

Utrecht Institute of Linguistics OTS, LABRI-CNRS Bordeaux
M.J.Moortgat@uu.nl,moot@labri.fr

Abstract

The Lambek-Grishin calculus (**LG**) is a multiple-conclusion extension of Lambek's categorial type logic with dual families of fusion ('merge') and fission operations, and linear distributivity principles relating these two. Thanks to the distributivity principles, **LG** captures dependency patterns beyond context-free, both in syntax and semantics. In this paper we represent the information flow in categorial derivations in terms of a proof net graphical calculus. We study the correspondence between the composition graphs for these nets and the terms associated with focused sequent derivations.

1 Background, motivation

In this paper, we study **LG**, a type logic based on the generalization of Lambek's Syntactic Calculus proposed in Grishin (1983). The formula language of this

logic is given in (1).

$$\begin{array}{lll} A, B ::= p \mid & & \text{atoms: } s, np, \ldots \\ & A \otimes B \mid B\backslash A \mid A/B \mid & \text{product, left vs right division} \\ & A \oplus B \mid A \oslash B \mid B \ominus A & \text{coproduct, right vs left difference} \end{array} \tag{1}$$

Algebraically, **LG** combines the residuated triple of (3) — fusion with its two residuals — with the dual residuated triple in (4): fission, left and right difference.

$$A \leq A \quad ; \quad \text{from } A \leq B \text{ and } B \leq C \text{ infer } A \leq C \tag{2}$$

$$A \leq C/B \quad \text{iff} \quad A \otimes B \leq C \quad \text{iff} \quad B \leq A\backslash C \tag{3}$$

$$B \ominus C \leq A \quad \text{iff} \quad C \leq B \oplus A \quad \text{iff} \quad C \oslash A \leq B \tag{4}$$

For the interaction between the fusion and fission families, we have the postulates of (5).[1] These postulates have come to be called *linear distributivity principles* (e.g. Cockett and Seely (1996)): linear, because they respect resources (no material gets copied).

$$\begin{array}{ll} (A \ominus B) \otimes C \leq A \ominus (B \otimes C) & C \otimes (B \oslash A) \leq (C \otimes B) \oslash A \\ C \otimes (A \ominus B) \leq A \ominus (C \otimes B) & (B \oslash A) \otimes C \leq (B \otimes C) \oslash A \end{array} \tag{5}$$

LG is attractive for syntactic and for semantic reasons. Syntactically, the interaction principles of (5) bring expressivity beyond context-free. Moot (2007) gives an **LG** encoding of the adjunction operation of Tree Adjoining Grammar, the most restricted formalism in the mildly context-sensitive hierarchy Kallmeyer (2010); Moortgat (2009) has an **LG** grammar for MIX, according to Salvati (p.c.) an instance of a non-wellnested 2-MCFG. The upper bound for the syntactic expressivity of **LG** grammars in their full generality is open; see Melissen (2010) for discussion.

At the semantic level, **LG** derivations can be given an interpretation in the continuation-passing style. The CPS interpretation leads to a considerable simplification of the syntax/semantics interface: semantic scope construal can

[1] There is a second set, with the inequalities reversed, which we'll not discuss here.

be obtained on the basis of simple first-order syntactic types, as shown in Bastenhof (2012) and discussed in §3.1.

Despite these attractions, working with the standard 'symbolic' presentations of **LG** involves rather formidable technical machinery. Our focus in this paper is on *proof nets* — an elegant graphical calculus that captures the essence of **LG** derivations without the bureaucracy of heavy symbol manipulation.

2 Display sequent calculus and proof nets

A sequent calculus for **LG**, in the 'display logic' style, can be found in Goré (1997). We present it in §2.1, using the notation of Moortgat (2009). In §2.2, we introduce the proof net graphical calculus, and show how it leads to a representation of **LG** derivations that is free of spurious ambiguities.

2.1 sLG: display sequent calculus

The characteristic feature of Display Logic is that for every logical connective, there is a corresponding structural connective, not just for conjunction and disjunction as in standard sequent calculus. We use the same symbols for the logical operations and their structural counterparts; structural operations are marked off by centerdots. Below the grammar for input (sequent left hand side), and output structures (sequent rhs). Atomic structures are formulas $\mathcal{F}$.

$$\begin{array}{lcl} \mathcal{I} & ::= & \mathcal{F} \mid \mathcal{I} \cdot \otimes \cdot \mathcal{I} \mid \mathcal{I} \cdot \oslash \cdot \mathcal{O} \mid \mathcal{O} \cdot \oslash \cdot \mathcal{I} \\ \mathcal{O} & ::= & \mathcal{F} \mid \mathcal{O} \cdot \oplus \cdot \mathcal{O} \mid \mathcal{I} \cdot \backslash \cdot \mathcal{O} \mid \mathcal{O} \cdot / \cdot \mathcal{I} \end{array} \tag{6}$$

Figures 1 and 2 then give the structural and logical rules of **sLG**. The (dual) residuation principles take the form of 'display postulates', so called because they allow any formula component of a structure to be displayed as the sole occupant of the sequent lhs or rhs. The logical rules apply to formulas thus displayed. The one-premise rules simply replace a logical connective by its structural counterpart; these rules are invertible. The non-invertible two-premise rules give expression to the monotonicity properties of the type-forming operations. The distributivity principles (5) appear in rule form here as $G1 - G4$ in the structural group.

$$\frac{}{A \Rightarrow A}\ \text{Ax} \qquad \frac{X \Rightarrow A \quad A \Rightarrow Y}{X \Rightarrow Y}\ \text{Cut}$$

$$\dfrac{\dfrac{X \Rightarrow Z \cdot / \cdot Y}{X \cdot \otimes \cdot Y \Rightarrow Z}\ rp}{Y \Rightarrow X \cdot \backslash \cdot Z}\ rp \qquad \dfrac{\dfrac{Y \cdot \oslash \cdot Z \Rightarrow X}{Z \Rightarrow Y \cdot \oplus \cdot X}\ drp}{Z \cdot \oslash \cdot X \Rightarrow Y}\ drp$$

$$\frac{X \cdot \otimes \cdot Y \Rightarrow Z \cdot \oplus \cdot W}{Z \cdot \oslash \cdot X \Rightarrow W \cdot / \cdot Y}\ G1 \qquad \frac{X \cdot \otimes \cdot Y \Rightarrow Z \cdot \oplus \cdot W}{Y \cdot \oslash \cdot W \Rightarrow X \cdot \backslash \cdot Z}\ G3$$

$$\frac{X \cdot \otimes \cdot Y \Rightarrow Z \cdot \oplus \cdot W}{Z \cdot \oslash \cdot Y \Rightarrow X \cdot \backslash \cdot W}\ G2 \qquad \frac{X \cdot \otimes \cdot Y \Rightarrow Z \cdot \oplus \cdot W}{X \cdot \oslash \cdot W \Rightarrow Z \cdot / \cdot Y}\ G4$$

Figure 1: **sLG**. Structural rules.

$$\frac{A \cdot \$ \cdot B \Rightarrow Y}{A\ \$\ B \Rightarrow Y}\ \$L \quad \$ \in \{\otimes, \oslash, \varnothing\} \qquad \frac{X \Rightarrow A \cdot \# \cdot B}{X \Rightarrow A\ \#\ B}\ \#R \quad \# \in \{\oplus, /, \backslash\}$$

$$\frac{X \Rightarrow A \quad Y \Rightarrow B}{X \cdot \otimes \cdot Y \Rightarrow A \otimes B}\ \otimes R \qquad \frac{A \Rightarrow X \quad B \Rightarrow Y}{A \oplus B \Rightarrow X \cdot \oplus \cdot Y}\ \oplus L$$

$$\frac{X \Rightarrow A \quad B \Rightarrow Y}{A \backslash B \Rightarrow X \cdot \backslash \cdot Y}\ \backslash L \qquad \frac{X \Rightarrow A \quad B \Rightarrow Y}{X \cdot \oslash \cdot Y \Rightarrow A \oslash B}\ \oslash R$$

$$\frac{X \Rightarrow A \quad B \Rightarrow Y}{B / A \Rightarrow Y \cdot / \cdot X}\ /L \qquad \frac{X \Rightarrow A \quad B \Rightarrow Y}{Y \cdot \oslash \cdot X \Rightarrow B \oslash A}\ \oslash R$$

Figure 2: **sLG**. Logical rules.

It is shown in Moortgat (2009) that the display sequent format **sLG** enjoys cut elimination and thus allows for decidable proof search. Still, there is room for improvement:

- spurious ambiguity: as with sequent calculi in general, one and the same matching of occurrences of atomic subformulae in a proof's axiom leaves may be obtained in different ways as a result of irrelevant rule permutations;

- no parsing: backward chaining sequent proof search requires the *structure* of the formulas making up the end sequent to be given in advance; for genuine **LG** *parsing*, one would like this structure to be computed as the outcome of the deduction process;
- display equivalences represent alternative views on one and the same structure: one would like to have a proof format where there is no need for the explicit structural manipulations of the display postulates.

The proof net approach to be discussed below removes these problematic aspects.

2.2 Proof nets

Proof nets are a graphical way of representing proofs, introduced first for linear logic Girard (1987). The proof nets for **LG** we present in this section are a simple extension of the proof nets for the multimodal Lambek calculus of Moot and Puite (2002). A proof structure is a (hyper)graph where the vertices are labeled by formulas and the edges connect these formulas.[2] The hyperedges correspond to the logical rules, linking the active formulas and the main formula of the rule and keeping track of whether one is dealing with a non-invertible two-premise rule or with an invertible one-premise rule. We'll call these *tensor* and *cotensor* links respectively.

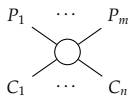
Definition 2.1. A link is a tuple $\langle t, p, c, m\rangle$ where

- t is the type of the link — tensor or cotensor
- p is the list of premises of the link,
- c is the list of conclusions of the link,
- m, the main vertex/formula of the link, is either a member of p, a member of c or the constant "nil".

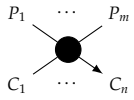
In case m is a member of p we speak of a *left* link (corresponding to the left rules of the sequent calculus, where the main formula of the link occurs in the antecedent) and in case m is a member of c we speak of a *right* link.

[2] In what follows we will often speak of formula occurrences (or simply *formulas* if there is no possibility of confusion) instead of vertices labeled by formulas.

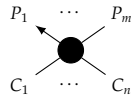
Graphically, links are displayed as shown below. A central node links together the premises and conclusions of the link; when we need to refer to the connections between the central node and the vertices, we will call them its *tentacles*. The interior of this central node is white for a tensor link and black for a cotensor link. The premises are drawn, in left-to-right order, above the central node and the conclusions, also in left-to-right order, are drawn below it. The main formula of cotensor links is drawn with an arrow towards it; the main formula of a tensor link can only be determined by inspection of the formulas.



tensor rule



cotensor rule (right rule)



cotensor rule (left rule)

Figure 3 shows the links for **LG**. The links for the fission connectives are up-down symmetric versions of the links for the fusion connectives.

Definition 2.2. A *proof structure* $\langle S, \mathcal{L} \rangle$ is a finite set of formula occurrences S and a set of links $\mathcal{L}$ from those shown in Figure 3 such that

- each formula is at most once the premise of a link,
- each formula is at most once the conclusion of a link.

Formulas which are not the conclusion of any link are called the *hypotheses* of the proof structure. Formulas which are not the premise of any link are called the *conclusions* of the proof structure.

We will say that a proof structure with hypotheses $H_1, \ldots, H_m$ and conclusions $C_1, \ldots, C_n$ is a proof structure of $H_1, \ldots, H_m \Rightarrow C_1, \ldots C_n$.

Example 1. Figure 4 shows the hypothesis unfolding of $(s \oslash s) \oslash np$ and the conclusion unfolding of $s \,/\, (np \setminus s)$. Both are obtained by simple application of the rules of Figure 3 until we reach the atomic subformulas.

Though the figure satisfies the condition on proof structures (connectedness is not a requirement), it is a proof structure of $(s \oslash s) \oslash np, s, s, np \Rightarrow s \,/\, (np \setminus s), s, s, np$. We obtain a proof structure of $(s \oslash s) \oslash np \Rightarrow s \,/\, (np \setminus s)$ by identifying atomic formulas.[3] In this case, we choose to identify the top s of the left subgraph

[3] This node identification corresponds to the "axiom links" of linear logic proof nets.

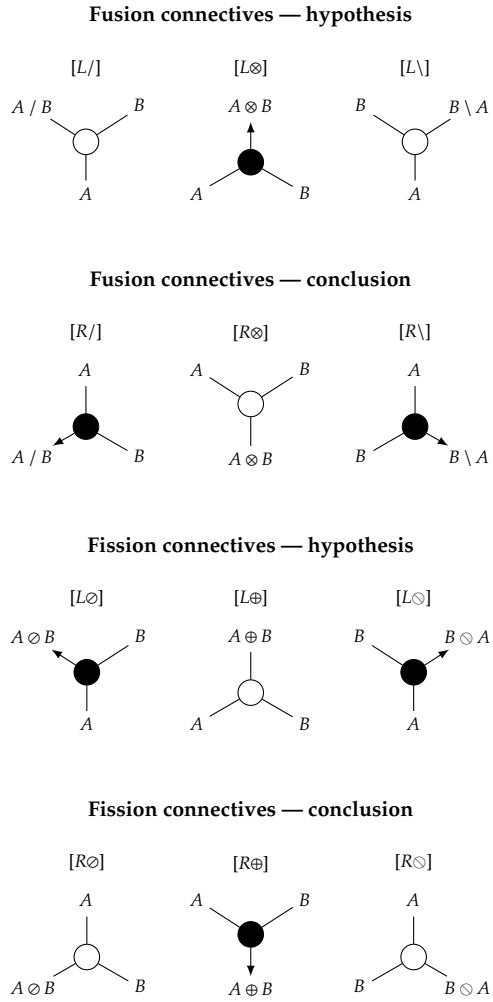


Figure 3: Links for proof structures of the Lambek-Grishin calculus

with the bottom s of the right subgraph and perform the unique choice for the remaining atomic formulas. The result is the proof structure shown in Figure 5

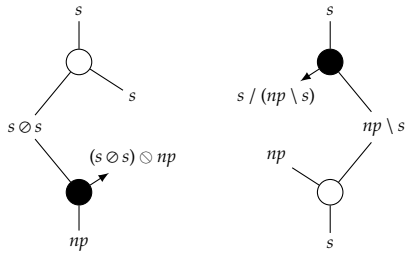


Figure 4: Lexical unfolding

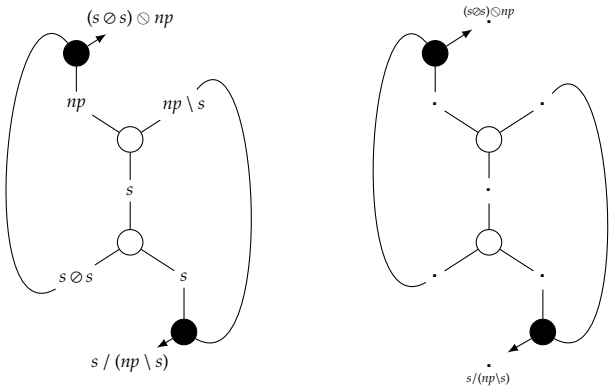


Figure 5: Proof structure of $(s \oslash s) \oslash np \Rightarrow s \,/\, (np \setminus s)$ corresponding to the lexical unfolding of Figure 4 and its corresponding abstract proof structure

on the left.

Due to the graphical constraints of writing these proof nets on the plane — we want to draw the $np \setminus s$ node *below* the cotensor link at the bottom of the figure, since it is a conclusion of this link, but would have to draw the figure on a cylinder to make this work — we need to use curved tentacles connect the minor premise of (co-)implication links to the rest of the proof structure.

Definition 2.3. An *abstract proof structure* $\langle V, \mathcal{L}, h, c \rangle$ is a set of vertices V, a set of (unlabeled) links $\mathcal{L}$ and two functions h and c, such that

- each formula is at most once the premise of a link,
- each formula is at most once the conclusion of a link,
- h and c are functions from the hypotheses resp. conclusions of the abstract proof structure to formulas

Note that the abstract proof structure corresponding to a two formula sequent $A \Rightarrow B$ has only a single vertex v, with $h(v) = A$ and $c(v) = B$.

The transformation from proof structure to abstract proof structure is a forgetful mapping: we transform a proof structure into an abstract proof structure by erasing all formula information on the internal vertices. Visually, we remove the formula labels from the graph and replace them by simple vertices (**.**). We indicate the results of the functions h and c above (resp. below) the vertices (for the hypotheses and conclusions respectively). As a result, we have to following four types of vertices in an abstract proof structure.

internal hypothesis conclusion both

Example 2. Figure 5 shows (on the right) the transformation of the proof structure on its left into an abstract proof structure.

Definition 2.4. A *tree* is an acyclic, connected abstract proof structure which does not contain any cotensor links.

The trees of Definition 2.4 correspond to sequents in a rather direct way. In fact, they have the pleasant property of "compiling away" the display rules of the sequent calculus. Or, in other words, trees represent a class of sequents which is equivalent up to the display postulates.

Definition 2.5. Given an abstract proof structure A, we say that A *contracts in one step* to A', written $A \to A'$ iff A' is obtained from A by replacing one of the subgraphs of the form shown in Figures 6 and 7 by a single vertex.

H represents the result of the function h for the indicated node (relevant only in case this node is a hypothesis of the abstract proof structure). Similarly, C represents the formula assigned by the function c to the indicated node.

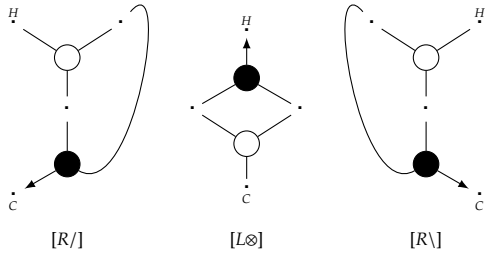


Figure 6: Contractions — Lambek connectives

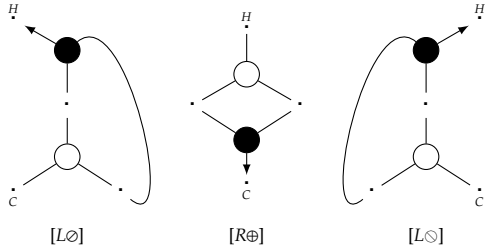


Figure 7: Contractions — Grishin connectives

Given an abstract proof structure A we say that A *contracts to* an abstract proof structure A' if there is a sequence of zero or more one step contractions from A to A'.

When we say that a *proof structure* P contracts to an abstract proof structure A' we will mean that the underlying abstract proof structure A of P contracts to A'.

To obtain expressivity beyond context-free, we are interested in **LG** with added interaction principles. Figures 8 and 9 give the additional rewrite rules on abstract proof structures that correspond to the rule form of Grishin's distributivity laws.

Definition 2.6. A proof structure P is a *proof net* iff its underlying abstract proof structure A converts to a tree using the contractions of Figures 6 and 7 and the

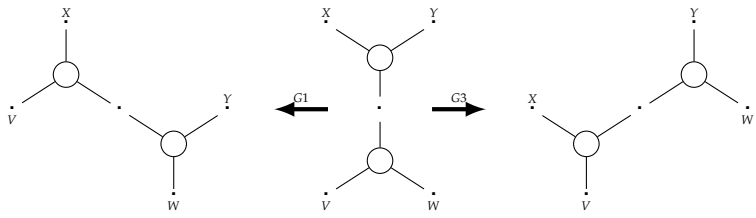


Figure 8: Grishin interactions I — "mixed associativity"

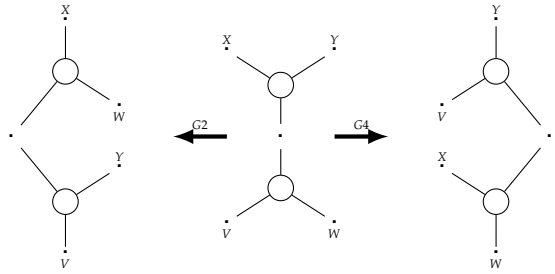


Figure 9: Grishin interactions II — "mixed commutativity"

structural rules of Figures 8 and 9.

Theorem 1. *A proof structure P is a proof net — that is, P converts to a tree T — iff there is a sequent proof of T.*

The proof is an easy adaptation of the proof of Moot and Puite (2002). A detailed proof can be found in Moot (2007).

Example 3. We show that the proof structure of Figure 5 is a proof net by contracting it to a tree. Starting with rule ($G1$), the two cotensor links can be contracted in any order. Figure 10 shows a complete sequence.

Example 4. For a second example (to be taken up again when we discuss focused proof search in §3) we turn to Figure 11 which shows the lexical proof structures for a generalized quantifier noun phrase, a transitive verb, a determiner and a lexical noun.

Consider the sentence 'everyone likes the teacher'. In the unfocused sequent calculus **sLG**, the sequent $(np \,/\, n) \otimes n, (np \setminus s) \,/\, np, np \,/\, n, n \Rightarrow s$ has at least seven

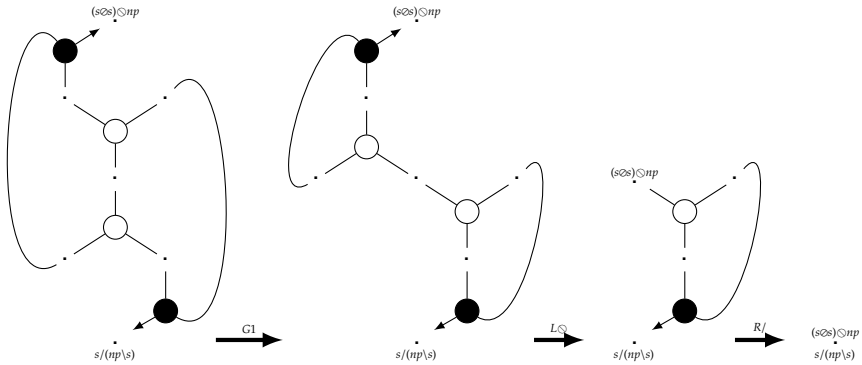


Figure 10: Reducing the abstract proof structure of Figure 5 to a tree.

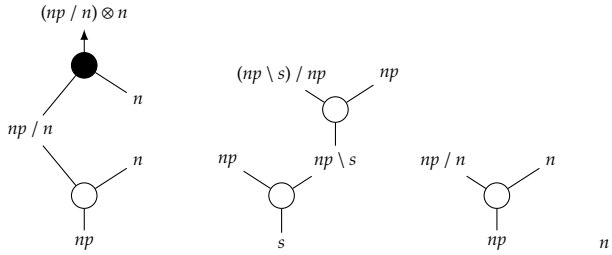


Figure 11: Lexical proof structures for a generalized quantifier noun phrase, a transitive verb, a determiner and a noun.

proofs, depending on the order of application of the introduction rules for the five occurrences of the logical connectives involved: ⊗ (once), / (three times), \ (once). Figure 12 gives, on the left, the *single* possible identification of *n* and *np* formulas that gives rise to a proof net with the lexical entries in the desired order. The corresponding abstract proof structure is given in the middle. This abstract proof structure allows us to apply a contraction directly, as shown on the right.

The table below summarizes the correspondence between proof nets and sequent proofs.

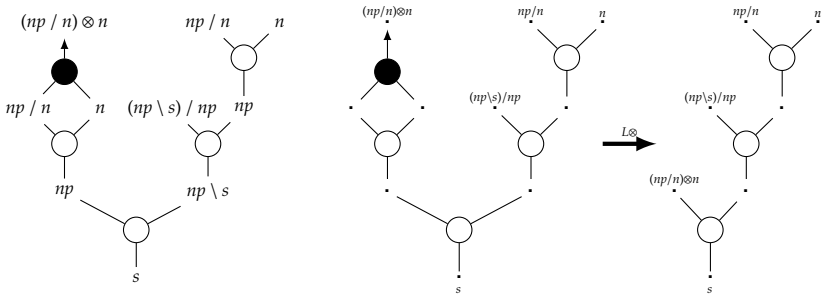


Figure 12: Judgement $(np \,/\, n) \otimes n, (np \setminus s) \,/\, np, np \,/\, n, n \Rightarrow s$: proof structure, abstract proof structure and contraction.

sequent calculus	proof structure	conversion
axiom	axiomatic formula	—
cut	cut formula	—
two-premise rule	tensor link	—
one-premise rule	cotensor link	contraction
interaction rule	—	rewrite

The invertible one-premise rules correspond to both a link and a contraction and the interaction rules are invisible in the proof structure, appearing only in the conversion sequence.

With a bit of extra effort in the sequentialization proof (and the exclusion of cuts on axioms) we can show that these correspondences are 1-to-1, that is each axiomatic formula in a proof net corresponds to exactly one axiom rule in the sequent proof, each non-invertible two-premise rule corresponds to exactly one link in the proof net and each invertible one-premise rule to exactly one link in the proof net and exactly one contraction in its conversion sequence.

Summing up, the proof net approach offers the following benefits in comparison to sequent proof search.

- Parsing. Whereas for sequent proof search the structure of the sequent has to be given, the contraction sequence that identifies a proof structure as a proof net actually *computes* this structure.
- Removal of spurious ambiguity. Proof nets, like (product-free) natural deduction, have different proof objects only for proofs of a judgement which differ essentially. The combinatorial possibilities for such readings, which are obtained by finding a complete matching of the premise and conclusion atomic formulas, can easily be enumerated for a given sequence of formulas.
- Display rules compiled away. The tensor trees associated with well-formed proof nets represent a class of sequents which is equivalent up to the display postulates.

3 Proof nets and focused display calculus

The spurious non-determinism of naive backward-chaining proof search can also be addressed within the sequent calculus itself, by introducing an appropriate notion of 'normal' derivations. In §3.1, we introduce **fLG**, a focused version of the sequent calculus for **LG**. In §3.2, we then study how to interpret focused derivations from a proof net perspective.

3.1 fLG: focused display calculus

The strategy of focusing has been well-studied in the context of linear logic, starting with the work of Andreoli Andreoli (2001). It is based on the distinction between *asynchronous* and *synchronous* non-atomic formulas. The introduction rule for the main connective of an asynchronous formula is *invertible*; it is non-invertible for the synchronous formulas. Backward chaining focused proof search starts with an asynchronous phase where invertible rules are applied deterministically until no more candidate formulas remain. At that point, a non-deterministic choice for a synchronous formula must be made: this formula is put 'in focus', and decomposed in its subformulae by means of non-invertible rules until no more non-invertible rules are applicable, at which point one reenters an asynchronous phase. The main result of Andreoli (2001) is that focused proofs are complete for linear logic.

Focused proof search for the Lambek-Grishin calculus has been studied by Bastenhof (2011) who uses a one-sided presentation of the calculus. In this section, we implement his focusing regime in the context of the two-sided sequent format of Bernardi and Moortgat (2010). We proceed in two steps. First we introduce **fLG**, the focused version of the sequent calculus of §2.1. **fLG** makes a distinction between focused and unfocused judgements, and has a set of inference rules to switch between these two. **fLG** comes with a term language that is in Curry-Howard correspondence with its derivations. This term language is a directional refinement of the $\overline{\lambda}\mu\widetilde{\mu}$ language of Curien and Herbelin (2000).

The second step is to give a constructive interpretation for **LG** derivations by means of a continuation-passing-style translation: a mapping $\lceil\cdot\rceil$ that sends derivations of the multiple-conclusion source logic to (natural deduction) proofs in a fragment of single-conclusion intuitionistic Linear Logic MILL (in the categorial terminology: **LP**). For the translation of Bastenhof (2011) that we follow here, the target fragment has linear products and negation $A^{\perp}$, i.e. a restricted form of linear implication $A \multimap \perp$, where $\perp$ is a distinguished atomic type, the response type. Focused source derivations then can be shown to correspond to distinct *normal* natural deduction proofs in the target calculus.

$$\mathbf{fLG}^{\mathcal{A}}_{/,\otimes,\backslash,\oslash,\oplus,\oslash} \xrightarrow{\lceil\cdot\rceil} \mathbf{LP}^{\mathcal{A}\cup\{\perp\}}_{\otimes,\cdot^{\perp}} \quad \left(\xrightarrow{\cdot^{\ell}} \mathbf{IL}^{\{e,t\}}_{\times,\rightarrow} \right)$$

For the linguistic illustrations in §3.2, we compose the CPS translation $\lceil\cdot\rceil$ with a second mapping $\cdot^{\ell}$, that establishes the connection with Montague-style semantic representations. This mapping sends the linear constructs to their intuitionistic counterparts, and allows *non-linear* meaning recipes for the translation of the lexical constants.

fLG: proofs and terms In the Curry-Howard proofs-as-programs tradition, we set up **fLG** starting from a term language for which the sequent logic then provides the type system. The term language encodes the *logical* steps of a derivation (left and right introduction rules, and the new set of left and right (de)focusing rules, to be introduced below); structural rules (residuation, distributivity) leave no trace in the proof terms.

Sequent structures, as in §2.1, are built out of formulas. Input formulas now are labeled with variables $x, y, z, \ldots$, output formulas with covariables $\alpha, \beta, \gamma, \ldots$.

To implement the focusing regime, we allow sequents to have one displayed formula *in focus*. Writing the focused formula in a box, **fLG** will have to deal with three types of judgements: sequents with no formula in focus (we'll call these *structural*), and sequents with a succedent or antecedent formula in focus.

$$X \vdash Y \qquad X \vdash \boxed{A} \qquad \boxed{A} \vdash Y$$

Corresponding to the types of sequents, the term language has three types of expressions: *commands*, *values* and *contexts* respectively. For commands, we use the metavariables c, C, for values v, V, for contexts e, E. The typing rules below provide the motivation for the subclassification.

$$\begin{gathered} v ::= \mu\alpha.C \mid V \quad ; \quad V ::= x \mid v_1 \otimes v_2 \mid v \oslash e \mid e \oslash v \\ e ::= \widetilde{\mu}x.C \mid E \quad ; \quad E ::= \alpha \mid e_1 \oplus e_2 \mid v\backslash e \mid e/v \\ c ::= \langle x \upharpoonleft E \rangle \mid \langle V \upharpoonright \alpha \rangle \\ C ::= c \mid \tfrac{x\ y}{z}.C \mid \tfrac{x\ \beta}{z}.C \mid \tfrac{\beta\ x}{z}.C \mid \tfrac{\alpha\ \beta}{\gamma}.C \mid \tfrac{x\ \beta}{\gamma}.C \mid \tfrac{\beta\ x}{\gamma}.C \end{gathered} \tag{7}$$

Typing rules To enforce the alternation between asynchronous and synchronous phases of focused proof search, formulas are associated with a polarity: *positive* for non-atomic formulas with invertible left introduction rule: $A \otimes B$, $A \oslash B$, $B \oslash A$; *negative* for non-atomic formulas with invertible right introduction rule: $A \oplus B$, $A\backslash B$, B/A. For atomic formulas, one can fix an arbitrary polarity. Different choices lead to different prooftheoretic behaviour (and to different interpretations, once we turn to the CPS translation). We will assume that atoms are assigned a bias (positive or negative) in the lexicon. Below the typing rules for **fLG** (restricting attention to the cut-free system).

(Co-)Axiom, (de)focusing First we have the focused version of the axiomatic sequents, and rules for focusing and defocusing which are new with respect to the unfocused presentation of §2.1. There is a polarity restriction on the formula A in these rules: the boxed formula has to be negative for CoAx, $\mu, \widetilde{\mu}^*$; for Ax, $\widetilde{\mu}, \mu^*$ it has to be positive. In the (Co-)Axiom cases, A can be required to be atomic.

$$\frac{}{x : A \vdash \boxed{x : A}}\ \mathrm{Ax} \qquad \frac{}{\boxed{\alpha : A} \vdash \alpha : A}\ \mathrm{CoAx}$$

$$\frac{X \vdash \boxed{V : A}}{\langle V \uparrow \alpha \rangle : (X \vdash \alpha : A)}\ \mu^* \qquad \frac{\boxed{E : A} \vdash X}{\langle x \downarrow E \rangle : (x : A \vdash X)}\ \widetilde{\mu}^* \tag{8}$$

$$\frac{C : (x : A \vdash X)}{\boxed{\widetilde{\mu} x.C : A} \vdash X}\ \widetilde{\mu} \qquad \frac{C : (X \vdash \alpha : A)}{X \vdash \boxed{\mu \alpha.C : A}}\ \mu$$

From a backward-chaining perspective, the $\mu, \widetilde{\mu}$ rules *remove* the focus from a focused succedent or antecedent formula. The result is an unfocused premise sequent, the domain of applicability of the invertible rules, i.e. one enters the asynchronous phase. From the same perspective, the rules $\mu^*, \widetilde{\mu}^*$ place a succedent or antecedent formula in focus, shifting control to the non-invertible rules of the synchronous phase. The $\mu^*, \widetilde{\mu}^*$ rules are in fact instances of Cut where one of the premises is axiomatic.

Invertible rules The term language makes a distinction between simple commands c (the image of the focusing rules $\widetilde{\mu}^*, \mu^*$: $\langle x \downarrow E \rangle, \langle V \uparrow \alpha \rangle$) from extended commands C. The latter start with a sequence of invertible rewrite rules replacing a logical connective by its structural counterpart. We impose the requirement that in the asynchronous phase all formulas to which an invertible rule is applicable are indeed decomposed.

$$\frac{C : (x : A \cdot \otimes \cdot y : B \vdash X)}{\frac{x\ y}{z}.C : (z : A \otimes B \vdash X)}\ \otimes L \qquad \frac{C : (X \vdash \alpha : A \cdot \oplus \cdot \beta : B)}{\frac{\alpha\ \beta}{\gamma}.C : (X \vdash \gamma : A \oplus B)}\ \oplus R$$

$$\frac{C : (x : A \cdot \oslash \cdot \beta : B \vdash X)}{\frac{x\ \beta}{z}.C : (z : A \oslash B \vdash X)}\ \oslash L \qquad \frac{C : (X \vdash x : A \cdot \backslash \cdot \beta : B)}{\frac{x\ \beta}{\gamma}.C : (X \vdash \gamma : A \backslash B)}\ \backslash R \tag{9}$$

$$\frac{C : (\beta : B \cdot \oslash \cdot x : A \vdash X)}{\frac{\beta\ x}{z}.C : (z : B \oslash A \vdash X)}\ \oslash L \qquad \frac{C : (X \vdash \beta : B \cdot / \cdot x : A)}{\frac{\beta\ x}{\gamma}.C : (X \vdash \gamma : B / A)}\ / R$$

Non-invertible rules When a positive (negative) formula has been brought into focus in the succedent (antecedent), one is committed to transfer the focus to its subformulae.

$$\frac{\boxed{e_1 : B} \vdash Y \qquad \boxed{e_2 : A} \vdash X}{\boxed{e_1 \oplus e_2 : B \oplus A} \vdash Y \cdot \oplus \cdot X}\,\oplus L \qquad \frac{X \vdash \boxed{v_1 : A} \qquad Y \vdash \boxed{v_2 : B}}{X \cdot \otimes \cdot Y \vdash \boxed{v_1 \otimes v_2 : A \otimes B}}\,\otimes R$$

$$\frac{X \vdash \boxed{v : A} \qquad \boxed{e : B} \vdash Y}{\boxed{v \backslash e : A \backslash B} \vdash X \cdot \backslash \cdot Y}\,\backslash L \qquad \frac{X \vdash \boxed{v : A} \qquad \boxed{e : B} \vdash Y}{X \cdot \oslash \cdot Y \vdash \boxed{v \oslash e : A \oslash B}}\,\oslash R \tag{10}$$

$$\frac{\boxed{e : B} \vdash Y \qquad X \vdash \boxed{v : A}}{\boxed{e / v : B / A} \vdash Y \cdot / \cdot X}\,/ L \qquad \frac{\boxed{e : B} \vdash Y \qquad X \vdash \boxed{v : A}}{Y \cdot \oslash \cdot X \vdash \boxed{e \oslash v : B \oslash A}}\,\oslash R$$

Derived inference rules: focus shifting To highlight the correspondence with the algorithm for proof net construction to be discussed in §2.2, we will use a derived rule format for shifting between a conclusion and premise focused formula. A branch from ($\widetilde{\mu}^*$) via a sequence (possibly empty) of structural rules and rewrite rules to (μ) is compiled in a derived inference rule with the $\widetilde{\mu}^*$ restrictions on A and the μ restrictions on B.

$$\frac{\dfrac{\boxed{E : A} \vdash Y}{\langle x \uparrow E \rangle : (x : A \vdash Y)}\,\widetilde{\mu}^*}{\begin{array}{c}\vdots \\ (res, distr, rewrite) \\ \vdots \\ (\div)\langle x \uparrow E\rangle : (X \vdash \beta : B)\end{array}}\,\mu \qquad \rightsquigarrow \qquad \frac{\boxed{E : A} \vdash Y}{X \vdash \boxed{\mu\beta.(\div)\langle x \uparrow E\rangle : B}}\,\leftrightarrows$$

For the combinations of $\mu^*, \widetilde{\mu}^*$ and $\mu, \widetilde{\mu}$, this results in the focus shifting rules below. We leave it to the reader to add the terms.

$$\frac{\boxed{A} \vdash Y}{X \vdash \boxed{B}} \leftrightarrows \qquad \frac{X' \vdash \boxed{A}}{X \vdash \boxed{B}} \Rightarrow \qquad \frac{X \vdash \boxed{A}}{\boxed{B} \vdash Y} \rightleftharpoons \qquad \frac{\boxed{A} \vdash Y'}{\boxed{B} \vdash Y} \Leftarrow \tag{11}$$

Example 5. We illustrate the effect of the focusing regime with some alternative ways of assigning a polarity bias to atomic formulas with a simple Subject-Transitive Verb-Object sentence. Examples with lexical material filled in would be 'everyone seeks/finds a unicorn'.

$$(np/n \otimes n) \cdot \otimes \cdot ((np\backslash s)/np \cdot \otimes \cdot (np/n \cdot \otimes \cdot n)) \vdash s \tag{12}$$

For the Object we have a Determiner-Noun combination. For the Subject, we take a product type $(np/n)\otimes n$, so that we have a chance to illustrate the working of the asynchronous phase of the derivation. In the discussion of Figure 12, we saw that (12) has multiple proofs in the unfocused sequent calculus, but only one proof net, i.e. one way of matching the premise and conclusion atoms.

What about the focused calculus **fLG**? Before answering this question, we have to decide on the polarization of the atomic types. Suppose we give them uniform negative bias. There is only one focused proof then, with proof term (13): 'goal driven', top-down, to use parsing terminology. In the proof term, we write tv for the transitive verb; det for the object determiner; noun for the object common noun; subj for the subject noun phrase.

$$\begin{gathered}\mu\beta.(\frac{y\ z}{\mathsf{subj}}.\langle\ \mathsf{tv} \upharpoonright ((Q \setminus \beta) \,/\, Q')\ \rangle) \quad \text{with} \\ Q : \mu\gamma.\langle\ y \upharpoonright (\gamma \,/\, \mu\gamma'.\langle\ z \upharpoonright \gamma'\rangle)\rangle\ ,\ Q' : \mu\alpha.\langle\ \mathsf{det} \upharpoonright (\alpha \,/\, \mu\alpha'.\langle\ \mathsf{noun} \upharpoonright \alpha'\rangle)\rangle\end{gathered} \tag{13}$$

As an alternative, suppose basic type s keeps its negative bias, resetting the sentence continuation for each clausal domain, but the other basic types are assigned positive bias. We now have *two* focused derivations: 'data driven', bottom-up. To make sense of this difference, we will have to look at the CPS translation of these proofs, to be introduced below.

$$\mu\alpha.(\frac{x'\ z}{\mathsf{subj}}.\langle\ x' \upharpoonright (\widetilde{\mu}x.\langle\ \mathsf{det} \upharpoonright (\widetilde{\mu}y.\langle\ \mathsf{tv} \upharpoonright ((x \setminus \alpha) \,/\, y)\ \rangle \,/\, \mathsf{noun})\ \rangle \,/\, z)\ \rangle) \tag{14}$$

Table 1: CPS translation: non-atomic types

pol(·)							
A	B	$\lceil A \otimes B \rceil$	$\lceil A/B \rceil$	$\lceil B\backslash A \rceil$	$\lceil A \oplus B \rceil$	$\lceil A \oslash B \rceil$	$\lceil B \oplus A \rceil$
$-$	$-$	$\lceil A \rceil^\perp \otimes \lceil B \rceil^\perp$	$\lceil A \rceil \otimes \lceil B \rceil^\perp$	$\lceil B \rceil^\perp \otimes \lceil A \rceil$	$\lceil A \rceil \otimes \lceil B \rceil$	$\lceil A \rceil^\perp \otimes \lceil B \rceil$	$\lceil B \rceil \otimes \lceil A \rceil^\perp$
$-$	$+$	$\lceil A \rceil^\perp \otimes \lceil B \rceil$	$\lceil A \rceil \otimes \lceil B \rceil$	$\lceil B \rceil \otimes \lceil A \rceil$	$\lceil A \rceil \otimes \lceil B \rceil^\perp$	$\lceil A \rceil^\perp \otimes \lceil B \rceil^\perp$	$\lceil B \rceil^\perp \otimes \lceil A \rceil^\perp$
$+$	$-$	$\lceil A \rceil \otimes \lceil B \rceil^\perp$	$\lceil A \rceil^\perp \otimes \lceil B \rceil^\perp$	$\lceil B \rceil^\perp \otimes \lceil A \rceil^\perp$	$\lceil A \rceil^\perp \otimes \lceil B \rceil$	$\lceil A \rceil \otimes \lceil B \rceil$	$\lceil B \rceil \otimes \lceil A \rceil$
$+$	$+$	$\lceil A \rceil \otimes \lceil B \rceil$	$\lceil A \rceil^\perp \otimes \lceil B \rceil$	$\lceil B \rceil \otimes \lceil A \rceil^\perp$	$\lceil A \rceil^\perp \otimes \lceil B \rceil^\perp$	$\lceil A \rceil \otimes \lceil B \rceil^\perp$	$\lceil B \rceil^\perp \otimes \lceil A \rceil$

$$\mu\alpha.(\frac{x'\ z}{\mathsf{subj}}.\langle\ \mathsf{det} \upharpoonright (\widetilde{\mu}y.\langle\ x' \upharpoonright (\widetilde{\mu}x.\langle\ \mathsf{tv} \upharpoonright ((x \setminus \alpha) / y)\ \rangle / z)\ \rangle / \mathsf{noun})\ \rangle) \tag{15}$$

CPS translation Let us turn then to the translation that associates the proofs of the multiple-conclusion source logic **fLG** with a constructive interpretation, i.e. a linear lambda term of the target logic MILL/**LP**. CPS translations for **LG** were introduced in Bernardi and Moortgat (2007; 2010), who adapt the call-by-value and call-by-name regimes of Curien and Herbelin (2000) to a directional environment. The translation of Bastenhof (2011) (following Girard (1991)) is an improvement in that it avoids the 'administrative redexes' of the earlier approaches: the image of **LG** source derivations, under the mapping from Bastenhof (2011) that we present below, are *normal* **LP** terms.

The target language, on the type level, has the same atoms as the source language, and in addition a distinguished atom $\perp$, the response type. Complex types are linear products $- \otimes -$ and a defined negation $A^\perp \doteq A \multimap \perp$. The CPS translation $\lceil \cdot \rceil$ maps **fLG** source types, sequents and their proof terms to the target types and terms in Curry-Howard correspondence with normal natural deduction proofs.

Types For positive atoms, $\lceil p \rceil = p$, for negative atoms $\lceil p \rceil = p^\perp$. For complex types, the value of $\lceil \cdot \rceil$ depends on the polarities of the subtypes as shown in Table 1.

Terms The action of $\lceil\cdot\rceil$ on terms is given in (16). We write $\widetilde{x}, \widetilde{\alpha}$ for the target variables corresponding to source x, α. The (de)focusing rules correspond to application/abstraction in the target language. Non-invertible (two premise) rules are mapped to linear pair terms; invertible rewrite rules to the matching deconstructor, the **case** construct (φ, ψ, ξ metavariables for the the (co)variables involved).

$$\begin{array}{rcc} \textit{(co)var} & \lceil x\rceil = \widetilde{x} \quad ; & \lceil\alpha\rceil = \widetilde{\alpha} \\ \textit{linear application} & \lceil\langle x \upharpoonright E\rangle\rceil = (\widetilde{x}\,\lceil E\rceil) \quad ; & \lceil\langle V \upharpoonright \alpha\rangle\rceil = (\widetilde{\alpha}\,\lceil V\rceil) \\ \textit{linear abstraction} & \lceil\widetilde{\mu}x.C\rceil = \lambda\widetilde{x}.\lceil C\rceil \quad ; & \lceil\mu\alpha.C\rceil = \lambda\widetilde{\alpha}.\lceil C\rceil \\ \textit{linear pair} & \lceil\varphi\#\psi\rceil = \langle\lceil\varphi\rceil, \lceil\psi\rceil\rangle \quad (\# \in \{\otimes, /, \backslash, \oplus, \oslash, \obslash\}) & \\ \textit{case} & \lceil\frac{\varphi\ \psi}{\xi}.C\rceil = \textbf{case}\ \widetilde{\xi}\ \textbf{of}\ \langle\widetilde{\varphi}, \widetilde{\psi}\rangle.\lceil C\rceil & \end{array} \tag{16}$$

Sequents For sequent hypotheses/conclusions, we have

$\mathsf{pol}(A)$	$\lceil x : A\rceil$	$\lceil\alpha : A\rceil$
$+$	$\widetilde{x} : \lceil A\rceil$	$\widetilde{\alpha} : \lceil A\rceil^{\perp}$
$-$	$\widetilde{x} : \lceil A\rceil^{\perp}$	$\widetilde{\alpha} : \lceil A\rceil$

(17)

Table 1 then specifies how the translation extends to sequents (replace logical connectives by their structural counterparts, and target $\otimes$ by the comma for multiset union).

$$\begin{array}{rcl} \lceil C : (X \vdash Y)\rceil & = & \lceil X\rceil, \lceil Y\rceil \vdash_{\mathbf{LP}} \lceil C\rceil : \perp \\ \left\lceil X \vdash \boxed{v : A}\right\rceil & = & \lceil X\rceil \vdash_{\mathbf{LP}} \lceil v\rceil : \lceil A\rceil \\ \left\lceil \boxed{e : A} \vdash Y\right\rceil & = & \lceil Y\rceil \vdash_{\mathbf{LP}} \lceil e\rceil : \lceil A\rceil^{\perp} \end{array} \tag{18}$$

Illustrations We return to our sample derivations. In (19) one finds the CPS image of the source types for transitive verb and determiner under the different assignments of bias to the atomic subformulas, and the composition with $\cdot^{\ell}$, assuming $np^{\ell} = e$ (entities), $s^{\ell} = \perp^{\ell} = t$ (truth values) and $n^{\ell} = e \to t$ (sets of

Table 2: Constants: lexical translations

$(np^+\backslash s^-)/np^+$	finds	$\lambda\langle\langle x, c\rangle, y\rangle.(c\ (\textsc{find}^{eet}\ y\ x))$
$(np^+/n^+)\otimes n^+$	everyone	$\langle\lambda\langle x, y\rangle.(\forall\ \lambda z.(\Rightarrow\ (y\ z)\ (x\ z))), \textsc{person}^{et}\rangle$
np^+/n^+	some	$\lambda\langle x, y\rangle.(\exists\ \lambda z.(\wedge\ (y\ z)\ (x\ z)))$
n^+	unicorn	$\textsc{unicorn}^{et}$
$(np^-\backslash s^-)/np^-$	needs	$\lambda\langle\langle q, c\rangle, q'\rangle.(q\ \lambda x.(\textsc{need}^{((et)t)et}\ q'\ x))$
$(np^-/n^-)\otimes n^-$	everyone	$\langle\lambda\langle x, w\rangle.(\forall\ \lambda z.(\Rightarrow\ (w\ \lambda y.(y\ z))\ (x\ z))), \lambda k.(k\ \textsc{person}^{et})\rangle$
np^-/n^-	some	$\lambda\langle x, w\rangle.(\exists\ \lambda z.(\wedge\ (w\ \lambda y.(y\ z))\ (x\ z)))$
n^-	unicorn	$\lambda k.(k\ \textsc{unicorn}^{et})$

entities). For the lexical constants of the illustration, Table 2 gives $\cdot^\ell$ translations compatible with the typing. In Table 3, these lexical recipes are substituted for the parameters of the CPS translation.

	LG	$\lceil\cdot\rceil^\perp$	$(\lceil\cdot\rceil^\perp)^\ell$
a.	$(np^+\backslash s^-)/np^+$	$((np\otimes s^\perp)\otimes np)^\perp$	$((e\times(tt))\times e)\to t$
b.	np^+/n^+	$(np^\perp\otimes n)^\perp$	$((et)\times(et))\to t$
c.	$(np^-\backslash s^-)/np^-$	$((np^{\perp\perp}\otimes s^\perp)\otimes np^{\perp\perp})^\perp$	$((((et)t)\times(tt))\times((et)t))\to t$
d.	np^-/n^-	$(np^\perp\otimes n^{\perp\perp})^\perp$	$((et)\times(((et)t)t))\to t$

(19)

Table 3: Compositional translations

$\lceil(13)\rceil = \lambda\widetilde{\beta}.(\mathsf{case\ subj}^\ell\ \mathsf{of}\ \langle\widetilde{y},\widetilde{z}\rangle.(\mathsf{tv}^\ell\ \langle\langle\lambda\widetilde{\gamma}.(\widetilde{y}\ \langle\widetilde{\gamma}, \lambda\widetilde{\gamma}'.(\widetilde{z}\ \widetilde{\gamma}')\rangle), \widetilde{\beta}\rangle, \lambda\widetilde{\alpha}.(\mathsf{det}^\ell\ \langle\widetilde{\alpha}, \lambda\widetilde{\alpha}'.(\mathsf{noun}^\ell\ \widetilde{\alpha}')\rangle)\rangle)))$

$\lceil(13)\rceil^\ell = \lambda c.(\forall\ \lambda x.((\Rightarrow\ (\mathsf{person}\ x))\ (c\ ((\textsc{needs}\ \lambda w.(\exists\ \lambda y.((\wedge\ (\mathsf{unicorn}\ y))\ (w\ y))))\ x))))$

$\lceil(14)\rceil = \lambda\widetilde{\alpha}.(\mathsf{case\ subj}^\ell\ \mathsf{of}\ \langle\widetilde{x}',\widetilde{z}\rangle.(\widetilde{x}'\ \langle\lambda\widetilde{x}.(\mathsf{det}^\ell\ \langle\lambda\widetilde{y}.(\mathsf{tv}^\ell\ \langle\langle\widetilde{x},\widetilde{\alpha}\rangle,\widetilde{y}\rangle), \mathsf{noun}^\ell\rangle), \widetilde{z}\rangle))$

$\lceil(14)\rceil^\ell = \lambda c.(\forall\ \lambda x.((\Rightarrow\ (\textsc{person}\ x))\ (\exists\ \lambda y.((\wedge\ (\textsc{unicorn}\ y))\ (c\ ((\textsc{likes}\ y)\ x))))))$

$\lceil(15)\rceil = \lambda\widetilde{\alpha}.(\mathsf{case\ subj}^\ell\ \mathsf{of}\ \langle\widetilde{x}',\widetilde{z}\rangle.(\mathsf{det}^\ell\ \langle\lambda\widetilde{y}.(\widetilde{x}'\ \langle\lambda\widetilde{x}.(\mathsf{tv}^\ell\ \langle\langle\widetilde{x},\widetilde{\alpha}\rangle,\widetilde{y}\rangle), \widetilde{z}\rangle), \mathsf{noun}^\ell\rangle))$

$\lceil(15)\rceil^\ell = \lambda c.(\exists\ \lambda y.((\wedge\ (\textsc{unicorn}\ y))\ (\forall\ \lambda x.((\Rightarrow\ (\textsc{person}\ x))\ (c\ ((\textsc{likes}\ y)\ x))))))$

3.2 Proof nets and focusing

We saw in §3.1 that **fLG** may allow multiple derivations from one and the same set of (co)axiom judgements. These derivations would be identified under the proof net perspective of §2.2. To establish the correspondence with **fLG** derivations, we introduce term-labeled proof nets, and show how a proof term can be read off from the *composition graph* associated with a net.

Our approach is comparable to the algorithm of de Groote and Retoré (1996), which computes a linear lambda term from a traversal of the dynamic graph associated with a proof net for a derivation in the Lambek calculus **L**. For single-conclusion **L**, the term associated with a given proof net is unique; in the case of multiple-conclusion **LG** the term computation algorithm may associate more than one term with a proof net. These multiple results will then be shown to correspond to the derivational ambiguity of focused proof search.

Reduction tree In order to analyse the structure of a conversion sequence in more detail, we introduce the notion of a proof net *component*:

Definition 3.1. Given a proof net P, a *component* C of P is a maximal subnet of P containing only tensor links.

From a proof net, we can obtain its components by simply erasing all cotensor links. The components will be the connected components (in the graph-theoretic sense) of the resulting graph. To simplify the following discussion, unless otherwise indicated, we will use the word *component* to refer only to components containing at least one tensor link.

When P is a proof net (and therefore converts to a tensor tree using a sequence ρ of conversions and contractions) the components of P can bee seen as a parallel representation of the synchronous phases in sequent proof search. In ρ, all interaction rules operate in one component C, the cotensor rules and the corresponding contractions join two different components (though the component connected to the main vertex can be trivial here). When multiple cotensor links have both active tentacles attached to a single component (Figure 10 shows an example), we apply all contractions simultaneously, repeating this process until no further contractions apply.

So instead of seeing ρ as a *sequence* of reductions, we can see it as a rooted *tree* of reductions: the initial components are its leaves (synchronous phases) and

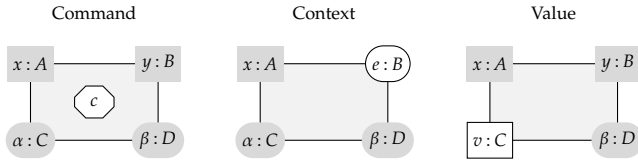


Figure 13: Proof nets with term labels: commands, context and values

the contractions, which join components, are its branches (the branches from the active components to their parents correspond to asynchronous phases) and the final tree — a single component — is its root (we will see an example in Figure 19 below). Note that this same observation is essential to the cut elimination proof of Moot and Puite (2002).

Nets and term labeling When assigning a term label to a proof net, our algorithms will assign labels to larger and larger subnets of a given proof net, until we have computed a term for the complete proof net. Like in the sequent calculus, we distinguish between subnets which are commands, contexts and values. Figure 13 shows how we will distinguish these visually: the main formula of a subnet is drawn white, other formulas are drawn in light gray, values are drawn inside a rectangle, contexts inside an oval.

Figure 14 gives the term-labeled version of the proof net links corresponding to the logical rules of the sequent calculus. The flow of information is shown by the arrows: information flow is always from the active formulas to the main formula of a link, and as a consequence the complex term can be assigned either to a conclusion or to a premise of the link. This is the crucial difference with term labeling for the single-conclusion Lambek calculus, where the complex term is always assigned to a conclusion. The cotensor rules, operating on commands, indicate the prefix for the command corresponding to the term assignment for the rule (we will see later how commands are formed).

The proof term of an **LG** derivation is computed on the basis of the *composition graph* associated with its proof net.

Definition 3.2. Given a proof net P, the associated *composition graph* $cg(P)$ is obtained as follows.

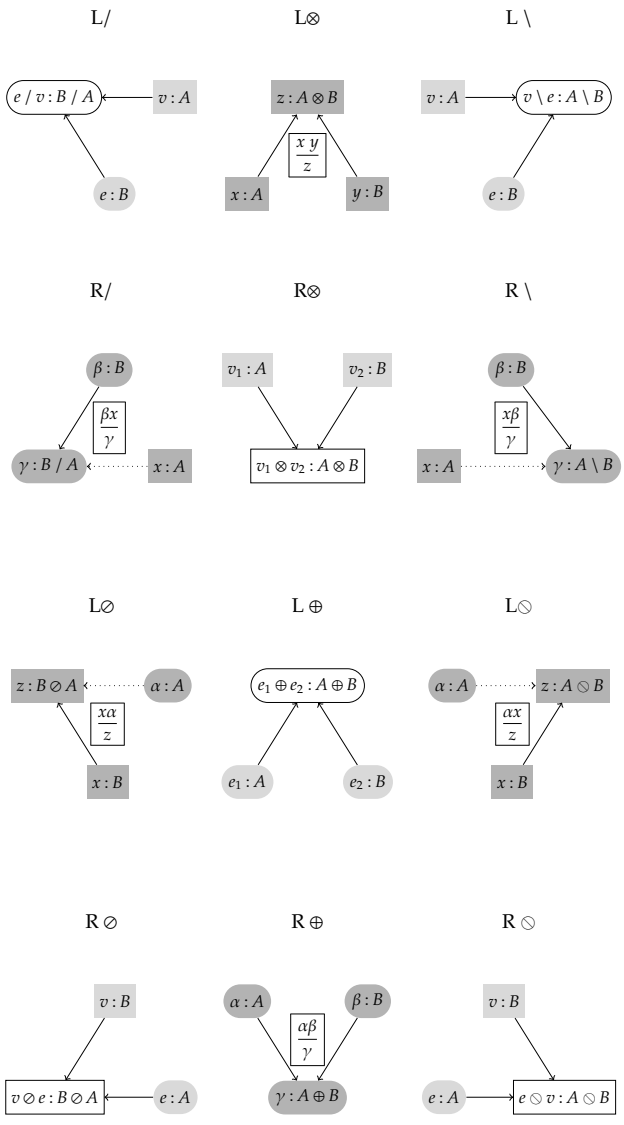


Figure 14: **LG** links with term labeling

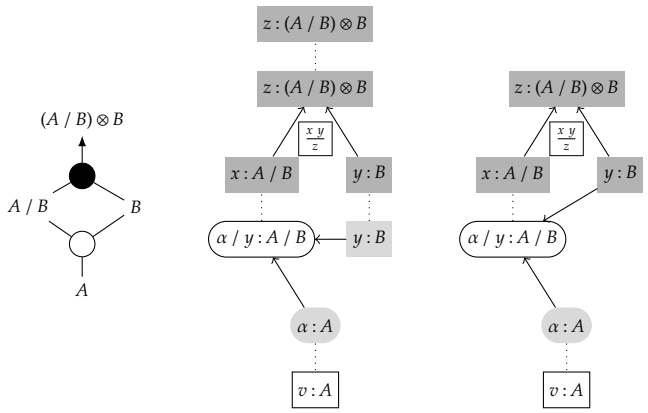


Figure 15: Proof net, initial composition graph, reduced composition graph.

1. all vertices of P with formula label A are expanded into *axiom links*: edges connecting two vertices with formula label A; all links are replaced by the corresponding links of Figure 14;
2. all vertices in this new structure are assigned atomic terms of the correct type (variable or covariable) and the terms for the tensor rules are propagated from the active formulas to the main formula;
3. all axiom links connecting terms of the same type (value or context) are collapsed.

Figure 15 gives an example of the composition graph associated with a net. In all, the expansion stage gives rise to four types of axiom links, depending on the type of the term assigned to the A premise and the A conclusion. These cases are summarized in Figure 16. The substitution links are collapsed in the final stage of the construction of the composition graph (shown on the right of Figure 15; the command and $\mu/\widetilde{\mu}$ cases are the ones that remain.

Given the composition graph $cg(P)$ associated with a proof net P, we compute terms for it as follows.

1. we compute all maximal subnets of $cg(P)$, which consist of a set of tensor links with a single main formula, marking all these links as visited;

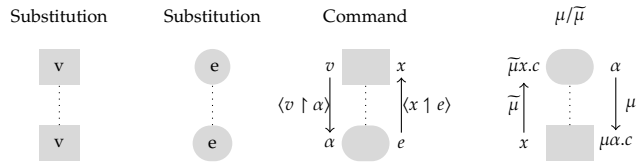


Figure 16: Types of axiom links

2. while $cg(P)$ contains unvisited links do the following:
   (a) follow an unvisited command link attached to a previously calculated maximal subnet, forming a correct command subnet; like before, we restrict to *active* subnets which do not contain (or allow us to reach through an axiom) the main formula of a negative link;
   (b) for each negative link with both active formulas attached to the current command subnet, pass to the main formula of the negative link, forming a new command, repeat this step until no such negative links remain attached;
   (c) follow a μ or $\widetilde{\mu}$ link to a new vertex, forming a larger value or context subnet and replacing the variable previously assigned to the newly visited vertex by the μ value or $\widetilde{\mu}$ context.

The algorithm stays quite close to the focused derivations of the previous section: the maximal subnets of step 1 are *rooted* versions of the components we have used before, with the directions of the arrows potentially splitting components into multiple rooted components (Figure 18 will give an example) and the asynchronous phases, which consisted of one or more contractions for cotensor links, will now consist of a passage through a command link, followed by zero or more cotensor links, followed by either a μ or a $\widetilde{\mu}$ link, the result being a new, larger subnet. The term assignment algorithm is a way to enumerate the non-equivalent proof terms of a net. Given that these terms are isomorphic to focused sequent proofs, it is no coincidence that the computation of the proof terms looks a lot like the sequentialisation algorithm.[4]

Lemma 1. *If P is a proof net (with a pairing of command and $\mu/\widetilde{\mu}$ links) and v is a term calculated for P using this pairing then there is a sequent proof π which is assigned v*

[4]The connection between proof net sequentialisation and focusing for linear logic is explored in Andreoli and Maieli (1999)

as well.

This lemma is easily proved by induction on the depth of the tree: it holds trivially for the leaves (which are rooted components), and, inductively, each command, cotensor, $\mu/\widetilde{\mu}$ sequence will produce a sequent proof of the same term: in fact each such step corresponds exactly to the derived inference rules for focus shifting discussed in §3.1.

To summarize, the difference between computing terms for proof nets in the Lambek calculus **L** and in **LG** can be characterized as follows:

L: the (potential) terms are given through a bijection between premise and conclusion atomic formulas (ie. a complete matching of the axioms),

LG: the (potential) terms are given through a bijection between premise and conclusion atomic formulas *plus* a bijection between command and $\mu/\widetilde{\mu}$ axioms.

We speak of *potential* terms, since in the case of the Lambek calculus only proof nets can be assigned a term, whereas in the **LG** case we need proof nets plus a coherent bijection between command and $\mu/\widetilde{\mu}$ axioms, where the μ or $\widetilde{\mu}$ rule is applied to one of the free variables of the command c.

Illustrations Figure 17 shows how to compute the term for the example proof net of Figure 15, starting from the composition graph (on the right). We first look for the components (step 1). Since there is only a single tensor link, this is simple. Figure 17 shows, on the left, the context subnet corresponding to this link.

Now, there is only one command to follow from here (step 2a), which produces the command shown in the middle of Figure 17. Applying the cotensor link (step 2b) produces the figure shown on the right. The final μ link (step 2c, not shown) produces the completed term for this proof net.

$$v = \mu\alpha.\frac{x\ y}{z}\langle x \upharpoonright \alpha \ / \ y\rangle$$

Some remarks about this example. First, some of the axioms can be traversed in only one of the two possible directions: in cut-free proof nets, command links move either towards the active formulas of cotensor links or towards

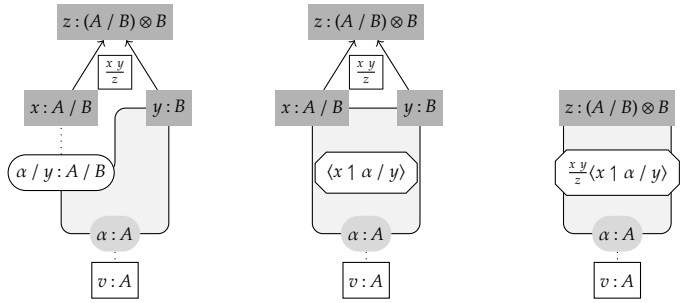


Figure 17: Computing the proof term from a composition graph

"dead ends": hypotheses or conclusions of the proof net. And since we want to compute the value of v for the example proof net, it only makes sense to apply a μ rule to compute this value: we always "exit" the proof net from a designated conclusion. With a slight modification to the algorithm that reads off terms from a composition graph, we could also compute *commands* for proof nets, or compute the *context* for a designated *premise* of the net.

Figure 18 returns to our "subj tv det noun" example. On the left we see the composition graph for the example of Figure 12.

The only cotensor link in the figure has the node subj : $(np \,/\, n) \otimes n$ as its main formula. When we compute the rooted components, we see that there are three, shown on the right of the figure.

There are three command axioms, one for the root node of each of the three components, C_1 to C_3 on the right hand side of the figure; these are numbered $c1$ to $c3$ next to the corresponding links with the same number as the corresponding component. There are also three $\mu/\widetilde{\mu}$ links (numbered $\mu 1$ to $\mu 3$).

Figure 19 gives a schematic representation of the proof net of Figure 18. The arrows next to the $\mu/\widetilde{\mu}$ links indicate the different possibilities for traversing the link and whether this traversal corresponds to a μ or a $\widetilde{\mu}$ link.

If both *np* arguments of the transitive verbs are lexically assigned a positive bias, then we can only pass the two axioms $\mu_2/\widetilde{\mu}_2$ and $\mu_3/\widetilde{\mu}_3$ in the $\widetilde{\mu}_2$ and $\widetilde{\mu}_3$ directions, following the arrows away from component C_2. Simple combinatorics will then give us two possible terms for this proof net: $c_2 - \widetilde{\mu}_3$, $c_3 - \widetilde{\mu}_2$ and

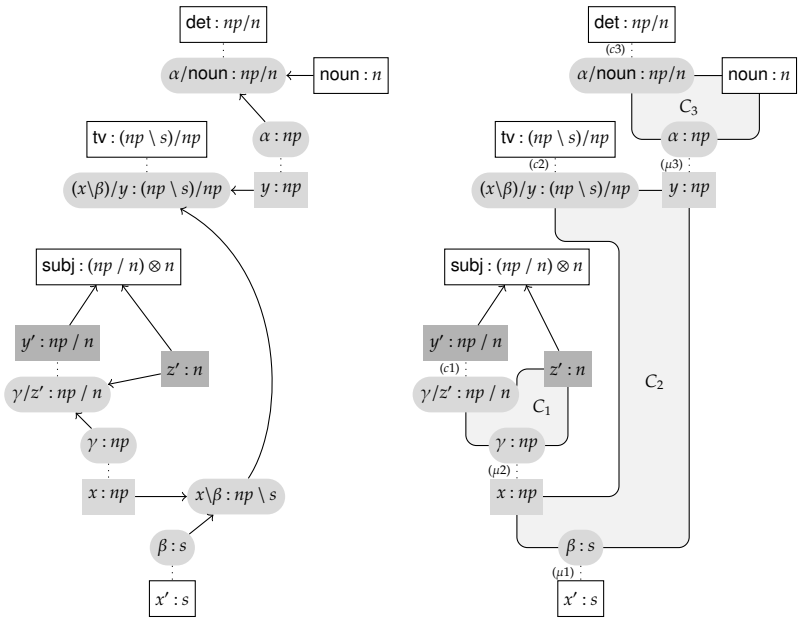


Figure 18: Composition graph (left) and initial components (right) for the "subj tv det noun" example

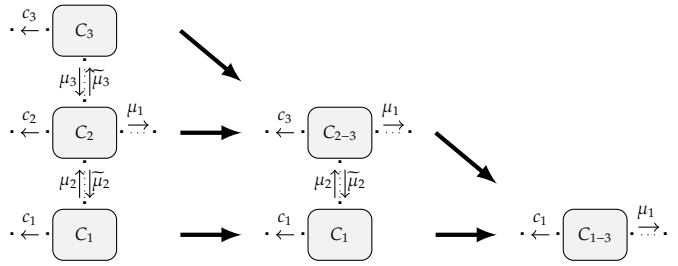


Figure 19: Matching: $c_2 - \widetilde{\mu}_3$, $c_3 - \widetilde{\mu}_2$, $c_2 - \mu_1$. Reading: subj < det < tv.

$c_1 - \mu_1$ (shown in Figure 19), producing term 20 below, and $c_2 - \widetilde{\mu}_2$, $c_1 - \widetilde{\mu}_3$ and $c_3 - \mu_1$, producing term 21.

$$\mu\gamma.\frac{y'z'}{\mathsf{subj}}.\langle y' \restriction (\widetilde{\mu}x.\langle \mathsf{det} \restriction (\widetilde{\mu}y.\langle \mathsf{tv} \restriction (x\backslash\beta)/y\rangle)/\mathsf{noun}\rangle)/z'\rangle \tag{20}$$

$$\mu\beta.\langle \mathsf{det} \restriction (\widetilde{\mu}y.\frac{y'z'}{\mathsf{subj}}.\langle y' \restriction (\widetilde{\mu}x.\langle \mathsf{tv} \restriction (x\backslash\beta)/y\rangle)/z'\rangle)/\mathsf{noun}\rangle \tag{21}$$

These are the only two readings available with positive bias for the two atomic *np* arguments of the transitive verb, and, as we have seen before, this gives the right quantifier scope possibilities for an extensional transitive verb such as "likes" we have seen in equations (14) and (15) (apart from the variable names, equation (21) differs from (15) in that the extended command fraction in the latter term is at the innermost position, but the terms are equivalent up to commutative conversions).

When we use a negative bias for the two *np* arguments of the transitive verb, we obtain the following term, corresponding to equation (13).

$$\mu\beta.\frac{y'z'}{\mathsf{subj}}.\langle y' \restriction (\widetilde{\mu}x.\langle \mathsf{tv} \restriction (x\backslash\beta)/(\mu\alpha.\langle \mathsf{det} \restriction \alpha/\mathsf{noun}\rangle))\rangle)/z'\rangle \tag{22}$$

4 Conclusions

The Lambek-Grishin calculus is a symmetric version of the Lambek calculus. Together with the interaction principles, it allows for the treatment of patterns beyond context-free which cannot be satisfactorily handled in the Lambek calculus. We have compared two proof systems for **LG**: focused sequent proofs and proof nets. Focused proofs avoid the spurious non-determinism of backward-chaining search in the sequent calculus; they provide a natural interface to semantic interpretation via their continuation-passing-style translation. Proof nets present the essence of a derivation in a visually appealing form; they do away with the syntactic clutter of sequent proofs, and compute the structure of the end-sequent in a data-driven manner where this structure has to be given before one can a start backward-chaining sequent derivation. Proof terms are read off from the composition graph associated with a net. The computation of these terms depends both on a bijection between premise and conclusion atomic formulas and between command and $\mu/\widetilde{\mu}$ axioms. As a re-

sult, one net can be associated with multiple construction recipes (proof terms), corresponding to multiple derivations in the focused sequent calculus.

Acknowledgements An extended version of this paper appears as a chapter in C. Heunen, M. Sadrzadeh and E. Grefenstette (eds.) *Compositional methods in quantum physics and linguistics*, OUP. We thank Arno Bastenhof for comments on an earlier draft.

References

J.-M. Andreoli. Focussing and proof construction. *Annals of Pure and Applied Logic*, 107(1-3):131–163, 2001.

J.-M. Andreoli and R. Maieli. Focusing and proof-nets in linear and non-commutative logic. In *International Conference on Logic for Programming and Automated Reasoning (LPAR)*, volume 1581 of *LNAI*. Springer, 1999.

A. Bastenhof. Polarized Montagovian semantics for the Lambek-Grishin calculus. *CoRR*, abs/1101.5757, 2011. To appear in the Proceedings of the 15th Conference on Formal Grammar (Copenhagen, 2010), Springer LNCS.

A. Bastenhof. Polarities in logic and semantics. In M. Aloni, V. Kimmelman, F. Roelofsen, K. Schulz, G. W. Sassoon, and M. Westera, editors, *Logic Language and Meaning*, volume 7218 of *LNCS*, pages 230–239. Springer, 2012. 18th Amsterdam Colloquium, December 2011. Revised selected papers.

R. Bernardi and M. Moortgat. Continuation semantics for symmetric categorial grammar. In D. Leivant and R. de Queiros, editors, *Proceedings 14th Workshop on Logic, Language, Information and Computation (WoLLIC'07)*, LNCS 4576. Springer, 2007.

R. Bernardi and M. Moortgat. Continuation semantics for the Lambek-Grishin calculus. *Information and Computation*, 208(5):397–416, 2010.

J. Cockett and R. Seely. Proof theory for full intuitionistic linear logic, bilinear logic and mix categories. In *Theory and Applications of Categories 3*, pages 85–131, 1996.

P. Curien and H. Herbelin. Duality of computation. In *International Conference on Functional Programming (ICFP'00)*, pages 233–243, 2000.

P. de Groote and C. Retoré. Semantic readings of proof nets. In G.-J. Kruijff, G. Morrill, and D. Oehrle, editors, *Formal Grammar*, pages 57–70, Prague, 1996. FoLLI.

J.-Y. Girard. Linear logic. *Theoretical Computer Science*, 50:1–102, 1987.

J.-Y. Girard. A new constructive logic: classical logic. *Mathematical Structures in Computer Science*, 1(3):255–296, 1991.

R. Goré. Substructural logics on display. *Logic Journal of IGPL*, 6(3):451–504, 1997.

V. Grishin. On a generalization of the Ajdukiewicz-Lambek system. In A. Mikhailov, editor, *Studies in Nonclassical Logics and Formal Systems*, pages 315–334. Nauka, Moscow, 1983. [English translation in Abrusci and Casadio (eds.) Proceedings 5th Roma Workshop, Bulzoni Editore, Roma, 2002].

L. Kallmeyer. *Parsing Beyond Context-Free Grammars*. Cognitive Technologies. Springer, 2010. ISBN 978-3-642-14845-3.

M. Melissen. The generative capacity of the Lambek-Grishin calculus: A new lower bound. In P. de Groote, M. Egg, and L. Kallmeyer, editors, *Proceedings 14th Conference on Formal Grammar*, volume 5591 of *Lecture Notes in Computer Science*, pages 118–132. Springer, 2010.

M. Moortgat. Symmetric categorial grammar. *Journal of Philosophical Logic*, 38 (6):681–710, 2009.

R. Moot. Proof nets for display logic. Technical report, CNRS and INRIA Futurs, 2007.

R. Moot and Q. Puite. Proof nets for the multimodal Lambek calculus. *Studia Logica*, 71(3):415–442, 2002.

A Dynamic Analysis of Interactive Rationality

Eric Pacuit and Olivier Roy

University of Maryland and Tilburg University, Munich Center of Mathematical Philosophy
e.j.pacuit@uvt.nl, olivier.roy@lmu.de

Abstract

Epistemic game theory has shown the importance of informational contexts in understanding strategic interaction. We propose a general framework to analyze how such contexts may arise. The idea is to view informational contexts as the fixed-points of iterated, "rational responses" to incoming information about the agents' possible choices. We show general conditions for the stabilization of such sequences of rational responses, in terms of structural properties of both the decision rule and the information update policy.

1 Background and Motivation

An increasingly popular[1] view is that "*the* fundamental insight of game theory [is] that a rational player must take into account that the players reason about

[1] But, of course, not uncontroversial. See, for example, (Kadane and Larkey 1982, pg. 239).

each other in deciding how to play" (Aumann and Dreze 2008, pg. 81). Exactly *how* the players (should) incorporate the fact that they are interacting with other (actively reasoning) agents into their own decision making process is the subject of much debate. A variety of frameworks explicitly model the *reasoning* of rational agents in a strategic situation. Key examples include Brian Skyrms' models of "dynamic deliberation" Skyrms (1990), Ken Binmore's analysis of "eductive reasoning" Binmore (1987), and Robin Cubitt and Robert Sugden's "common modes of reasoning" Cubitt and Sugden (2011a). Although the details of these frameworks are quite different they share a common line of thought: In contrast to classical game theory, *solution concepts* are no longer the basic object of study. Instead, the "rational solutions" of a game are the result of individual (rational) decisions in specific informational "contexts".

This perspective on the foundations of game theory is best exemplified by the so-called epistemic program in game theory (cf. Brandenburger (2007)). The central thesis here is that the basic mathematical model of a game should include an explicit parameter describing the players' *informational attitudes*. However, this broadly decision-theoretic stance does not simply *reduce* the question of decision-making in interaction to that of rational decision making in the face of uncertainty or ignorance. Crucially, *higher-order* information (belief about beliefs, etc.) are key components of the informational context of a game[2]. Of course, different contexts of a game can lead to drastically different outcomes, but this means that the informational contexts themselves are open to rational criticism:

> "It is important to understand that we have two forms of irrationality [...]. For us, a player is rational if he optimizes and also rules nothing out. So irrationality might mean not optimizing. But it can also mean optimizing while not considering everything possible." (Brandenburger et al. 2008, pg. 314)

[2]That is, strategic behavior *depends*, in part, on the players' higher-order beliefs. However, the question of what precisely is being claimed should be treated with some care. The well-known *email game* of Ariel Rubinstein Rubinstein (1989) demonstrates that misspecification of arbitrarily high-orders of beliefs can have a great impact on (predicted) strategic behavior. So there are simple examples where (predicted) strategic behavior is *too sensitive* to the players' higher-order beliefs. We are not claiming that a rational agent is *required* to consider *all* higher-order beliefs, but only that a rational player recognizes that her opponents are actively reasoning, rational agents, which means that a rational player does take into account *some* of her higher-order beliefs (e.g., what she believes her opponents believe she will do) as she deliberates. Precisely "how much" higher-order information should be taken into account is a very interesting, open question which we set aside in this paper.

Thus, a player can be rationally criticized for not choosing what is *best given their information*, but also for not reasoning *to* a "proper" context. Of course, what counts as a "proper" context is debatable. There might be rational pressure for or against making certain *substantive assumptions*[3] about the beliefs of one's opponents, for instance, always entertaining the possibility that one of the players might not choose optimally.

Recently, researchers using methods from dynamic-epistemic logic have taken steps to understanding this idea of reasoning *to* a "proper" or "rational" context van Benthem (2007), Baltag et al. (2009), Baltag and Smets (2009a), van Benthem and Gheerbrant (2010). Building on this literature[4], we provide a general characterization of when players can or cannot rationally reason to an informational context.

2 Belief Dynamics for Strategic Games

Our goal is to understand well-known solutions concepts, not in terms of fixed informational contexts—for instance, models (e.g., type spaces or epistemic models) satisfying rationality and common belief of rationality—but rather as a result of a dynamic, interactive process of "information exchanges". It is important to note that we do *not* see this work as an attempt to represent some type of "pre-play communication" or form of "cheap talk". Instead, the idea is to represent the process of *rational deliberation* that takes the players from the *ex ante* stage to the *ex interim* stage of decision making. Thus, the "informational exchanges" are the result of the players' *practical reasoning* about what they should do, given their current beliefs. This is in line with the current research program using dynamic epistemic and doxastic logics to analyze well-known solution concepts (cf. Apt and Zvesper (2010a), Baltag et al. (2009), van Benthem (2007) where the "rationality announcements" do not capture any type of communication between the players, but rather internal observations about which outcomes of the game are "rational").

[3] The notion of substantive assumption is explored in more detail in Roy and Pacuit (2010).

[4] The reader not familiar with this area can consult the recent textbook van Benthem (2010) for details.

2.1 Describing an Informational Context

Let $G = \langle N, \{S_i\}_{i\in N}, u_i\rangle$ be a strategic game (where N is the set of players and for each $i \in N$, S_i is the set of actions for player i and $u_i : \Pi_i S_i \to \mathbb{R}$ is a utility function).[5] The informational context of a game describes the players' *hard* and *soft* information about the possible outcomes of the game. Many different formal models have been used to represent an informational context of a game (for a sample of the extensive literature, see Bonanno and Battigalli (1999), van Benthem (2007) and references therein). In this paper we employ one such model: a *plausibility structure* consisting of a set of states and a single plausibility ordering (which is reflexive, transitive and connected) $w \preceq v$ that says "v is at least as plausible as w." Originally used as a semantics for conditionals (cf. Lewis (1973)), these *plausibility models* have been extensively used by logicians (van Benthem 2004; 2010, Baltag and Smets 2009a), game theorists (Board 2004) and computer scientists (Boutilier 1992, Lamarre and Shoham 1994) to represent rational agents' (all-out) beliefs. We thus take for granted that they provide a natural model of beliefs in games:

Definition 2.1. Let $G = \langle N, \{S_i\}_{i\in N}, u_i\rangle$ be a strategic form game. An **informational context** of G is a plausibility model $\mathcal{M}_G = \langle W, \preceq, \sigma\rangle$ where $\preceq$ is a connected, reflexive, transitive and well-founded[6] relation on W and σ is a **strategy function**: a function $\sigma : W \to \Pi_i S_i$ assigning strategy profiles to each state. To simplify notation, we write $\sigma_i(w)$ for $(\sigma(w))_i$ (similarly, write $\sigma_{-i}(w)$ for the sequence of strategies of all players except i).

A few comments about this definition are in order. First of all, note that there is only one plausibility ordering in the above models, yet we are interested in games with more than one player. There are different ways to interpret the fact that there is only one plausibility ordering. One is that the models represent the beliefs of a single player before she has made up her mind about which option to choose in the game. A second interpretation is to think of a model as representing the modeler's or game theorist's point of view about which outcomes are more or less plausible given the reasoning of the players. Thus,

[5] We assume the reader is familiar with the basic concepts of game theory. For example, strategic games and various solution concepts, such as iterated removal of strictly (weakly) dominated strategies.

[6] Well-foundedness is only needed to ensure that, for any set X, the set of minimal elements in X is nonempty. This is important only when W is infinite – and there are ways around this in current logics. Moreover, the condition of connectedness can also be lifted, but we use it here for convenience.

a model describes a stage of the rational deliberation of *all* the players starting from an initial model where the players have the same beliefs (i.e., the *common prior*). The private information about which outcomes the *players* consider possible given their actual choice can then be defined from the *conditional beliefs*.[7] Our second comment on the above definition is that since we are representing the rational deliberation process, we do not assume that the players have made up their minds about which actions they will choose. Finally, note that the strategy functions need not be onto. Thus, the model represents the player's(s') opinions about which outcomes of the game are more or less plausible *among the ones that have not been ruled out*.

Of course, this model can be (and has been: see Baltag and Smets (2009a), van Benthem (2010)) extended to include beliefs for each of the players, an explicit relation representing the player(s) hard information or by making the plausibility orders state-dependent. In order to keep things simple we focus on models with a single plausibility ordering.

We conclude this brief introduction to plausibility models by giving the well-known definition of a conditional belief. For $X \subseteq W$, let $Min_{\preceq}(X) = \{v \in X \mid v \preceq w \text{ for all } w \in X\}$ be the set of minimal elements of X according to $\preceq$.

Definition 2.2 (Belief and Conditional Belief)**.** Let $\mathcal{M}_G = \langle W, \preceq, \sigma \rangle$ be a model of a game G. Let E and F be subsets of W, we say:

- E is **believed conditional on** F in $\mathcal{M}_G$ provided $Min_{\preceq}(F) \subseteq E$.

Also, we say E is **believed** in $\mathcal{M}_G$ if E is believed conditional on W. Thus, E is believed provided $Min_{\preceq}(W) \subseteq E$

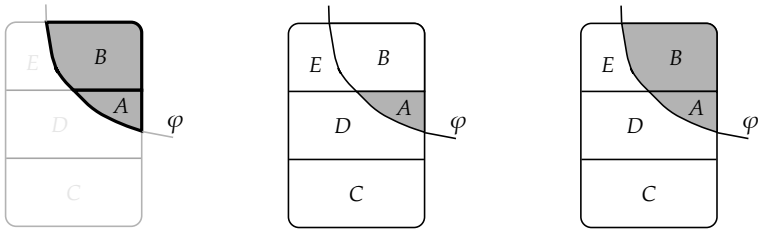
2.2 A Primer on Belief Dynamics

We are not interested in informational contexts *per se*, but rather how the informational context changes during the process of rational deliberation. The type of change we are interested in is how a model $\mathcal{M}_G$ of a game G incorporates new information about what the players *should* do (according to a particular choice rule). As is well known from the belief revision literature, there are

[7] The suggestion here is that one can define a partition model á la Aumann Aumann (1999) from a plausibility model. Working out the details is left for future work, but we note that such a construction blurs the distinction between so-called *belief*-based and *knowledge*-based analyses of solution concepts (cf. the discussion in Brandenburger (2007)).

many ways to transform a plausibility model given some new information Rott (2006). We do not have the space to survey the entire body of relevant literature here (cf., van Benthem (2010), Baltag and Smets (2009b)). Instead we sketch some key ideas, assuming the reader is already familiar with this approach to belief revision.

The general approach is to define a way of *transforming* a plausibility model $\mathcal{M}_G$ given a proposition φ. A transformation τ maps plausibility models and propositions to plausibility models (we write $\mathcal{M}_G^{\tau(\varphi)}$ for $\tau(\mathcal{M}_G, \varphi)$). Different definitions of τ represent the different attitudes an agent can take to the incoming information. The picture below provides three typical examples:



$!(\varphi) : A \prec B$ $\quad\uparrow(\varphi) : A \prec C \prec D \prec B \cup E$ $\quad\Uparrow(\varphi) : A \prec B \prec C \prec D \prec E$

The operation on the left is the well-known *public announcement* operation Plaza (1989), Gerbrandy (1999), which assumes that the source of φ is *infallible*, ruling out any possibilities that are inconsistent with φ. For the other transformations, while the players do *trust* the source of φ, they do not treat the source as infallible. Perhaps the most ubiquitous policy is *conservative upgrade* ($\uparrow\varphi$), which allows the player(s) only tentatively to accept the incoming information φ by making the best φ-worlds the new minimal set while keeping the old plausibility ordering the same on all other worlds. The operation on the right, *radical upgrade* ($\Uparrow\varphi$), is stronger, moving *all* φ worlds before all the $\neg\varphi$ worlds and otherwise keeping the plausibility ordering the same. These dynamic operations satisfy a number of interesting logical principles van Benthem (2010), Baltag and Smets (2009b), which we do not discuss further here.

We are interested in the operations that transform the informational context as the players deliberate about what they should do in a game situation. In each informational context (viewed as describing one stage of the deliberation process), the players determine which options are "*rationally permissible*" and

which options the players ought to avoid (which is guided by some fixed choice rule). This leads to a transformation of the informational context as the players adopt the relevant beliefs about the outcome of their *practical reasoning*. The different types of transformation mentioned above then represent how confident the player(s) (or modeler) is (are) in the assessment of which outcomes are rational. In this new informational context, the players again think about what they should do, leading to another transformation. The main question is does this process *stabilize*?

The answer to this question will depend on a number of factors. The general picture is

$$\mathcal{M}_0 \stackrel{\tau(D_0)}{\Longrightarrow} \mathcal{M}_1 \stackrel{\tau(D_1)}{\Longrightarrow} \mathcal{M}_2 \stackrel{\tau(D_2)}{\Longrightarrow} \cdots \stackrel{\tau(D_n)}{\Longrightarrow} \mathcal{M}_{n+1} \Longrightarrow \cdots$$

where each D_i is some proposition and τ is a model transformer. Two questions are important for the analysis of this process. First, what type of transformations are the players using? For example, if τ is a public announcement, then it is not hard to see that, for purely logical reasons, this process must eventually stop at a limit model (see Baltag and Smets (2009a) for a discussion and proof). The second question is where do the propositions D_i come from? To see why this matters, consider the situation where you iteratively perform a radical upgrade with p and $\neg p$ (i.e., $\Uparrow(p), \Uparrow(\neg p), \ldots$). Of course, this sequence of upgrades never stabilizes. However, in the context of reasoning about what to do in a game situation, this situation may not arise thanks to special properties of the choice rule that is being used to describe (or guide) the players' decisions.

2.3 Deliberating about What to Do

It is not our intention to have the dynamic operations of belief change discussed in the previous section directly represent the players' (practical) *reasoning*. Instead, we treat practical reasoning as a "black box" and focus on general *choice rules* that are intended to describe rational decision making (under ignorance). To make this precise, we need some notation:

Definition 2.3 (Strategies in Play)**.** Let $G = \langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$ be a strategic game and $\mathcal{M}_G = \langle W, \leq, \sigma\rangle$ an informational context of G. For each $i \in N$, the strategies **in play for** i is the set

$$S_{-i}(\mathcal{M}_G) = \{s_{-i} \in \Pi_{j\neq i} S_j \mid \text{ there is } w \in Min_{\leq}(W) \text{ with } \sigma_{-i}(w) = s_{-i}\}$$

This set $S_{-i}(\mathcal{M}_G)$ is the set of strategies that are believed to be available for player i at some stage of the deliberation process represented by the model $\mathcal{M}_G$. Given $S_{-i}(\mathcal{M}_G)$, different choice rules offer recommendations about which options to choose. There are many choice rules that could be analyzed here (e.g., strict dominance, weak dominance or admissibility, minimax, minmax regret, etc.). For the present purposes we focus primarily on weak dominance (or admissibility), although our main theorem in Seciton 1 applies to all choice rules.

Weak Dominance (pure strategies[8]) Let $G = \langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$ be a strategic game and $\mathcal{M}_G$ an model of G. For each i and $a \in S_i$, put $a \in S_i^{wd}(\mathcal{M}_G)$ provided there is $b \in S_i$ such that for all $s_{-i} \in S_{-i}(\mathcal{M}_G)$, $u_i(s_{-i}, b) \geq u_i(s_{-i}, a)$ and there is some $s_{-i} \in S_{-i}(\mathcal{M}_G)$ such that $u_i(s_{-i}, b) > u_i(s_{-i}, a)$.

So an action a is weakly dominated for player i if it is weakly dominated with respect to all of i's available actions and the (joint) strategies believed to be still in play for i's opponents.

More generally, we assume that given the beliefs about which strategies are in play the players *categorize* their available options (i.e., the set S_i) into "good" (or "rationally permissible") strategies and those strategies that are "bad" (or "irrational"). Formally, a **categorization** for player i is a pair $\mathbf{S}_i(\mathcal{M}_G) = (S_i^+, S_i^-)$ where $S_i^+ \cup S_i^- \subseteq S_i$. (We write $\mathbf{S}_i(\mathcal{M}_G)$ to signal that the categorization depends on current beliefs about which strategies are in play.) Note that, in general, a categorization need not be a partition (i.e., $S_i^+ \cup S_i^- \neq S_i$) . See Cubitt and Sugden (2011b) for an example of such a categorization algorithm. However, in the remainder of this paper we focus on familiar choice rules where the categorization does form a partition. For example, for weak dominance we let $S_i^- = S_i^{wd}(\mathcal{M}_G)$ and $S_i^+ = S_i - S_i^-$.

Given a model of a game $\mathcal{M}_G$ and for each player i a categorization is $\mathbf{S}_i(\mathcal{M}_G)$; the next step is to incorporate this information into $\mathcal{M}_G$ using some model transformation. We start by introducing a simple propositional language to describe a categorization.

Definition 2.4 (Language for a Game)**.** Let $G = \langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$ be a strategic game. Without loss of generality, assume that each of the S_i is disjoint and let

[8]This definition can be modified to allow for dominance by mixed strategies, but we leave issues about how to incorporate probabilities to another occasion.

$\mathsf{At}_G = \{P_a^i \mid a \in S_i\}$ be a set of atomic formulas (one for each $a \in S_i$). The propositional language for G, denoted $\mathcal{L}_G$, is the smallest set of formulas containing At_G and closed under the Boolean connectives $\neg$ and $\wedge$.

Formulas of $\mathcal{L}_G$ are intended to describe possible outcomes of the game. Given an informational context of a game $\mathcal{M}_G$, the formulas $\varphi \in \mathcal{L}_G$ is can be associated with subsets of the set of states in the usual way:

Definition 2.5. Let G be a strategic game, $\mathcal{M}_G = \langle W, \preceq, \sigma\rangle$ an informational context of G and $\mathcal{L}_G$ a propositional language for G. We define a map $[\![\cdot]\!]_{\mathcal{M}_G} : \mathcal{L}_G \to \wp(W)$ by induction as follows: $[\![P_a^i]\!]_{\mathcal{M}_G} = \{w \mid \sigma(w)_i = a\}$, $[\![\neg\varphi]\!]_{\mathcal{M}_G} = W - [\![\varphi]\!]_{\mathcal{M}_G}$ and $[\![\varphi \wedge \psi]\!]_{\mathcal{M}_G} = [\![\varphi]\!]_{\mathcal{M}_G} \cap [\![\psi]\!]_{\mathcal{M}_G}$.

Using the above language, for each informational context of a game $\mathcal{M}_G$, we can define $Do(\mathcal{M}_G)$, which describes what the players are going to do according to a fixed categorization procedure. To make this precise, suppose that $\mathbf{S}_i(\mathcal{M}_G) = (S_i^+, S_i^-)$ is a categorization for each i and define:

$$Do_i(\mathcal{M}_G) := \bigvee_{a \in S_i^+} P_i^a \wedge \bigwedge_{b \in S_i^-} \neg P_i^b$$

Then, let $Do(\mathcal{M}_G) = \bigwedge_i Do_i(\mathcal{M}_G)$.[9]

The general project is to understand the interaction between types of categorizations (eg., choice rules) and types of model transformations (representing the rational deliberation process). One key question is: Does a deliberation process *stabilize*(and if so, under what conditions)? (See Baltag and Smets (2009a) for general results here.) In this paper there are two main reasons why an upgrade stream would stabilize. The first is from properties of the transformation. The second is because the choice rule satisfies a monotonicity property so that, eventually, the categorizations stabilize and no new transformations can change the plausibility ordering. We are now ready to give a formal definition of a "deliberation sequence":

Definition 2.6 (Deliberation Sequence)**.** Given a game G and an informational context $\mathcal{M}_G$, a deliberation sequence of type τ (which we also call an upgrade sequence), induced by $\mathcal{M}_G$ is an infinite sequence of plausibility models $(\mathcal{M}_m)_{m \in \mathbb{N}}$ defined as follows:

[9] There are other ways to describe a categorization, but we leave this for further research.

$$\mathcal{M}_0 = \mathcal{M}_G \qquad \mathcal{M}_{m+1} = \tau(\mathcal{M}_m, Do(\mathcal{M}_m))$$

An upgrade sequence **stabilizes** if there is an $n \geq 0$ such that $\mathcal{M}_n = \mathcal{M}_{n+1}$.

3 Case Study: Iterated Admissibility

A key issue in the epistemic foundations of game theory is the epistemic analysis of iterated removal of *weakly* dominated strategies. Many authors have pointed out puzzles surrounding such an analysis Asheim and Dufwenberg (2003), Samuelson (1992), Brandenburger et al. (2008). For example, Samuelson Samuelson (1992) showed (among other things) that "common knowledge of admissibility" may be an inconsistent concept (in the sense that there is a game which does not have a model with a state satisfying 'common knowledge of rationality' (Samuelson 1992, Example 8, pg. 305)).[10] This is illustrated by the following game:

		Bob	
		L	R
Ann	u	1,1	1,0
	d	1,0	0,1

The key issue is that the assumption that players only play *admissible* strategies conflicts with the logic of iteratively removing strategies deemed "irrational". The general framework introduced above offers a new, dynamic perspective on this issue, and on reasoning with admissibility more generally.[11] Dynamically, Samuelson's non-existence result corresponds to the fact that the players' rational upgrade streams do not stabilize. That is, the players are not able to

[10] Compare with strict dominance: it is well known that common knowledge that players do not play weakly dominated strategies *implies* that the players choose a strategy profile that survives iterated removal of strictly dominated strategies.

[11] We do not provide an alternative epistemic characterization of this solution concept. Both Brandenburger et al. (2008) and Halpern and Pass (2009) have convincing results here. Our goal is to use this solution concept as an illustration of our general approach.

deliberate their way to a stable, common belief in admissibility. In order to show this we need the "right" notion of model transformation.

Our first observation is that the model transformations we discussed in Section 2.2 do not explain Samuelson's result.

Observation 1. Suppose that the categorization method is weak dominance and that $Do(\mathcal{M})$ is defined as above. For each of the model transformations discussed in Section 2.2 (i.e., public announcement, radical upgrade and conservative upgrade), any deliberation sequence for the above game stabilizes.

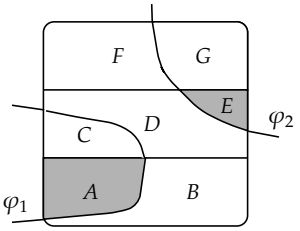
The proof of this Observation is straightforward since the language used to describe the categorization does not contain belief modalities[12]. This observation is nice, but it does not explain the phenomena noticed by Samuelson Samuelson (1992). The problem lies in the way we incorporate information when there is more than one element of $S_i^+(\mathcal{M})$ for some agent i.

It is well known that, in general, there are no rational principles of decision making (under ignorance or uncertainty) which *always* recommend a *unique* choice. In particular, it is not hard to find a game and an informational context where there is at least one player without a *unique* "rational choice". How should a rational player incorporate the information that more than one action is classified as "choice-worthy" or "rationally permissible" (according to some choice rule) for her opponent(s)? Making use of a well-known distinction due to Edna Ullmann-Margalit and Sidney Morgenbesser Ullmann-Margalit and Morgenbesser (1977), the assumption that all players are rational can help determine which options the player will *choose*, but rationality alone does not help determine which of the rationally permissible options will be "picked"[13]. What interests us is how to transform a plausibility model to incorporate the fact that there is a *set* of choice-worthy options for (some of) the players.

[12] An interesting extension would be to start with a multiagent belief model and allow players not only to incorporate information about which options are "choice-worthy", but also what beliefs their opponents may have. We leave this extension for future work, focusing here on setting up the basic framework.

[13] This line of thought led Cubitt and Sugden to impose a "privacy of tie breaking" property which says that players cannot *know* that her opponent will not pick an option that is classified as "choice-worthy" (Cubitt and Sugden 2011a, pg. 8) (cf. also Asheim and Dufwenberg (2003)'s "no extraneous restrictions on beliefs" property). Wlodeck Rabinovich takes this even further and argues that from the principle of indifference, players must assign equal probability to all choice-worthy options Rabinowicz (1992).

We suggest that a generalization of *conservative upgrade* is the notion we are looking for (see Holliday (2009) for more on this operation). The idea is to do an upgrade with a *set* of propositions $\{\varphi_1, \ldots, \varphi_n\}$ by letting the most plausible worlds be the union of each of the most plausible φ_i worlds:



$\uparrow\{\varphi_1, \varphi_2\} : A \cup E \prec B \prec C \cup D \prec F \cup G$

We do not give the formal definition here, but it should be clear from the example given above. It is not hard to see that this is not the same as $\uparrow\varphi_1 \vee \cdots \vee \varphi_n$, since, in general, $Min_{\preceq}([\![\varphi_1]\!] \cup \cdots \cup [\![\varphi_n]\!]) \neq \bigcup_i Min_{\preceq}([\![\varphi_i]\!])$. We must modify our definition of $Do(\mathcal{M})$: for each $i \in N$ let:

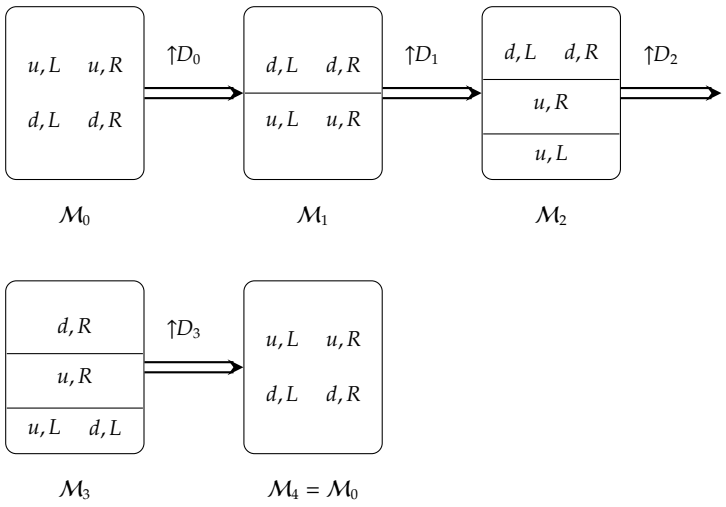
$$Do_i(\mathbf{S}_i(\mathcal{M}_G)) = \{P_a^i \mid a \in \mathbf{S}_i^+(\mathcal{M}_G)\} \cup \{\neg P_b^i \mid b \in \mathbf{S}_i^-(\mathcal{M}_G)\}$$

Then define $Do(\mathbf{S}(\mathcal{M}_G)) = Do_i(\mathbf{S}_i(\mathcal{M}_G)) \bigwedge Do_2(\mathbf{S}_2(\mathcal{M}_G)) \cdots \bigwedge Do_n(\mathbf{S}_n(\mathcal{M}_G))$, where if X and Y are two sets of propositions, then let $X \wedge Y := \{\varphi \wedge \psi \mid \varphi \in X, \psi \in Y\}$.

Observation 2. Suppose that the categorization method is weak dominance as explained in Section 2.3 and that $Do(\mathcal{M})$ is defined as above. Then, starting with the initial full model of the above game,[14] a generalized conservative upgrade stream does not stabilize.

The following upgrade stream illustrates this observation:

[14] A full model is one where it is common knowledge that each outcome of the game is equally plausible.



Intuitively, from $\mathcal{M}_0$ to $\mathcal{M}_2$ the agents have reasons to exclude d and R, leading them to the common belief that u, L is played. At that stage, however, d is admissible for Ann, canceling the reason the agents had to rule out this strategy. The rational response here is thus to suspend judgment on d, leading to $\mathcal{M}_3$. In this new model the agents are similarly led to suspend judgment on not playing R, bringing them back to M_0. This process loops forever: the agents' reasoning does not stabilize.

A corollary of this observation is that common belief in admissibility is not sufficient for the stabilization of upgrade streams. Stabilization also requires that all *and only* those profiles that are most plausible are admissible.

4 Stabilization Theorem

In this section we informally state and discuss a number of abstract principles which guarantee that a rational deliberation sequence will *stabilize*. The principles ensure that the categorizations are "sensitive" to the players' beliefs and that the players respond to the categorizations in the appropriate way.

We start by fixing some notation. Let U be a fixed set of states and G a fixed

strategic game. We confine our attention to transformations between models of G whose states come from the universe of states U. Let $\mathbb{M}_G$ be the set of all such plausibility models. A model transformation is then a function that maps a model of G and a finite set of formulas of $\mathcal{L}_G$ to a model in $\mathbb{M}_G$:

$$\tau : \mathbb{M}_G \times \wp_{<\omega}(\mathcal{L}_G) \to \mathbb{M}_G$$

where $\wp_{<\omega}(\mathcal{L}_G)$ is the set of finite subsets of $\mathcal{L}_G$. Of course, not all transformations τ make sense in this context.

The first set of principles that τ must satisfy ensure that the categorizations and belief transformation τ are connected in the "right way". One natural property is that the belief transformations treat *equivalent* formulas the same way. A second property we impose is that receiving exactly the same (ground) information twice does not have any effect on the players' beliefs. These are general properties of the belief transformation. Certainly, there are other natural properties that one may want to impose (for example, variants of the AGM postulates Alchourrón et al. (1985)), but for now we are interested in the minimal principles needed to prove a stabilization result.

The next set of properties ensure that the transformations respond "properly" to a categorization. First, we need a property to guarantee that the categorizations depend only on the players' beliefs. Second, we need to ensure that all upgrade sequences respond to the categorizations in the right way:

C2$^-$ For any upgrade sequence $(\mathcal{M}_n)_{n\in\mathbb{N}}$ in τ, if $a \in S_i^-(\mathcal{M}_n)$ then $\neg P_i^a$ is believed in $\mathcal{M}_{n+1}$.

C2$^+$ For any upgrade sequence $(\mathcal{M}_n)_{n\in\mathbb{N}}$ in τ, if $a \in S_i^+(\mathcal{M}_n)$ then $\neg P_i^a$ is not believed in $\mathcal{M}_{n+1}$

Finally, we need to assume that the categorizations are monotonic:

Mon$^-$ For any upgrade sequence $(\mathcal{M}_n)_{n\in\mathbb{N}}$, for all $n \geq 0$, for all players $i \in N$, $\mathbf{S}_i^-(\mathcal{M}_n) \subseteq \mathbf{S}_i^-(\mathcal{M}_{n+1})$

Mon$^+$ Either for all models $\mathcal{M}_G$, $\mathbf{S}_i^+(\mathcal{M}_G) = S_i - \mathbf{S}_i^-(\mathcal{M}_G)$ or for any upgrade sequence $(\mathcal{M}_n)_{n\in\mathbb{N}}$, for all $n \geq 0$, for all players $i \in N$, $\mathbf{S}_i^+(\mathcal{M}_n) \subseteq \mathbf{S}_i^+(\mathcal{M}_{n+1})$

In particular, **Mon**$^-$ means that once an option for a player is classified as "not rationally permissible", it cannot drop this classification at a later stage of the deliberation process.

Theorem 1. *Suppose that G is a finite game and all of the above properties are satisfied. Then every upgrade sequence* $(\mathcal{M}_n)_{n\in\mathbb{N}}$ *stabilizes.*

The proof can be found in the full version of the paper. The role of monotonicity of the choice has been noticed by a number of researchers (see Apt and Zvesper (2010b) for a discussion). This theorem generalizes van Benthem's analysis of rational dynamics van Benthem (2007) to soft information, both in terms of attitudes and announcements. It is also closely related to the result in Apt and Zvesper (2010b) (a complete discussion can be found in the full paper).

5 Concluding remarks

In this paper we have proposed a general framework to analyze how "proper" informational contexts my arise. We have provided general conditions for the stabilization of deliberation sequences in terms of structural properties of both the decision rule and the information update policy. We have also applied the framework to admissibility, giving a dynamic analysis of Samuelson's non-existence result.

Throughout the paper we have worked with (logical) models of *all out* attitudes, leaving aside probabilistic andgraded beliefs, even though the latter are arguably most widely used in the current literature on epistemic foundations of game theory. It is an important but non-trivial task to transpose the dynamic perspective on informational contexts that we advocate here to such probabilistic models. This we leave for future work.

Finally, we stress that the dynamic perspective on informational contexts is a natural complement and not an alternative to existing epistemic characterizations of solution concepts van Benthem et al. (2011), which offer rich insights into the consequences of taking seriously the informational contexts of strategic interaction. What we have proposed here is a first step towards understanding how or why such contexts might arise.

References

C. E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510 – 530, 1985.

K. Apt and J. Zvesper. Public announcements in strategic games with arbitrary strategy sets. In *Proceedings of LOFT 2010*, 2010a.

K. Apt and J. Zvesper. The role of monotonicity in the epistemic analysis of strategic games. *Games*, 1(4):381 – 394, 2010b.

G. Asheim and M. Dufwenberg. Admissibility and common belief. *Game and Economic Behavior*, 42:208 – 234, 2003.

R. Aumann. Interactive epistemology I: Knowledge. *International Journal of Game Theory*, 28:263–300, 1999.

R. Aumann and J. Dreze. Rational expectations in games. *American Economic Review*, 98:72 – 86, 2008.

A. Baltag and S. Smets. Group belief dynamics under iterated revision: Fixed points and cycles of joint upgrades. In *Proceedings of Theoretical Aspects of Rationality and Knowledge*, 2009a.

A. Baltag and S. Smets. ESSLLI 2009 course: Dynamic logics for interactive belief revision. Slides available online at `http://alexandru.tiddlyspot.com/#%5B%5BESSLLI09%20COURSE%5D%5D`, 2009b.

A. Baltag, S. Smets, and J. Zvesper. Keep 'hoping' for rationality: a solution to the backwards induction paradox. *Synthese*, 169:301–333, 2009.

K. Binmore. Modeling rational players: Part I. *Economics and Philosophy*, 3:179 – 214, 1987.

O. Board. Dynamic interactive epistemology. *Games and Economic Behavior*, 49: 49 – 80, 2004.

G. Bonanno and P. Battigalli. Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics*, 53(2):149–225, June 1999.

C. Boutilier. *Conditional Logics for Default Reasoning and Belief Revision*. PhD thesis, University of Toronto, 1992.

A. Brandenburger. The power of paradox: some recent developments in interactive epistemology. *International Journal of Game Theory*, 35:465–492, 2007.

A. Brandenburger, A. Friedenberg, and H. J. Keisler. Admissibility in games. *Econometrica*, 76:307–352, 2008.

R. Cubitt and R. Sugden. Common reasoning in games: A Lewisian analysis of common knowledge of rationality. CeDEx Discussion Paper, 2011a.

R. Cubitt and R. Sugden. The reasoning-based expected utility procedure. *Games and Economic Behavior*, page In Press, 2011b.

J. Gerbrandy. *Bisimulations on Planet Kripke*. PhD thesis, University of Amsterdam, 1999.

J. Halpern and R. Pass. A logical characterization of iterated admissibility. In A. Heifetz, editor, *Proceedings of the Twelfth Conference on Theoretical Aspects of Rationality and Knoweldge*, pages 146 – 155, 2009.

W. Holliday. Trust and the dynamics of testimony. In *Logic and Interaction Rationality: Seminar's Yearbook 2009*, pages 147 – 178. ILLC Technical Reports, 2009.

J. B. Kadane and P. D. Larkey. Subjective probability and the theory of games. *Management Science*, 28(2):113–120, 1982.

P. Lamarre and Y. Shoham. Knowledge, certainty, belief and conditionalisation. In *Proceedings of the International Conference on Knoweldge Representation and Reasoning*, pages 415 – 424, 1994.

D. Lewis. *Counterfactuals*. Blackwell Publishers, Oxford, 1973.

J. Plaza. Logics of public communications. In M. L. Emrich, M. S. Pfeifer, M. Hadzikadic, and Z. Ras, editors, *Proceedings, 4th International Symposium on Methodologies for Intelligent Systems*, pages 201–216 (republished as Plaza (2007)), 1989.

J. Plaza. Logics of public communications. *Synthese: Knowledge, Rationality, and Action*, 158(2):165 – 179, 2007.

W. Rabinowicz. Tortous labyrinth: Noncooperative normal-form games between hyperrational players. In C. Bicchieri and M. L. D. Chiara, editors, *Knowledge, Belief and Strategic Interaction*, pages 107 – 125, 1992.

H. Rott. Shifting priorities: Simple representations for 27 itereated theory change operators. In H. Lagerlund, S. Lindström, and R. Sliwinski, editors, *Modality Matters: Twenty-Five Essays in Honour of Krister Segerberg*, volume 53 of *Uppsala Philosophical Studies*, pages 359 – 384, 2006.

O. Roy and E. Pacuit. Substantive assumptions and the existence of universal knowledge structures: A logical perspective. Under submission, 2010.

A. Rubinstein. The electronic mail game: A game with almost common knowledge. *American Economic Review*, 79:385 – 391, 1989.

L. Samuelson. Dominated strategies and common knowledge. *Game and Economic Behavior*, 4:284 – 313, 1992.

B. Skyrms. *The Dynamics of Raitonal Deliberation*. Harvard University Press, 1990.

E. Ullmann-Margalit and S. Morgenbesser. Picking and choosing. *Social Research*, 44:757 – 785, 1977.

J. van Benthem. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics*, 14(2):129 – 155, 2004.

J. van Benthem. Rational dynamics and epistemic logic in games. *International Game Theory Review*, 9(1):13–45, 2007.

J. van Benthem. *Logical Dynamics of Information and Interaction*. Cambridge University Press, 2010.

J. van Benthem and A. Gheerbrant. Game solution, epistemic dynamics and fixed-point logics. *Fund. Inform.*, 100:1–23, 2010.

J. van Benthem, E. Pacuit, and O. Roy. Towards a theory of play: A logical perspective on games and interaction. *Games*, 2(1):52–86, 2011.

Learning in a changing world, an algebraic modal logical approach

Prakash Panangaden and Mehrnoosh Sadrzadeh

School of Computer Science, McGill University, Department of Computer Science, University of Oxford
prakash@cs.mcgill.ca, mehrs@cs.ox.ac.uk

Abstract

We develop an algebraic modal logic that combines epistemic and dynamic modalities with a view to modelling information acquisition (learning) by automated agents in a changing world. Unlike most treatments of dynamic epistemic logic, we have transitions that "change the state" of the underlying system and not just the state of belief of the agents. The key novel feature that emerges is the need to have a way of "inverting transitions" and distinguishing between transitions that "really happen" and transitions that are possible.

Our approach is algebraic, rather than being based on a Kripke-style semantics. The semantics are given in terms of quantales. We study a class of quantales with the appropriate inverse operations and prove properties of the setting. We illustrate the ideas with toy robot-navigation problems. These illustrate how an agent learns information by taking actions.

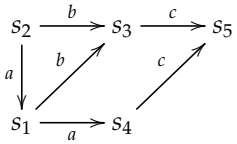
1 Introduction

Epistemic logic has proved very important in the analysis of protocols in distributed systems (see, for example, Fagin et al. (1995)) and, more generally in any situation where there is some notion of cooperation or "agreement" between agents. The original work in distributed systems, by Halpern and Moses Halpern and Moses (2000) and several others modelled the knowledge and belief of agents using Kripke-style models Kripke (1963) . In these models there are a set of states (often called "possible worlds") in which the agent could be and, for each agent, an equivalence relation on the states. If two states are equivalent to an agent then that agent cannot "tell them apart". An agent "believes" a fact φ in the state s if, in all states t that the agent "thinks" is equivalent to s, the fact φ holds. The quoted words in the preceding sentences are, of course, unnecessary anthropomorphisms that are intended to give an intuition for the definitions.

A vital part of any analysis is how processes "learn" as they participate in the protocol. The main literature on distributed systems treat this as a change in the Kripke equivalence relations and argue about these changes *only in the semantics*. For instance, the "interpreted system" setting of Fagin et al. (1995) uses runs of protocols to model how the local and global information states of the system change, but this kind of dynamics has no counterpart in the syntax of logic: epistemic logic does not have the "dynamic" modalities that refer to updating of the state of belief. On the other hand, dynamic epistemic logic has indeed been studied; see, for example the original papers Plaza (2007), Gerbrandy and Groeneveld (1997), Baltag and Moss (2004) and the recent book Ditmarsch et al. (2007). In the second author's doctoral dissertation an algebraic approach to dynamic epistemic logic was studied in depth Baltag et al. (2007), Sadrzadeh (2006). Relational models and logics where the temporal structure is chosen over the dynamic one have also been studied, for example in the context of Alternating Temporal Time Epistemic Logic (ATEL) Ågotnes (2006). But in this paper our focus is on the structures that prefer the dynamic structure over the temporal ones.

The bulk of the work in this area (apart from Baltag (2002), van Ditmarsch et al. (2005), van Benthem et al. (2006), the differences with which we will discuss below), concerns situations where the *state of belief* is changed by broadcasts but not situations where the *state of the system* is changed. An illuminating concrete example of such situations arises in robot navigation. A general feature of

these protocols is that an agent is given the description of a place, but cannot determine exactly where it is; however, it can move and as a result may acquire information that allows it to infer its present location. Consider a robot that is given the map of a small computing laboratory with 5 rooms accessible via 3 actions, as follows:



The robot cannot see the whole path in front of it; it can only see one step ahead of itself, hence only knows about the one-step actions that he can take in each state. Since it can do the same immediate actions in the pair s_1, s_2, it cannot tell s_1 and s_2 apart, and similarly for the pair s_3, s_4. Once in s_1 (similarly for s_2), it believes that it could be in s_1 or s_2, and once in s_3 (similarly for s_4), it believes that it could be in s_3 or s_4. But if it is in s_1 and it performs an a action, then it reaches s_4 and learns where it is and where it had been just before the a action. So after doing an a in s_1, he believes that he is in s_4.

A deeper investigation of such situations reveals that it is not a question of "patching up" the existing theory of dynamic epistemic logics and in particular the algebraic approach in which the second author was involved Baltag and Moss (2004). There are some interesting fundamental changes that need to be made. First of all, one has to distinguish between transitions that exist in the agent's "mental model" of the system and actions that *actually occur*. Second, one has to introduce a converse dynamic modality in order to correctly formulate the axioms for updating the robot's belief about his whereabouts[1]. To see why, let us reason as we think the robot whould: when it reaches s_4, it checks with its map and reasons that the only way it could have reached s_4 would be that it was originally in s_1. It rules out s_3 from its uncertainty set about s_4, because, according to the map, it could not have reached s_3 via an a action. We have two types of data here, the locations and actions described on the map versus the ones in reality. The data on the map are hard-coded in the robot and there is no uncertainty about it, the map fully describes the system. But there is some uncertainty about the real locations. The robot is

[1]Note that here we are only talking about propositional belief, as we have assumed that the robot does not have any uncertainties about the one-step actions that it can take and knows about all of them all the time. Thus we do not fall into the problems discussed in Ågotnes (2006), where expressing the knowledge of agents about all possible actions is impossible.

uncertain about its location but the actions it takes change its uncertainties. The other issue is that to be able to encode what actions *could have* led the robot to where it is, it needs to look back, so we need a converse operation to reason about the past. Now by moving from s_1 to s_4, the robot has changed its uncertainty, acquired information, and learned where it is located. This is exactly the manner in which our new *uncertainty reduction* axiom formalizes the elimination of past uncertainties: after performing a certain move in the real world, the robot consults its description, considers its possibilities and eliminates the ones that could not have been reached as a result of the action it just performed. Furthermore with this converse operation, we can also derive information about the past, that the robot was in s_1 before doing action a.

This paper presents an algebraic theory with these features. The advantage of working in the algebraic setting is that it abstracts over the details of the Kripke structures and showcases the high-level structure of the actions and their updates. It turns out that epistemic update is the action of the quantale of programs/actions on the module of propositions (factual and epistemic), hence it is the left adjoint to the dynamic modality which encodes the weakest precondition of Hoare Logic. Epistemic modalities are also encoded as an adjoint pair: the belief modality is the right adjoint of the appearance map, which is the lifting to subsets of the accessibility relation of the Kripke structure. This results in a simple method of computing belief acquisition after an action: uniform unfolding of epistemic and dynamic adjunctions, which simplifies, to a great extent, the proofs of complex protocols and puzzles, such as the muddy children, even the versions with dishonest children, see Baltag et al. (2007), Sadrzadeh (2006).

Regarding related work, the reason the setting of Baltag et al. (2007) fails for the navigation situations is that its key learning axiom is only geared towards epistemic actions and is not powerful enough for fact-changing actions. It requires that the uncertainty about (possible states of) a location after an action to be included in the result of applying the action to the uncertainty about the location beforehand, a property similar to *perfect recall* in protocol models of Halpern and Moses (2000). This fails here, since after performing an a at s_1 one ends up in s_4, hence uncertainty about s_1 after an a is the same as uncertainty about s_4, consisting of s_3 and s_4. But performing a on the set of uncertainties about s_1, consisting of s_1 and s_2, results in both s_4 and s_1. However, $\{s_3, s_4\}$ is not included in $\{s_4, s_1\}$. Moreover, after the robot moved to s_4, it can conclude that it *was* in s_1 before moving; the language of Baltag et al. (2007) simply cannot express these *past tense* properties. Finally, dynamic epistemic logic has been

extended with *assignments* and *post-conditions* to model certain types of fact-changing actions Baltag (2002), van Ditmarsch et al. (2005), van Benthem et al. (2006). Location-changing actions have not been studied there and indeed we faced difficulties trying to apply their setting following our own intuitions. In a nutshell, one has to divide the transition system into two separate models: a state model whose states are the states of the transition system and whose accessibility relations are the uncertainties, and an action model with a labeled superset of transitions as states and an extra notion of uncertainty about them as accessibility relations. One loses the above simple image, but more important is that the usual *update product* of these two models does not satisfy our desired belief properties. For details and examples see Horn (2011).

We develop an algebraic setting to formalize information acquisition from such navigation protocols. We study special cases of the past and future deterministic action and converse action operations of the algebra and prove some of their axiomatic properties. We apply our algebra to model a grid and a map-based navigation protocol and use the axioms to prove that the agent learns where he is and was after moving about. Further applications of our setting are to AI, mobile communication, security, and control theory.

2 The Algebra of di-systems

We need to model "actions" and "formulas". The actions are modelled by a quantale while the propositions are a module over the quantale; i.e. actions modify propositions.

Definition 2.1. A **quantale** $(Q, \bigvee, \bullet, 1)$ is a complete sup-lattice equipped with a unital monoid structure satisfying $q\bullet(\bigvee_i q_i) = \bigvee_i (q\bullet q_i)$ and $(\bigvee_i q_i)\bullet q = \bigvee_i (q_i \bullet q)$.

Instead of arbitrary complete sup-lattices, we take our complete sup-lattices to be *complete completely distributive prime-algebraic lattices*. Recall that a prime element, or simply "prime", p in lattice has the property that for any x, y in the lattice, $p \leq x \vee y$ implies that $p \leq x$ or $p \leq y$; "prime algebraic" means that every element is the supremum of the primes below it. The restriction to prime algebraic lattices, while definitely a serious restriction, includes a large class of interesting examples, for instance all transition systems. Prime algebraicity would be a restriction for extensions to probabilistic systems; we will address such issues in future work. The use of algebraicity is to be able to use simple

set-theoretic arguments via the representation theorem for such lattices Gehrke and Jónsson (1994), Winskel (2009). For finite distributive lattices it is not a restriction at all because of Birkhoff's classical duality theory. Henceforth, we will not explicitly state that we are working with (completely) distributive prime-algebraic lattices.

Definition 2.2. A **right-module** over Q is a sup-lattice M with an action of Q on M, denoted by $- \cdot -\colon M \times Q \to M$ and satisfying the following axioms:

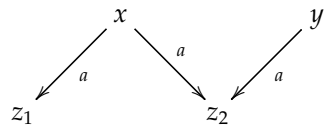
$$(m \cdot q) \cdot q' = m \cdot (q \bullet q')$$
$$m \cdot (\bigvee_i q_i) = \bigvee_i (m \cdot q_i) \qquad (\bigvee_i m_i) \cdot q = \bigvee_i (m_i \cdot q)$$
$$m \cdot 1 = 1 \cdot m = m$$

We call the collection of actions and propositions a *system*.

Definition 2.3. A **system** is a pair consisting of a quantale Q and a right-module M over Q. We write $(M, Q, \cdot)$ for a system.

This is closely related to the definition of Abramsky and Vickers who have also argued for the application to Computer Science of quantales of actions, see Abramsky and Vickers (1993). As is usually done, we interpret elements of the module as *propositions* and the order as entailment, thus $m \vee m'$ is the logical disjunction and $\bot$ is the falsum. The elements of the quantale are interpreted as actions and the order is the order of non-determinism, thus $q \vee q'$ is the non-deterministic choice and $\bot$ is crash, monoid multiplication $q \bullet q'$ is sequential composition, and its unit 1 is the action that does nothing.

Example 1. Consider the following transition system



We model it as a system $(\mathcal{L}(S), \mathcal{M}(A^*), \cdot)$, where A^* is the free monoid generated from the set $A = \{a\}$ with the multiplication being juxtaposition and its unit the empty string. $\mathcal{M}(A^*)$ is the quantale generated on that monoid and $\mathcal{L}(S)$ is the sup-lattice generated from the set $S = \{x, y, z_1, z_2\}$. The most concrete

examples of $\mathcal{L}(S)$ and $\mathcal{M}(A^*)$ are $\mathcal{P}(S)$ and $\mathcal{P}(A^*)$. The action on atoms is given by $x \cdot a = z_1 \vee z_2$ and $y \cdot a = z_2$, whereas $z_1 \cdot a = z_2 \cdot a = \bot$. This is extended to juxtaposition and choice (subsets of actions), as well as subsets of states pointwise.

Example 2. The powerset $\mathcal{P}(S)$ of a set S is the right module of the quantale of all the relations thereon $\mathcal{P}(S \times S)$. Relational composition is the monoid multiplication, the diagonal relation is its unit, and the join is set union. The action is the pointwise image of the relation, i.e. for $W \subseteq S$ and $R \subseteq S \times S$

$$W \cdot R = \bigcup_{w \in W} R[w] = \{z \in S \mid \exists w \in W, (w, z) \in R\}$$

Since the action preserves all the joins of its module, the map $- \cdot q \colon M \to M$, obtained by fixing the quantale argument, has a Galois right adjoint that preserves all the meets. This is denoted by $- \cdot q \dashv [q]-$ and defined in the canonical way, as follows:

$$[q]m := \bigvee \{m' \mid m' \cdot q \leq m\}$$

The right adjoints stand for the "dynamic modality" of Hoare logic, encoding the "weakest preconditions" of programs. Each of $[q]m$ is read as "after doing action q or running program q, proposition m holds". This is, in effect, all the propositions such that if true at the input of q, then at its output m holds. One gets very nice logical properties, relating the action and its adjoint to each other and to the $\vee$ and $\wedge$ operators of the lattice and their units $\bot$ and $\top$. Some examples are exhibited in the following proposition.

Proposition 1. *The following inequalities hold in any system* $(M, Q, \cdot)$*:*

(1) $([q]m) \cdot q \leq m$ (2) $m \leq [q](m \cdot q)$

(3) $(m \wedge m') \cdot q \leq m \cdot q \ \wedge \ m' \cdot q$ (4) $m \cdot (q \wedge q') \leq m \cdot q \ \wedge \ m \cdot q'$

(5) $[q](m \vee m') \geq [q]m \ \vee \ [q]m'$

(6) $q \leq q' \implies [q']m \leq [q]m$ (7) $[\bot]m = \top$

(8) $[q \vee q']m = [q]m \ \wedge \ [q']m$ (9) $[q \vee q']m \leq [q]m \ \vee \ [q']m$

(10) $[q \wedge q']m \geq [q]m \ \vee \ [q']m$ (11) $[q \wedge q']m \geq [q]m \ \wedge \ [q']m$

(12) $[\bigvee_i q_i]m = \bigwedge_i [q_i]m$

Proof. (1) and (2) are immediate consequences of the definition of $[q]$ as a right adjoint. (3), (4), (5) follow from monotonicity. For (6), assume $q \leq q'$ and we have to show

$$\bigvee \{m' \mid m' \cdot q' \leq m\} \leq \bigvee \{m'' \mid m'' \cdot q \leq m\}.$$

It suffices to show that an arbitrary element of the lhs set is in the rhs set. Take one such element m', we have $m' \cdot q' \leq m$, but since $q \leq q'$, we also have that $m' \cdot q \leq m' \cdot q'$, hence $m' \cdot q \leq m$, i.e. m' is also in the rhs set. For (7), the direction $[\bot]m \leq \top$ is trivial, the other direction $\top \leq [\bot]m$ is equivalent to $\top \cdot \bot \leq m$, which holds since $\top \cdot \bot = \bot$. For (8), the $\leq$ direction follows from (6) and definition of meet, for the $\geq$ direction we have to show $[q]m \ \wedge \ [q']m \leq [q \vee q']m$, which is by adjunction equivalent to $([q]m \ \wedge \ [q']m) \cdot (q \vee q') \leq m$. By join preservation of action, this is equivalent to $([q]m \ \wedge \ [q']m) \cdot q \ \vee ([q]m \ \wedge \ [q']m) \cdot q' \leq m$. To show this, we have to show that both disjuncts are less than or equal to m. Consider the first one, by (3) and transitivity, it suffices to show $[q]m \cdot q \ \wedge \ [q']m \cdot q \leq m$, by the definition of meet and transitivity it suffices to show either of the conjuncts satisfy the inequality, now $[q]m \cdot q \leq m$, is true by (1). The proofs of the remaining items follow from these in a similar way, e.g. (12) follows from (7) and (8). □

Definition 2.4. Consider a quantale Q with a right action $\cdot$ on the sup lattice M and the converse of the action – written $\cdot^c$ for the purposes of this definition. If $\cdot^c$ preserves the arbitrary joins and 1 in both arguments and satisfies the following axioms:

$$(i) \quad p \leq p' \cdot q \Leftrightarrow p' \leq p \cdot^c q \qquad p, p' \text{ primes}$$
$$(ii) \quad m \cdot^c (q \bullet q') = (m \cdot^c q') \cdot^c q$$

then we refer to the system as a **converse di-system** and denote it by $(M, Q, \cdot, \cdot^c)$.

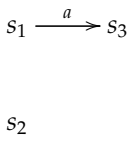
Proposition 2. *The following hold in any converse di-system, for p a prime:*

$$(i) \quad m \cdot q \leq m' \implies m \leq m' \cdot^c q \qquad \text{whenever } \forall p \leq m, p \cdot q \neq \bot$$
$$(ii) \quad m \cdot^c q \leq m' \implies m \leq m' \cdot q \qquad \text{whenever } \forall p \leq m, p \cdot^c q \neq \bot$$

Proof. Consider (i) and assume the antecedent and the side condition. Take an arbitrary prime $p \leq m$, by the antecedent and monotonicity of action $p \cdot q \leq m \cdot q \leq m'$. By the side condition $p \cdot q \neq \bot$, hence there is a prime below it $p' \leq p \cdot q$. By axiom (i) of the above definition this is equivalent to $p \leq p' \cdot^c q$. Since $p' \leq p \cdot q \leq m'$, by monotonicity we obtain that $p' \cdot^c q \leq m' \cdot^c q$, and by transitivity it follows that $p \leq m' \cdot^c q$. We are in a prime algebraic lattice, hence m is the join of the primes p below it, so we obtain that $m \leq m' \cdot^c q$. Proof of (ii) is *mutatis mutandis*. □

In a non prime-algebraic lattice the side conditions can be generalized to $\forall n \leq m, n \neq \bot \Rightarrow n \cdot q \neq \bot$. It would not have sufficed to just require this for m as it would be false even for examples that are transition systems.

Example 3.



In this example if we take the states to be primitive propositions then we have $(s_1 \vee s_2) \cdot a \neq \bot$, we take $m = s_1 \vee s_2$ and $m' = s_3$, so $m \cdot a \leq m'$ but $m' \cdot^c a = s_1$ and clearly $m \not\leq s_1$[2].

To avoid repeating the non-emptiness side conditions in the preceding definitions and propositions, in the rest of the paper we assume that our transition systems have a null start state and a nulll end state. There is an arrow to every state from the null start state and arrow to the null end state from every state. In algebraic terms, for all primes $p \in M$ and all actions $q \in Q$, we have that $p \cdot q \neq \bot$ and $p \cdot^c q \neq \bot$.

[2]We thank Mai Gehrke and Sam van Gool for pointing this out.

Proposition 3. *Items* (i), (ii) *of proposition 2 are equivalent to the following:*

$$(i')\ m \leq (m \cdot q) \cdot^c q \qquad\qquad (ii')\ m \leq (m \cdot^c q) \cdot q$$

Proof. From (i) to (i'): apply (i) to $m \cdot q \leq m \cdot q$ by taking m' to be $m \cdot q$. From (ii) to (ii'): apply (ii) to $m \cdot^c q \leq m \cdot^c q$ and take m' to be $m \cdot^c q$. From (i') to (i): assume the premise of (i), i.e. $m \cdot q \leq m'$, apply the converse action to both sides $m \cdot q \cdot^c q \leq m' \cdot^c q$, from this and (i') by transitivity obtain $m \leq m' \cdot^c q$. From (ii') to (ii): assume the premise of (ii), i.e. $m \cdot^c q \leq m'$, apply the action to both sides $m \cdot^c q \cdot q \leq m' \cdot q$, from this and (ii') by transitivity obtain $m \leq m' \cdot q$. □

Definition 2.5. A converse di-system is past-deterministic iff $m \leq m' \cdot q \implies m \cdot^c q \leq m'$. It is future-deterministic iff $m \leq m' \cdot^c q \implies m \cdot q \leq m'$.

Example 4. Consider the transition system of example 1, this is moreover an example of a converse di-system $(\mathcal{L}(S), \mathcal{M}(A^*), \cdot, \cdot^c)$, where the converse action is given by $z_1 \cdot^c a = x$, and $z_2 \cdot^c a = x \vee y$. It is easy to check that these satisfy the inequalities of definition 2.4. It is also easy to see that they satisfy $(i), (ii)$ of proposition 2 but not their converses: the transition system is neither past-deterministic nor future-deterministic. A counterexample for the converse of part (i) is $x \leq z_2 \cdot^c a$ but $x \cdot a \not\leq z_2$. If we eliminate the leftmost edge, then the system becomes future-deterministic and the converse of (i) holds. A counterexample for the converse of part (ii) is $z_2 \leq y \cdot a$ but $z_2 \cdot^c a \not\leq y$. If we eliminate the rightmost edge, then the system becomes past-deterministic and the converse of (i) holds.

Example 5. The transition system of the introduction is a future-deterministic converse di-system, in the same way as the above example, where $S = \{s_1, \ldots, s_5\}$ and $A = \{a, b, c\}$. It is not past-deterministic, since $s_3 \cdot^c b = s_1 \vee s_2$, also $s_5 \cdot^c c = s_3 \vee s_4$.

Example 6. Consider the setting of example 2, this is also an example of a converse di-system, where the converse action is the point wise image of the converse relation, i.e. for $W \subseteq S$ and $R^c \subseteq S \times S$ converse of R, we have:

$$W \cdot^c R = \bigcup_{w \in W} R^c[w] = \{z \in W \mid \exists w \in W, (w, z) \in R^c\}$$

It is easy to see that $W \cdot^c R = W \cdot R^c$. If $R^c[w]$ is a singleton then this di-system becomes a past-deterministic one, if R is a singleton, it becomes future-deterministic.

Proposition 4. *The axioms of definition 2.4 are valid for any labelled transition system.*

Proof. By an elementary argument, e.g. see Abramsky and Vickers (1993), there is a canonical bijection between labelled transition systems $(S, \rightarrow^a)_{a\in A}$ and pairs $(\mathcal{P}(S), \mathcal{P}(A^*))$ where $\mathcal{P}(S)$ is the right module of quantale $\mathcal{P}(A^*)$ under the direct-image action. Given any labelled transition system, one can define the converse of a transition as $s'(\rightarrow^a)^c s$ iff $s \rightarrow^a s'$. Consider axiom (i), suppose for $s, s' \in S$ and $a \in A$ we have $\{s\} \subseteq \{s'\}\cdot a$, i.e. $s' \rightarrow^a s$, by definition of converse of transition system, this is equivalent to $s(\rightarrow^a)^c s'$, i.e. $\{s'\} \subseteq \{s\} \cdot^c a$. The proof for axiom (ii) is also routine. □

The converse action preserves all the joins of the module, thus similar to the action, it has a Galois right adjoint denoted by $- \cdot^c q \dashv [q]^c -$, canonically defined using the converse action:

$$[q]^c m := \bigvee \{m' \in M \mid m' \cdot^c q \leq m\}$$

Similar to $[q]m$, we read $[q]^c m$ as "before doing action q, proposition m held".

For the special cases of past/future deterministic systems, the action and its converse relate to each other in stronger ways. From definition 2.5, it easily follows that:

Proposition 5. *In a past-deterministic converse di-system we have* $m \leq m' \cdot q \iff m \cdot^c q \leq m'$ *. In a future-deterministic converse di-system we have* $m \leq m' \cdot^c q \iff m \cdot q \leq m'$ *.*

As a result we have that:

Proposition 6. *In a past-deterministic converse di-system we have* $m = m \cdot q \cdot^c q$*. In a future-deterministic converse di-system we have* $m = m \cdot^c q \cdot q$*.*

Putting the above two propositions together, we obtain:

Proposition 7. *In a past and future-deterministic converse di-system we have* $- \cdot^c q \dashv - \cdot q$ *and* $- \cdot q \dashv - \cdot^c q$*.*

The following propositions are of particular theoretical interest, since it turns out that in the presence of a Boolean negation on the module, the de Morgan dual of the right adjoint to the action is the converse action, and the de Morgan

dual of the right adjoint to the converse action is the action. In other words $- \cdot q$ and $[q]^c -$ are de Morgan duals and so are $- \cdot^c q$ and $[q]-$. Our modules need not necessarily be Boolean, nevertheless, these connections can be expressed using the following properties, which axiomatize de Morgan duality in the absence of negation, see Dunn (1995).

Before the next proposition we need a lemma.

Lemma 1. *If p is a prime and l is any element of the lattice of propositions we have*

$$(p \cdot q) \wedge l \neq \bot \iff p \leq l \cdot^c q.$$

Proof. For the forward direction we assume that p' is a prime less than $(p \cdot q) \wedge l$. Since $p' \leq p \cdot q$ we have from (i) of Def. 2.4 that $p \leq p' \cdot^c q$ and from $p' \leq l$ and monotonicity of $\cdot^c$ we have $p' \cdot^c q \leq l \cdot^c q$ hence $p \leq l \cdot^c q$.

For the reverse direction show the contrapositive, i.e. that

$$p \leq l \cdot^c q \implies (p \cdot q) \wedge l \neq \bot.$$

We note that l is the sup of the primes below it, so we can write $l = \bigvee_i p_i$. Then using distributivity of $\cdot^c$ over sups we calculate

$$p \leq (\bigvee_i p_i) \cdot^c q = \bigvee_i (p' \cdot^c q).$$

Since p is a prime this means that there is some p', a prime less than l, such that $p \leq p' \cdot^c q$. Using part (i) of Def. 2.4 again we get $p' \leq p \cdot q$. Thus $p' \leq (p \cdot q) \wedge l$ so $(p \cdot q) \wedge l \neq \bot$. □

Proposition 8. *In any converse di-system,*

$$[q](l \vee l') \leq [q]l \ \vee \ (l' \cdot^c q) \qquad \textit{and} \qquad (l \cdot^c q) \ \wedge \ [q]l' \leq (l \wedge l') \cdot^c q$$

Proof. We will show the inequalities by showing that any prime less than the left hand side is also less than the right hand side. Let p be a prime such that $p \leq [q](l \vee l')$. Then $p \cdot q \leq l \vee l'$, which implies $p \leq (l \vee l') \cdot^c q$ and by distributivity we have $p \leq (l \cdot^c q) \vee (l' \cdot^c q)$. Since p is a prime, either $p \leq l \cdot^c q$ or $p \leq l' \cdot^c q$. If $p \leq l' \cdot^c q$ we have nothing left to prove. Suppose, therefore that $p \not\leq l' \cdot^c q$; of course, this means that $p \leq l \cdot^c q$. Now suppose $p \not\leq [q]l$. This implies that there exists a prime p' such that $p' \leq p \cdot q$ and $p' \not\leq l$. Since we have assumed that

$p \not\leq l' \cdot^c q$ we can use Lemma 1 to deduce that $(p \cdot q) \wedge l' = \bot$; thus, no prime below p, in particular p', can be below l'. Hence, since p' is a prime, p' is not below $l \vee l'$ which contradicts $p' \leq p \cdot q \leq l \vee l'$. Thus $p \leq [q]l$. Combining this with the other case, we have shown that $p \leq [q]l \vee (l' \cdot^c q)$. Since this is true for any prime, the required inequality follows.

For the second inequality we proceed in a similar fashion. Assume that p is a prime below $[q]l \wedge (l' \cdot^c q)$. From $p \leq [q]l$ we get $p \cdot q \leq l$. We want to show that $p \leq (l \wedge l') \cdot^c q$. Using Lemma 1 this is equivalent to $(p \cdot q) \wedge (l \wedge l') \neq \bot$. Since $p \cdot q \leq l$ this simplifies to showing $(p \cdot q) \wedge l' \neq \bot$. Again by Lemma 1 this is equivalent to showing $p \leq l' \cdot^c q$ which follows from the assumption on p. □

Proposition 9. *In any converse di-system we have*

$$[q]^c(l \vee l') \leq [q]^c l \ \vee \ (l' \cdot q) \qquad \textit{and} \qquad (l \cdot q) \ \wedge \ [q]^c l' \leq (l \wedge l') \cdot q$$

Proof. Similar to the above. □

In a Boolean module, these de Morgan dualities become explicit, as follows:

Proposition 10. *If the module of a past and future deterministic converse di-system is a Boolean algebra with negation operator $\neg -\colon M \to M$, we have $l \cdot q = \neg[q]^c \neg l$ and $l \cdot^c q = \neg[q]\neg l$.*

Proof. Consider the first equation, we unfold the definitions of $l \cdot q$ as the left adjoint to $[q]-$ and of $[q]^c \neg l$ as the right adjoint to $- \cdot^c q$, hence equivalently show the following

$$\bigwedge \{l' \in M \mid l \leq [q]l'\} \quad = \quad \neg \bigvee \{l' \in M \mid l' \cdot^c q \leq \neg l\}$$

Boolean negation turns joins to meets and negates the argument, hence the rhs is equivalent to $\bigwedge\{l' \in M \mid l' \cdot^c q \not\leq \neg l\} = \bigwedge\{l' \in M \mid l' \cdot^c q \leq \neg\neg l\}$. This equal to lhs, since Boolean negation is involutive and by adjunction $l \leq [q]l'$ is equivalent to $l \cdot q \leq l'$, which by definitions 2.4 and 2.5 is equivalent to $l' \cdot^c q \leq l$. Here we are using the part of the definition based on the fact that the di-system is future-deterministic. The proof of the second equation above is similar and uses the fact that the di-system is past-deterministic. □

Following Karger (1996), one can define a Kleene star for iteration as $m \cdot^* q := m \vee m \cdot q \vee m \cdot (q \bullet q) \vee \cdots$ and $[q]^b m := m \wedge [q]m \wedge [q \bullet q]m \wedge \cdots$, and similarly for

the converse action and its right adjoint. It easily follows that these preserve the adjunctions $- \cdot^* q \dashv [q]^\flat -$ and $- \cdot^{c*} q \dashv [q]^{c\flat} -$, and that in a Boolean module they also preserve the de Morgan dualities $m \cdot^* q = \neg[q]^{c\flat}\neg m$ and $m \cdot^{c*} q = \neg[q]^\flat \neg m$.

3 Navigation di-systems

To distinguish the "potential" actions that happen in the model used by the agent, for example, actions described by a map, from the "real" actions that take place in the real world, we use a formalism in which there is a family of systems indexed by action sequences representing how the model is modified as actions occur The protocols that we are interested in modeling have a set of locations or states and a set of actions, so our navigation di-systems are the concrete ones described in example 1. These states give rise to the module on which the quantale of actions acts. Let A be a set of actions and let Q be the free quantale generated by A, i.e. $\mathcal{P}(A^*)$, where A^*, is, as usual, the free monoid generated by A, in other words the set of sequences of actions. We write α for a typical element of A^* and we write αa for the sequence α with a appended. We let S be a set of states and we define M to be $\mathcal{P}(S)$, the powerset of S. With this choice of M and Q we assume that we have a di-system $(M, Q, \cdot, \cdot^c)$.

Definition 3.1. A **navigation pre di-system** is a di-system $(M, Q, \cdot, \cdot^c)$ together with a second action of Q and its converse $((M, Q, \cdot, \cdot^c), \odot, \odot^c)$, where $s \odot a$ is an element of the module and $(s \odot a) \odot b$ can be regarded as $s \odot ab$ and, in general $(s \odot \alpha) \odot a = s \odot \alpha a$. In order to describe the action of a general member of Q we lift the action from primes to arbitrary sets in a point wise fashion, i.e. $s \odot (a \vee b) = (s \odot a) \vee (s \odot b)$.

Potential and real actions have the same labels and both live in the quantale Q. Potential actions change the state of the map via the actions $\cdot$ and $\cdot^c$, real actions change the state of the world via the actions $\odot$ and $\odot^c$. The reason potential and real actions are distinguished from one another is that their targets have different uncertainties. For example, consider the scenario of the introduction, modeled as a converse di-system in example 5. There, the uncertainty of $s_1 \cdot a$ is $s_3 \vee s_4$, whereas the uncertainty of $s_1 \odot a$ is only s_4. So the real actions have an extra significance: they also change the uncertainty of the states.

Both the potential and real changes are actions of Q on M, so they both have right adjoints. We abuse the notation and denote both of these right adjoints

by the squared bracket notation [ ]. In practice, there are no ambiguities, as the properties that we are interested in only involve the right adjoints of the real changes, that is, $- \odot q \dashv [q]-$ and $- \odot^c q \dashv [q]^c-$. The potential changes and their converses (and not their right adjoints) are used in the proofs of these properties.

To encode the uncertainties, we use endomorphisms of the system. We read $u^M(m)\colon M \to M$ as the uncertainty about proposition m, the join of all propositions that are possibly true when in reality m is true. For example $u^M(m) = m \vee m'$, says that in reality m is true, but the agent considers it possible that either m or m' might be true. Similarly, we read $u^Q(q)\colon Q \to Q$ is the uncertainty about action q, the join of all actions that are possibly happening when in reality action q is happening. For example, $u^Q(q) = q \vee q'$ says that in reality action q is happening but the agent considers it possible that either q or q' is happening.

The real action $-\odot q$ changes the uncertainty of a proposition m via an *uncertainty reduction* axiom. The intuition behind it is as follows: when one does actions in reality, they change our uncertainty. In navigation systems this change is as follows: the uncertainty after performing an action in reality $u^M(m \odot q)$ is the uncertainty of performing a potential action according to the description of the system, i.e. $u^M(m \cdot q)$ minus the choices which one could not have reached via a q action (according to the description). Note that, in our navigation applications there are no uncertainties about actions, hence u^Q's are identities. Nonetheless we introduce them to keep our setting general and to be able formally to compare it to that of previous work. For example, $u^M(m \cdot q)$ can be a choice of $m' \vee m''$ and it is not possible to reach m' via a q action, i.e. $m' \cdot^c q = \bot$. Hence m' is removed from the choices in $u^M(m \odot q)$, hence $u^M(m \odot q) = m''$. The formal expression of this reasoning is the following axiom:

$$(*) \qquad u^M(m \odot q) \leq \bigvee \left\{ m' \in M \mid m' \leq u^M(m \cdot q), \quad m' \cdot^c u^Q(q) \neq \bot \right\}$$

Lemma 2. *In any converse di-system, if for all $m \in M, q \in Q$ and a prime x we have $x \leq m \cdot q$, then there exists a prime $y \leq m$ such that $x \leq y \cdot q$.*

Proof. Suppose not, i.e. for all $y \leq m$ we have $x \not\leq y \cdot q$, hence $x \not\leq (\bigvee y) \cdot q$, hence we get a contradiction $x \not\leq m \cdot q$, since every m is the join of the primes below it. □

Proposition 11. *In any converse di-system the following holds:*

$$m \wedge (m \cdot^c q \cdot q) = \bigvee \{m' \in M \mid m' \leq m, m' \cdot^c q \neq \bot\}$$

for all $m \in M, q \in Q$.

Proof. If the lhs is $\bot$, then it is easy to see that the rhs is also $\bot$ and the other way around, so suppose not. That is, there is a prime $x \leq m \wedge (m \cdot^c q \cdot q)$, hence $x \leq m$ and $x \leq m \cdot^c q \cdot q$ by the definition of meet. From the latter and the above lemma 2 we obtain that there is a prime $y \leq m \cdot^c q$ such that $x \leq y \cdot q$, hence $y \leq x \cdot^c q$ by our converse axiom (*i*), hence $x \cdot^c q \neq \bot$, hence $x \leq \bigvee \{m' \in M \mid m' \leq m, m' \cdot^c q \neq \bot\}$. For the other direction, take a prime x to be less than or equal the rhs, hence $x \leq m$ and $x \cdot^c q \neq \bot$, that is there is a prime y such that $y \leq x \cdot^c q$, and by converse axiom (*i*) we have $x \leq y \cdot q$. Now since $x \leq m$ by order preservation of action we have that $x \cdot^c q \leq m \cdot^c q$, by transitivity since $y \leq x \cdot^c q$ we obtain that $y \leq m \cdot^c q$, again by order preservation of action we obtain $y \cdot q \leq m \cdot^c q \cdot q$ and since $x \leq y \cdot q$ by transitivity we obtain $x \leq m \cdot^c q \cdot q$. □

A direct corollary of this proposition is that our above $(*)$ axiom is equivalent to the following (more algebraic) one:

$$u^M(m \odot q) \leq u^M(m \cdot q) \ \wedge \ (u^M(m \cdot q) \cdot^c u^Q(q) \cdot u^Q(q))$$

by taking m to be $u^M(m \cdot q)$ and q to be $u^Q(q)$.

Definition 3.2. A **lax endomorphism** u of a navigation pre di-system consists of a pair of endomorphisms $u = (u^M \colon M \to M, u^Q \colon Q \to Q)$, where u^M preserves joins of M and u^Q preserves joins of Q, moreover we have

$$u^M(m \odot q) \leq u^M(m \cdot q) \ \wedge \ (u^M(m \cdot q) \cdot^c u^Q(q) \cdot u^Q(q)) \tag{1}$$

$$u^Q(q \bullet q') \leq u^Q(q) \bullet u^Q(q') \tag{2}$$

$$1 \leq u^Q(1) \tag{3}$$

The reason the endomorphism axioms are inequalities has been motivated in Sadrzadeh (2006). In a nutshell, this is done in order to be able to encode the process of learning as a decrease in the uncertainty, hence an increase in information. The other two inequalities are for coherence of uncertainty with regard to composition, the motivations for these are as in Baltag et al. (2007).

Putting it all together, we define:

Definition 3.3. A **navigation di-system (Nav-diSys)** is a navigation pre di-system
$((M, Q, \cdot, \cdot^c), \odot, \odot^c, u)$ endowed with a di-system lax endomorphism $u = (u^M, u^Q)$.

Example 7. The transition system of example 5 is modeled in the following Nav-diSys

$$((\mathcal{P}(\Sigma), \mathcal{P}(A^*), \cdot, \cdot^c), \odot, \odot^c, u)$$

Here, Σ is obtained by closing the set of states S under product with A, i.e. $\Sigma := \bigcup_i S \times A^i$. So it contains states $s \in S$, pairs of states and actions $(s, a) \in S \times A$, pairs of pairs of states and actions $((s, a), b) \in (S \times A) \times A$) and so on. The potential action on states $s \cdot a$ is given by the transitions. This is extended to pairs by consecutive application of the action, i.e. $(s, a) \cdot b$ is given by $(s \cdot a) \cdot b$. The pairs encode the real actions, i.e. $s \odot a := (s, a)$, $(s \odot a) \odot b := ((s, a), b)$, ... for the atoms and extend it to all the other elements pointwisely, e.g. $s \odot (a \vee b) := (s \odot a) \vee (s \odot b)$ and $s \odot (a \bullet b) := ((s, a), b)$. Since real actions cannot be reversed, their corresponding converse action is taken to be the same as the converse of the potential action, i.e. for all actions a and states s, we have that $s \odot^c a = s \cdot^c a$. The converse of real action $\odot^c$, is introduced for reasons of symmetry with the real action, so that we can uniformly use their right adjoints to express the logical properties "after" and "before".

The lax di-system endomorphism on the module u^M are determined by indistinguishability of states as follows: s, s' are indistinguishable iff the same action a can be performed on them. In formal terms

$$u^M(s) := \bigvee \{s' \in M \mid \forall a \in A, \quad s \cdot a \neq \bot \quad \text{iff} \quad s' \cdot a \neq \bot\}$$

The u^M of the states updated by potential actions is the u^M of the image, i.e. for the transition system of the introduction, we have $u^M(s_1 \cdot a) = u^M(s_4) = s_3 \vee s_4$. The u^M of states updated by the real action is determined by inequality (1) of definition 3.2, e.g. $u^M(s_1 \odot a) = u^M(s_1, a) \leq s_4$. The uncertainties of actions, i.e. u^Q, can be set similar to that of states: by indistinguishability under application to states. Since for our navigation applications these do not play a crucial role, we assume them to be the identity, i.e. $u^Q(q) = q$ for all $q \in \mathcal{P}(A^*)$.

Finally, recall that since each projection of u is join preserving, it has a Galois right adjoint, we focus on the right adjoint of u^M, which we denote by the epistemic modality $\Box$. This is canonically defined as follows

$$\Box m := \bigvee \{m' \in M \mid u^M(m') \leq m\}$$

We read $\Box m$ as 'according to the information available m holds in reality'. Alternatively, one can use the belief modality of doxastic logic and read it as 'it is believed that, or the (sole) agent believes that, m holds in reality'. Putting these modalities together with the dynamic ones, we can express properties such as $[q]\Box m$, read as "after action q the agent believes that m holds", and such as $[q]\Box[q]^c m$, read as "after action q the agent believes that before action q proposition m held", and so on.

4 Applications to Navigation

4.1 Map-based Navigation

In the protocols of this section, the agent has a map of a place, it moves accordingly to be able to find out where it is, hence if it already knows where it is, there is no reason to move. Consider the navigation protocol of introduction, we encode it in a Nav-diSys with the set of locations $S = \{s_1, s_2, s_3, s_4, s_5\}$, the set of actions $Ac = \{a, b, c\}$ and show that after doing an a action on s_1, the robot knows where it is and where it was before moving.

Proposition 12. *The following hold in a Nav-diSys $\mathcal{N}$ based on the above data.*

$$s_1 \leq [a]\Box s_4 \qquad\qquad s_1 \leq [a]\Box[a]^c s_1$$

Proof. Consider the first one: by the adjunction $- \odot a \dashv [a]-$, it is equivalent to $s_1 \odot a \leq \Box s_4$. By the adjunction $u^M \dashv \Box$, this is equivalent to $u^M(s_1 \odot a) \leq s_4$. Now by the uncertainty reduction inequality, it suffices to show that

$$u^M(s_1 \cdot a) \wedge u^M(s_1 \cdot a) \cdot^c a \cdot a \leq s_4$$

By the assumptions we have that

$$u^M(s_1 \cdot a) = u^M(s_4) = s_3 \vee s_4$$

also that

$$u^M(s_1 \cdot a) \cdot^c a \cdot a = (s_3 \vee s_4) \cdot^c a \cdot a = (\bot \vee s_1) \cdot a = s_1 \cdot a = s_4$$

Hence the lhs of the above is $(s_3 \vee s_4) \wedge s_4$, less than or equal to s_4. Similarly, the second inequality becomes equivalent to $u^M(s_1 \odot a) \odot^c a \leq s_1$, by a series of 3 unfoldings of adjunctions. We have shown that $u^M(s_1 \odot a) \leq s_4$, so it suffices to show $s_4 \odot^c a \leq s_1$, which is true since $s_4 \odot^c a = s_4 \cdot^c a = s_1 \leq s_1$. □

4.2 Staircase Navigation

Navigating on the staircase is one of the simplest cases of robot navigation: if the robot is anywhere except for the first and last floor, it does not know where it is. But if it moves to any of these location, it learns where it is and was before moving. We model an n-floor stair case for $n \in \mathbb{N}$, by a set of locations $S = \{f_n \mid n \in \mathbb{N}\}$. The atomic actions available to the robot are $Ac = \{up, down\}$.

$$f_1 \underset{\text{down}}{\overset{\text{up}}{\rightleftarrows}} f_2 \underset{\text{down}}{\overset{\text{up}}{\rightleftarrows}} \cdots \underset{\text{down}}{\overset{\text{up}}{\rightleftarrows}} f_{n-1} \underset{\text{down}}{\overset{\text{up}}{\rightleftarrows}} f_n$$

The floors f_2 to f_{n-1} are indistinguishable from one another, i.e. for $1 < i < n$, we have $u^M(f_i) = \bigvee_{1<i<n} f_i$, the first and last floor and the actions have no uncertainty.

Proposition 13. *The following hold in a Nav-diSys $\mathcal{N}$ based on the above data*

$$f_k \leq [up^{n-k}]\Box f_n \qquad f_k \leq [up^{n-k}]\Box[up^{n-k}]^c f_k$$

$$f_k \leq [down^{k-1}]\Box f_1 \qquad f_k \leq [down^{k-1}]\Box[down^{k-1}]^c f_k$$

$$f_k \leq [up^{n-k}]\Box[down^{n-k}]f_k \qquad f_k \leq [down^{k-1}]\Box[up^{k-1}]f_k$$

for $1<n<k$, where $up^{n-k} = \underbrace{up \bullet \cdots \bullet up}_{n-k}$ and similarly for others.

Proof. Consider the first one, by the adjunction $- \odot q \dashv [q]-$ it's equivalent to

$$f_k \odot (\underbrace{up \bullet \cdots \bullet up}_{n-k}) \leq \Box^M f_n$$

equivalent to the following by the associativity of $\odot$ over $\bullet$

$$f_k \odot \underbrace{up \odot \cdots \odot up}_{n-k} \leq \Box^M f_n$$

equivalent to the following by the adjunction $u^M \dashv \Box^M$

$$u^M(f_k \odot \underbrace{up \odot \cdots \odot up}_{n-k}) \leq f_n$$

By the uncertainty reduction axiom it suffices to show

$$u^M(f_k \cdot \underbrace{up \cdot \cdots \cdot up}_{n-k}) \wedge u^M(f_k \cdot \underbrace{up \cdot \cdots \cdot up}_{n-k}) \cdot^c (\underbrace{up \bullet \cdots \bullet up}_{n-k}) \cdot (\underbrace{up \bullet \cdots \bullet up}_{n-k}) \quad \leq \quad f_n$$

The left hand side above is equal to f_n, since

$$u^M(f_k \cdot \underbrace{up \cdot \cdots \cdot up}_{n-k}) = u^M(f_n) = f_n$$

and also that

$$f_n \cdot^c (\underbrace{up \cdot \cdots \cdot up}_{n-k}) \cdot (\underbrace{up \cdot \cdots \cdot up}_{n-k}) = f_n$$

Consider the second inequality, by adjunction and definition 3.1 it is equivalent to

$$u^M(f_k \odot \underbrace{up \odot \cdots \odot up}_{n-k}) \odot^c (\underbrace{up \odot^c \cdots \odot^c up}_{n-k}) \leq f_k$$

Since in the first inequality we have shown that $u^M(f_k \odot \underbrace{up \odot \cdots \odot up}_{n-k}) \leq f_n$, it suffices to show that $f_n \odot^c (\underbrace{up \odot^c \cdots \odot^c up}_{n-k}) \leq f_k$. This is indeed the case since $f_n \odot^c (\underbrace{up \odot^c \cdots \odot^c up}_{n-k}) = f_k$. The other inequalities are proven similarly. □

4.3 Grid Navigation

A more complex robot navigation protocol happens on the grid: a robot is in a grid with n rows and m columns, it can go up, down, left, and right and is supposed to move about and find out where it is. The grid cells look alike to it as long as it can do the same movements in them, hence it knows where it is iff it ends up in one of the four corner cells. We model this protocol in a Nav-diSys and show that no matter where the robot is, there is always some sequence of movements that it can do to get it to one of the corners. After doing either of these it learns where it is and where it was beforehand.

Each grid cell is modeled by a state s_{ij} in the i'th row and j'th column. Uncertainty of corner states $s_{11}, s_{1m}, s_{n1}, s_{nm}$ is identity, i.e.

$$u^M(s_{11}) = s_{11} \quad u^M(s_{1m}) = s_{1m} \quad u^M(s_{n1}) = s_{n1} \quad u^M(s_{nm}) = s_{nm}$$

For the non-corner cells of the first row and first column, we have

$$u^M(s_{1j}) = \bigvee_{1<y<m} s_{1y} \qquad u^M(s_{i1}) = \bigvee_{1<x<n} s_{x1}$$

For the non-corner cells of last row n and last column m , we have

$$u^M(s_{nj}) = \bigvee_{1<y<m} s_{ny} \qquad u^M(s_{im}) = \bigvee_{1<x<n} s_{xm}$$

For the rest of the cells we have $u^M(s_{ij}) = \bigvee_{\substack{1<x<n \\ 1<y<m}} s_{xy}$. The set of actions is $Ac = \{u, d, l, r\}$, their non-applicability is as follows

$$s_{1j} \cdot u = s_{1j} \cdot^c d = s_{i1} \cdot l = s_{i1} \cdot^c r = s_{nj} \cdot d = s_{nj} \cdot^c u = s_{im} \cdot r = s_{im} \cdot^c l = \bot$$

All the other actions are applicable in all the other states.

Proposition 14. *The following hold in a Nav-diSys $\mathcal{N}$ based on the above data.*

$$s_{ij} \leq [\alpha]\Box(s_{11} \vee s_{1m} \vee s_{n1} \vee s_{nm}) \qquad s_{ij} \leq [\alpha]\Box[\alpha]^c s_{ij}$$

for $1 < i < n, 1 < j < m$ and α the following choices of sequences of movements

$$(u^{i-1} \vee d^{n-i}) \bullet (l^{j-1} \vee r^{m-j}) \vee (l^{j-1} \vee r^{m-j}) \bullet (u^{i-1} \vee d^{n-i})$$

Proof. By adjunctions $- \odot q \dashv [q]-$ and $u^M \dashv \Box$ the first inequality is equivalent to:

$$u^M(s_{ij} \odot \alpha) \leq s_{11} \vee s_{1m} \vee s_{n1} \vee s_{nm}$$

By join preservation of $\odot$ and u^M, the above becomes equivalent to showing a join of 8 terms on the left to be less than or equal to the a join of 4 locations on the right. So by definition of join, we must show that all of the following 8 cases hold

$$u^M(s_{ij} \odot (u^{i-1} \bullet l^{j-1})) \leq s_{11} \qquad u^M(s_{ij} \odot (u^{i-1} \bullet r^{m-j})) \leq s_{1m}$$
$$u^M(s_{ij} \odot (d^{n-i} \bullet l^{j-1})) \leq s_{n1} \qquad u^M(s_{ij} \odot (d^{n-i} \bullet r^{m-j})) \leq s_{nm}$$
$$u^M(s_{ij} \odot (l^{j-1} \bullet u^{i-1})) \leq s_{11} \qquad u^M(s_{ij} \odot (l^{j-1} \bullet d^{n-i})) \leq s_{n1}$$
$$u^M(s_{ij} \odot (r^{m-j} \bullet u^{i-1})) \leq s_{1m} \qquad u^M(s_{ij} \odot (r^{m-j} \bullet d^{n-i})) \leq s_{nm}$$

Consider the first one, by uncertainty reduction and associativity of action and quantale multiplication it suffices to show that

$$u^M(s_{ij} \cdot u^{i-1} \cdot l^{j-1}) \ \wedge \ u^M(s_{ij} \cdot u^{i-1} \cdot l^{j-1}) \cdot^c (u^{i-1} \bullet l^{j-1}) \cdot (u^{i-1} \bullet l^{j-1}) \quad \leq \quad s_{11}$$

By the grid assumptions we have that

$$u^M(s_{ij} \cdot u^{i-1} \cdot l^{j-1}) = u^M(s_{11}) = s_{11}$$

also that

$$s_{11} \cdot^c (u^{i-1} \bullet l^{j-1}) \cdot (u^{i-1} \bullet l^{j-1}) = s_{11}$$

Hence the left hand side is $\leq s_{11}$. Proofs of the other 7 inequalities are similar.

The second property $s_{ij} \leq [\alpha]\Box[\alpha]^c s_{ij}$ is equivalent to $u^M(s_{ij} \odot \alpha) \odot^c \alpha \leq s_{ij}$ by adjunction. Like above $u^M(s_{ij} \odot \alpha)$ breaks down to 8 terms and since $\odot^c$ is also join preserving, one has to show 8×8 inequalities similar to those in the above, but updated with the $\odot^c$ of the 8 combinations of composition of actions on the left and s_{ij} on the right. That is, we have 8 inequalities of the form

$$u^M(s_{ij} \odot (u^{i-1} \bullet l^{j-1})) \odot^c ((u^{i-1} \vee d^{n-i}) \bullet (l^{j-1} \vee r^{m-j}) \vee (l^{j-1} \vee r^{m-j}) \bullet (u^{i-1} \vee d^{n-i})) \leq s_{ij}$$

We have shown that $u^M(s_{ij} \odot (u^{i-1} \bullet l^{j-1})) = s_{11}$, so the above is equivalent to

$$s_{11} \odot^c ((u^{i-1} \vee d^{n-i}) \bullet (l^{j-1} \vee r^{m-j}) \vee (l^{j-1} \vee r^{m-j}) \bullet (u^{i-1} \vee d^{n-i})) \leq s_{ij}$$

To show the above, one must show all of the following 8 inequalities

$$s_{11} \odot^c (u^{i-1} \bullet l^{j-1}) \leq s_{ij} \qquad s_{11} \odot^c (u^{i-1} \bullet r^{m-j}) \leq s_{ij}$$
$$s_{11} \odot^c (d^{n-i} \bullet l^{j-1}) \leq s_{ij} \qquad s_{11} \odot^c (d^{n-i} \bullet r^{m-j}) \leq s_{ij}$$
$$s_{11} \odot^c (l^{j-1} \bullet u^{i-1}) \leq s_{ij} \qquad s_{11} \odot^c (l^{j-1} \bullet d^{n-i}) \leq s_{ij}$$
$$s_{11} \odot^c (r^{m-j} \bullet u^{i-1}) \leq s_{ij} \qquad s_{11} \odot^c (r^{m-j} \bullet d^{n-i}) \leq s_{ij}$$

For two of these we have $s_{11} \odot^c (d^{n-i} \bullet r^{m-j}) = s_{11} \odot^c (r^{m-j} \bullet d^{n-i}) = s_{ij}$, and the rest are equal to $\perp$ (by the inapplicability assumptions of the grid), which is fine since $\perp$ is less than or equal anything, in particular s_{ij}. □

5 Conclusions and future work

We have developed an algebraic framework for dynamic epistemic logic in which the dynamic and epistemic modalities appear as right adjoints. The key new feature in the present work relative to previous work Sadrzadeh (2006), Baltag et al. (2007) is the presence of converse actions and the algebraic laws that govern uncertainty reduction. The navigation protocols discussed in this paper, as well as the three-player game in Phillips's thesis Phillips (2009), give examples in which the old learning inequality was violated, showing that there were new subtleties that arise when there are actions that really change the state of the world.

A number of directions for future work naturally suggest themselves. On the purely theoretical side, we would like to relate boolean converse di-systems to Kleene algebras with test and converse Desharnais et al. (2006). A logic for NavdiSys would be based on the positive fragment of propositional dynamic logic with assignments Tiomkin and Makowsky (1985) and with converse Parikh (1978), extended with epistemic modalities and related axioms. Capturing the same reasoning in dynamic epistemic logic, e.g. with assignment van Benthem et al. (2006), van Ditmarsch et al. (2005) or converse Aucher and Herzig (2007) is worth further investigation. We are also particularly interested in extending this work to apply to examples that involve security protocols where "belief" and "learning" play evident roles. A fundamental extension, and one in which we have begun preliminary investigations, is the extension to the probabilistic case. Here belief and information theory may well merge in an interesting and not obvious way.

Acknowledgements

We have benefited greatly from discussions with Caitlin Phillips and Doina Precup. The latter invented the three-player game and the former discovered the violation of the update inequality. We also thank Mai Gehrke and Sam

van Gool for invaluable comments and discussions regarding properties of the converse action. This research was supported by EPSRC (UK) and NSERC (Canada).

References

S. Abramsky and S. Vickers. Quantales, observational logic and process semantics. *Mathematical Structures in Computer Science*, 3(2):161–227, 1993.

G. Aucher and A. Herzig. From del to edl : Exploring the power of converse events. In K. Mellouli, editor, *ECSQARU*, volume 4724 of *Lecture Notes in Computer Science*, pages 199–209. Springer, 2007. ISBN 978-3-540-75255-4.

A. Baltag. A logic for suspicious players: Epistemic actions and belief updates in games. *Bulletin of Economic Research*, 54:1–46, 2002.

A. Baltag and L. S. Moss. Logics for Epistemic Programs. *Synthese*, 139:165–224, 2004.

A. Baltag, R. Baltag, B. Coecke, and M. Sadrzadeh. Epistemic actions as resources. *Journal of Logic and Computation*, pages 555–585, 2007.

J. Desharnais, B. Möller, and G. Struth. Kleene algebra with domain. *ACM Trans. Comput. Log.*, 7(4):798–833, 2006.

H. v. Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*. Springer Publishing Company, Incorporated, 1st edition, 2007.

J. M. Dunn. Positive modal logic. *Studia Logica*, 55(2):301–317, 1995.

R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning About Knowledge*. MIT Press, 1995.

M. Gehrke and B. Jónsson. Bounded distributive lattices with operators. *Math. Japon.*, 40(2):207–215, 1994.

J. Gerbrandy and W. Groeneveld. Reasoning about information change. *Journal of Logic, Language and Information*, 6(2):147–169, 1997.

J. Y. Halpern and Y. Moses. Knowledge and common knowledge in a distributed environment. *CoRR*, cs.DC/0006009, 2000.

A. Horn. Dynamic epistemic algebra with post-conditions to reason about robot navigation. In L. D. Beklemishev and R. de Queiroz, editors, *WoLLIC*, volume 6642 of *Lecture Notes in Computer Science*, pages 161–175. Springer, 2011.

B. V. Karger. Temporal algebra. *Mathematical Structures in Computer Science*, pages 32–0, 1996.

S. A. Kripke. Semantical Analysis of Modal Logic I Normal Modal Propositional Calculi. *Mathematical Logic Quarterly*, 9(5-6):67–96, 1963.

R. Parikh. The completeness of propositional dynamic logic. In J. Winkowski, editor, *MFCS*, volume 64 of *Lecture Notes in Computer Science*, pages 403–415. Springer, 1978.

C. Phillips. An algebraic approach to dynamic epistemic logic. Master's thesis, School of Computer Science; McGill University, 2009.

J. Plaza. Logics of public communications. *Synthese*, 158(2):165–179, Sept. 2007.

T. Ågotnes. Action and Knowledge in Alternating-Time Temporal Logic. *Synthese*, 149(2), 2006.

M. Sadrzadeh. *Actions and Resources in Epistemic Logic*. PhD thesis, Université du Québec à Montréal, 2006.

M. L. Tiomkin and J. A. Makowsky. Propositional dynamic logic with local assignments. *Theor. Comput. Sci.*, 36:71–87, 1985.

J. van Benthem, J. van Eijck, and B. Kooi. Logics of communication and change. *Inf. Comput.*, 204(11):1620–1662, Nov. 2006.

H. P. van Ditmarsch, W. van der Hoek, and B. P. Kooi. Dynamic epistemic logic with assignment. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, AAMAS '05, pages 141–148, New York, NY, USA, 2005. ACM.

G. Winskel. Prime algebraicity. *Theor. Comput. Sci.*, 410(41):4160–4168, 2009.

Reasoning about Multiagent Resource Allocation in Linear Logic

Daniele Porello

Institute for Logic, Language & Computation (ILLC)
danieleporello@gmail.com

Abstract

In this paper, we present a general treatment of multiagent resource allocation (MARA) by means of the proof-theoretical insights provided by linear logic. We will see how linear logic provides a versatile language to express agents' preferences over multisets of goods. Moreover, we will interpret the problem of finding an optimal allocation as a proof-search problem in suitable fragments of linear logic. We apply this approach to two well known paradigms in MARA, combinatorial auctions and multilateral negotiation. This presentation is based on two joint papers with Ulle Endriss, Porello and Endriss (2010b) and Porello and Endriss (2010a).

1 Introduction

Multiagent resource allocation is generally intended as the process of distributing a number of resources to a number of agents, Chevaleyre et al. (2006). The type of resources and the type of agents may vary according to the domain

of application of MARA, as similar situations occur in several problems at the interface of AI and Economics. For example, in cooperative problem solving, we need to find an allocation of resources to agents that will allow each agent to complete the tasks she has been assigned; in the context of electronic commerce applications, the system objectives will often be defined in terms of properties of the allocations of resources that are being negotiated. Studies of resource allocation in AI may range from the design of negotiation strategies, over the game-theoretical analysis of allocation problems, to the complexity-theoretic study of relevant optimisation problems. Moreover, abstract frameworks for the precise representation and formal study of systems for MARA are important to deal with problems of knowledge representation and reasoning concerning a certain domain of objects. Logic is an important tool for this purpose, and there have been a number of contributions of this kind, cf. Endriss and Pacuit (2006), Fisher (2000), Harland and Winikoff (2002), Küngas and Matskin (2004), Leite et al. (2009), Porello and Endriss (2010b), Sadri et al. (2002).

In this paper, we develop a framework for MARA within *linear logic*, Girard (1987), a constructive approach to logic that provides a resource sensitive account of proofs. In particular, we will show how to apply proof-theoretical methods to embed a framework for *combinatorial auctions*, cf. Nisan (2006), and *distributed resource allocation*, cf. Chevaleyre et al. (2010), Endriss et al. (2006). The first reason that motivates the use of linear logic is that it allows for an intuitive representation of agents' preference over *multisets* of goods. Standard applications of MARA usually implicitly refer to goods that are available in multi-units, namely, where several indistinguishable copies of a same item are traded by agents; however no general and logically grounded approach to preferences defined over multisets of resources has been developed.

Moreover, the aim of our approach is to interpret preference satisfaction, allocations of resources, and deals between agents concerning allocations, as reasoning tasks. In particular, we will show how to interpret several problems in MARA as proof-search in linear logic.

Several authors have recognised that, due to its resource-sensitive nature, linear logic is particularly suited to modelling resource allocation problems; cf. Harland and Winikoff (2002), Küngas and Matskin (2004). Two contributions on logic-based approaches to resource allocation relate to the same kind of resource allocation framework we shall be working with here: Endriss and Pacuit (2006), develop a modal logic to study the convergence problem in distributed resource allocation; and Leite et al. (2009) show how to map the problem of finding an

allocation that is socially optimal (for a wide variety of fairness and efficiency criteria) into the framework of answer-set programming.

This paper is organised as follows. In Section 2, we present the relevant background on MARA, by reviewing combinatorial auctions, distributed negotiation, and by discussing them as reasoning tasks. Section 3 contains the essential background on linear logic. Section 4 introduces our model for representing preferences over multisets of goods. Section 5 presents our modelling of combinatorial auctions and the relationship between allocations and proofs in linear logic. Section 6 contains our model of distributed negotiation. Section 7 presents some possible extension and concludes.

2 Multiagent Resource Allocation

Assume agents have to allocate bundles of goods from the set $\mathcal{S}$ and let $\mathcal{W}$ be a set of values that the agents use to express their cardinal preferences. In the standard approach, the set of values $\mathcal{W}$ is usually $\mathbb{R}$. A *valuation* over bundles in $\mathcal{S}$ is a function $v : \mathcal{P}(\mathcal{S}) \to \mathbb{R}$. An *allocation* of goods to agents is a function $\alpha : \mathcal{S} \to \mathcal{N} \cup \{*\}$ from resources to agents; we indicate for each item who receives it or whether it does not get allocated at all ($*$): let $A_i = \alpha^{-1}(i)$ and $\alpha^{-1}(*)$ are the unallocated goods. The *winner determination problem* is to find an allocation that maximizes the value of an allocation, according to the definition of the value of α. In the next two paragraphs, we shall substantiate this definitions by applying them to combinatorial auctions and multilateral negotiation.

2.1 Combinatorial Auctions

A *combinatorial auction* (CA) is a (centralized) mechanism for one agent (the *auctioneer*) to sell goods to a number of other agents. In CAs, agents usually express their valuations over bundles of goods by means of a *bidding language*. Thus, agents submit bid expressions BID that specifies the price they are willing to pay for a given bundle of goods. The value of an allocation α is usually given by the sum of the prices associated to the satisfied bids: $v(\alpha) = \sum_i \{v_{\text{BID}_i}(A_i)\}$ The value of an allocation specifies the revenue that the auctioneer receives.

An *atomic bid* is an expression BID $= (S, w)$ where $S \subseteq \mathcal{S}$ and $w \in \mathcal{W}$ is the price associated to S. Each atomic bid defines a valuation $v_{\text{BID}} : \mathcal{P}(S) \to \mathcal{W}$ by means

of the following definition that specifies the semantics of a bid.

$$v_{\text{BID}}(S') = \begin{cases} w & \text{if } S \subseteq S' \\ 0 & \text{otherwise} \end{cases}$$

Several languages to encode valuations have been developed in the literature on CA. Here, we present three representative bidding languages.

Given a set of bids (S_1, w_1), ..., (S_l, w_l), an *XOR-bid* is defined as follows:

$$(S_1, w_1) \text{ XOR} \ldots \text{XOR}(S_l, w_l)$$

The intended meaning of such an expression is that bidders are willing to obtain at most one of these bids at its specified price. Thus, the semantics of an *XOR-bid* is defined as follows as follows:

$$v_{XOR}(S) = max\{v(S_i) \ : \ S_i \subseteq S\}$$

XOR-bids generate all valuations, as basically we can express the graph of a function. However, the size of the representation can be exponential.

An *OR-bid* is defined as follows:

$$(S_1, w_1) \text{ OR} \ldots \text{OR}(S_l, w_l)$$

The intended meaning of an OR-bid is that bidders are willing to obtain any number of disjoint atomic bids for the sum of their respective prices. Thus, the semantics of an OR-bid is the following:

$$v_{OR}(S) = max \sum_i \{v(S_i) \ : \ S_i \in F\}$$

where F is a collection of bids such that for all $i \neq j$, $S_i \cap S_j = \emptyset$ and $S_i \subseteq S$. The expressive power of OR-bids is characterized by the class of valuations such that $v(X \cup Y) \geq v(X) + v(Y)$, whenever $X \cap Y = \emptyset$, cf. Nisan (2006).

Next, we introduce the *goalbase* languages, a family of languages based on weighted propositional formulas; cf. Uckelman et al. (2009), Uckelman et al.

(2009). Such languages have been widely studied in the AI literature; for the specific use in CAs they have first been proposed by Boutilier and Hoos (2001). A *goalbase* G is a set of pairs (φ_i, w_i), where φ_i is a proposition (in classical logic) and w_i is a weight. G induces a valuation that maps any assignment of truth values to atoms to the sum of the weights of the formulas that are satisfied by that assignment (which we can think of as a bundle of goods):

$$v_{GB}(S) = \sum_i \{w_i \mid S \models \varphi_i\}$$

By restricting the language of the weighted proposition and the set of weights, goalbases characterize important classes of functions. In particular, the class of k-additive functions is proved to be equivalent to the class of functions generated by goalbases of *positive cubes*, i.e., conjunctions of positive literals, $(p_1 \wedge \cdots \wedge p_\ell, w)$, cf. Uckelman et al. (2009).

The language of *k-additive valuations*, cf. Chevaleyre et al. (2008), is based on the idea of specifying weights for the *marginal* valuations derived from sets of goods, rather than directly specifying the values of full bundles.

A valuation v is called *k-additive* if there exists a mapping $v' : \mathcal{S}[k] \to \mathbb{R}$, where $\mathcal{S}[k]$ denotes the set of all subsets with at most k elements, such that $v(X) = \sum\{v'(Y) \mid Y \subseteq X \text{ and } Y \in \mathcal{S}[k]\}$.

The notion of k-additivity gives rise to a bidding language: by specifying a (marginal, possibly negative) price for each bundle of size $\leq k$ (as an atomic bid) we can represent v' and thus v. An important difference between the OR language and goalbase languages (including k-additive languages) is that the accepted atomic bids may overlap. For example, in $G = \{(p \wedge q, 5), (p, 3)\}$, the allocation of p and q will satisfy both atomic bids.

2.2 Multilateral Negotiation

Combinatorial auctions provides a centralized mechanism for finding optimal allocations by satisfying an optimal combination of bids that the agents submit. Multilateral negotiation provides a distributed approach to the allocation problem where agents endowed with valuations over bundles of goods exchange deals in order to reach an optimal allocation, cf. Chevaleyre et al. (2010), Endriss et al. (2006). A *deal* takes us from one allocation to the next; i.e., we can think of

it as a pair of allocations. Note that there are no restrictions as to the number of agents or resources involved in a single deal. Of special interest are structurally simple deals: for instance, *1-deals* are deals involving the reassignment of a single resource only.

Given an allocation α, we will concentrate on two economic efficiency criteria, cf. Chevaleyre et al. (2006): (1) the (utilitarian) *social welfare* of an allocation α is defined as $sw_u(\alpha) = \sum_{i \in N} v_i(A_i)$ and we shall be interested in finding allocations that maximise social welfare; (2) an allocation is *Pareto optimal* if no other allocation gives higher valuation to some agents without giving less to any of the others (this is a considerably less demanding criterion).

What kinds of allocations can be reached from a given initial allocation depends on the range of deals we permit. First, we may ask whether a desirable allocation A' is reachable from the initial allocation A by means of a given class of deals. A very simple result shows that, quite clearly, the class of all 1-deals is always sufficient to take us from any A to any A' (Sandholm 1998, Prop. 1).

A deal is called *individually rational* if it is possible to arrange side payments for the agents involved such that for each agent her gain in valuation outweighs her loss in money (or her gain in money outweighs her loss in valuation). The payments of all agents need to add up to 0. In case we don't allow for side payments, rational deals are modelled by *cooperative rationality*: agents agree to a deal, if this at least maintains their utility and at least one agent strictly increases his utility.

2.3 Reasoning about types of goods

The goalbase languages show how to view the satisfaction of a demand as a certain type of reasoning. In particular, the matching between demand and offer is modelled as logical inference. By viewing goods as propositional atoms and preferences over bundles as logical formulas, we have:

$$v_{\{(a \wedge b, w)\}}(\{a, b, c\}) = w \text{ iff } \{a, b, c\} \models a \wedge b$$

However, if goods are available in multi-unit, or lists, classical entailment is problematic. This is due to structural rules of sequent calculus that basically force us to view premises as sets:

$$\frac{\{a,b\} \vdash a \wedge b}{\{a,a,b\} \vdash a \wedge b}\,\mathrm{W} \qquad \frac{\{a,a\} \vdash a \wedge a}{a \vdash a \wedge a}\,\mathrm{C} \qquad \frac{\{a,b\} \vdash a \wedge b}{\{b,a\} \vdash a \wedge b}\,\mathrm{E}$$

For example, contraction (C) shows that a single copy of a can satisfy the demand of two copies of a ($a \wedge a$). The motivation of this paper is to extend the logical treatment to other types of goods. Linear logic, as we shall see, provides a good candidate for expressing preferences over multisets or lists of goods as it is capable of controlling the application of structural rules.

3 Linear logic

Removing the structural rules from the classical sequent calculus, we are lead to split the usual connectives into two classes, since, for example, the following rules are no longer equivalent:

$$\frac{\Gamma \vdash A \qquad \Gamma' \vdash B}{\Gamma, \Gamma' \vdash A \wedge B}\,\mathrm{R}\wedge \qquad \frac{\Gamma \vdash A \qquad \Gamma \vdash B}{\Gamma \vdash A \wedge B}\,\mathrm{R}\wedge$$

Without structural rules, sequents behave as multisets of formula occurrences and we have to distinguish connectives that take the concatenation of contexts (*multiplicatives*) and connectives that demand a shared context (*additives*).

Given a set of atoms $\mathcal{A}$, the language of LL is defined as follows (where $p \in \mathcal{A}$):

$$L \;::= p \mid \mathbf{1} \mid \bot \mid \top \mid \mathbf{0} \mid L^{\bot} \mid L \otimes L \mid L \parr L \mid L \oplus L \mid L \,\&\, L \mid !L \mid ?L$$

Linear negation $(\cdot)^{\bot}$ is involutive and each formula in LL can be transformed into an equivalent formula where negation occurs only at the atomic level. The intuitive meaning of LL connective can be presented as follows. The conjunction $A \otimes B$ ("tensor") means that we have exactly one copy of A and one copy of B, no more no less. Thus, e.g., $A \otimes B \nvdash A$. We might say that in order to sell A and B, we need someone who buys A and B, while here there is just a buyer for A. We will not directly use the disjunction $A \parr B$ ("par"); rather we use linear implication: $A \multimap B := A^{\bot} \parr B$. Linear implication can be seen as a form of deal: "for A, I sell you B". The additive conjunction $A \,\&\, B$ ("with") introduces a form of choice: we have one of A and B and we can choose which one. For example, $A \,\&\, B \vdash A$, but we do not have them both: $A \,\&\, B \nvdash A \otimes B$. The

additive disjunction $A \oplus B$ ("plus") means that we have one of A and B, but we cannot choose, e.g., $A \vdash A \oplus B$ but $A \oplus B \nvdash A \,\&\, B$. The exponentials $!A$ and $?A$ reintroduce structural rules in a local way: !-formulas licence (C) and (W) on the lefthand side of $\vdash$; ?-formulas licence (C) and (W) on the right. Intuitively, exponential formulas can be copied and erased; they are relieved from their linear status.

We will use the intuitionistic version of linear logic (ILL), obtained by restricting the righthand side of the sequent to a single formula, so for example we will not have ? and ⅋ in the language. In fact, we will mostly use ILL augmented with the global weakening rule (W). The reasons for these choices will become clear later. The rules of the sequent calculus for ILL are shown in Table 1, cf. Troelstra (1992).

To control complexity, we can restrict attention to certain fragments: *intuitionistic multiplicative linear logic* (IMLL) using only $\otimes$ and $\multimap$; *intuitionistic multiplicative additive linear logic* (IMALL) using only $\otimes$, $\multimap$, $\&$ and $\oplus$; and *Horn linear logic* (HLL). In the latter, sequents must be of the form $X, \Gamma \vdash Y$ Kanovich (1994), where X and Y are tensors of positive atoms, and Γ is one of the following (with X_i, Y_i being tensors of positive atoms):

(*i*) Horn implications: $(X_1 \multimap Y_1) \otimes \cdots \otimes (X_n \multimap Y_n)$

(*ii*) &-Horn implications: $(X_1 \multimap Y_1) \,\&\, \cdots \,\&\, (X_n \multimap Y_n)$

For these fragments we can rely on the following proof-search complexity results. MLL is NP-complete and so is MLL with full weakening (W), Lincoln (1995). The same results apply for the intuitionistic versions. HLL is NP-complete, and so is HLL + W, Kanovich (1994). MALL and IMALL are PSPACE-complete, Lincoln et al. (1992).

4 Modelling preferences

There is an isomorphism between multisets and tensor formulas of atoms (up to associativity and commutativity): $\{m_1, \ldots, m_k\} \;\cong\; m_1 \otimes \cdots \otimes m_k$. Thus, we can represent each subset $X \subseteq M$ as a tensor product. Moreover, if $M \cong A$ and $N \cong B$, then the (disjoint) union of M and N is isomorphic to $A \otimes B$.

We now want to define languages to encode *valuations* $v : \mathcal{P}(\mathcal{M}) \to \mathbb{N}$, mapping

$$\frac{}{A \vdash A}\ \text{ax} \qquad \frac{\Gamma, A \vdash C \qquad \Gamma' \vdash A}{\Gamma, \Gamma' \vdash C}\ \text{cut}$$

Multiplicatives

$$\frac{\Gamma, A, B \vdash C}{\Gamma, A \otimes B \vdash C} \otimes\text{L} \qquad \frac{\Gamma \vdash A \qquad \Gamma' \vdash B}{\Gamma, \Gamma' \vdash A \otimes B} \otimes\text{R}$$

$$\frac{\Gamma \vdash A \qquad \Gamma', B \vdash C}{\Gamma', \Gamma, A \multimap B \vdash C} \multimap\text{L} \qquad \frac{\Gamma, A \vdash B}{\Gamma \vdash A \multimap B} \multimap\text{R}$$

$$\frac{\Gamma \vdash C}{\Gamma, \mathbf{1} \vdash C}\ \mathbf{1}\text{L} \qquad \frac{}{\vdash \mathbf{1}}\ \mathbf{1}\text{R}$$

Additives

$$\frac{\Gamma, A_i \vdash C}{\Gamma, A_0 \& A_1 \vdash C} \&\text{L} \qquad \frac{\Gamma \vdash A \qquad \Gamma \vdash B}{\Gamma \vdash A \& B} \&\text{R}$$

$$\frac{\Gamma, A \vdash C \qquad \Gamma, B \vdash C}{\Gamma, A \oplus B \vdash C} \oplus\text{L} \qquad \frac{\Gamma \vdash A_i}{\Gamma \vdash A_0 \oplus A_1} \oplus\text{R}$$

$$\frac{}{\Gamma, \mathbf{0} \vdash C}\ \mathbf{0}\text{L} \qquad \frac{}{\Gamma \vdash \top}\ \top\text{R}$$

Exponentials

$$\frac{\Gamma, A \vdash C}{\Gamma, !A \vdash C}\ !\text{L} \qquad \frac{!\Gamma \vdash A}{!\Gamma \vdash !A}\ !\text{R}$$

Structural Rules

$$\frac{\Gamma, A, B, \Gamma' \vdash C}{\Gamma, B, A, \Gamma' \vdash C}\ \text{E} \qquad \frac{\Gamma, !A, !A, \vdash C}{\Gamma, !A \vdash C}\ !\text{C} \qquad \frac{\Gamma \vdash C}{\Gamma, !A \vdash C}\ !\text{W} \qquad \frac{\Gamma \vdash C}{\Gamma, A \vdash C}\ \text{W}$$

Table 1: Sequent Calculus for Intuitionistic LL

subsets of $\mathcal{M}$ to prices. [1]

To model prices symbolically, we assume a finite set of distinct weight atoms $\mathcal{W} = \{w_1, ..., w_p\}$. In fact, often we will use just one weight atom u. We write u^k for the tensor product $u \otimes \cdots \otimes u$ (k times). To associate weights with numbers, we define a function $val : \mathcal{W} \to \mathbb{N}$, with $val(u) = 1$. Let $\mathcal{W}^{\otimes}$ be the set of all finite tensor products of atoms in $\mathcal{W}$, modulo commutativity (including the "empty" product $\mathbf{1}$). That is, $\mathcal{W}^{\otimes} = \{\mathbf{1}, w_1, w_2, w_1 \otimes w_2, \ldots\}$. We extend val to $\mathcal{W}^{\otimes}$ by stipulating $val(\mathbf{1}) = 0$ and $val(\varphi \otimes \psi) = val(\varphi) + val(\psi)$. In particular, this means that $val(u^k) = k$.

Definition 1. *An atomic bid is a formula of the form $B \multimap w$, where B is a tensor product of atoms in $\mathcal{A}$ and $w \in \mathcal{W}$.*

In a CA, given a bid $B \multimap w$, we can work with two alternative assumptions: *no free disposal* at the bidder's side, meaning that the bidder will pay w if she receives *exactly* B, and *free disposal* at the bidder's side, meaning that the bid is satisfied whenever the bidder receives *at least* B. In the sequel, unless otherwise stated, we will always assume free disposal. To model free disposal, we will use ILL with weakening (W). [2]

Definition 2. *Every bid formula* BID *generates a valuation v_{BID} mapping multisets $X \subseteq \mathcal{M}$ to prices:*

$$v_{\text{BID}}(X) = \max\{val(w') \mid w' \in \mathcal{W}^{\otimes} \text{ and } X, \text{BID} \vdash w'\}$$

Definition 2 applies to atomic bids as well as to the more powerful bidding languages we will define in the sequel. In the case of atomic bids $\text{BID} = (B \multimap w)$, it simply says that $v_{B \multimap w}(X) = w$ whenever X is equal to a superset of the multiset isomorphic to B, and $v_{B \multimap w}(X) = 0$ otherwise.

In case the only weight atom used is u, i.e., if $\mathcal{W} = \{u\}$, then Definition 2 can be simplified and we obtain:[3]

$$v_{\text{BID}}(X) = \max\{k \mid X, \text{BID} \vdash u^k\}$$

[1] For ease of notation, we shall assume $0 \in \mathbb{N}$.

[2] Alternatively, we could use the additive constant of linear logic $\top$ and write bids $B \otimes \top \multimap w$ to make it explicit in the syntax that a bidder has free disposal.

[3] We can define $u^0 = \mathbf{1}$. Using weakening (to represent free disposal), from $\vdash \mathbf{1}$ we get $\Gamma \vdash \mathbf{1}$, for any Γ. So every bid produces u^0, since it will always be satisfied by any allocation (also by allocating nothing), e.g., $p, p \otimes q \multimap u^k \vdash \mathbf{1}$ will be provable.

4.1 XOR/&-bids

An XOR-bid $\langle B_1, w_1\rangle \text{ XOR} \cdots \text{XOR} \langle B_\ell, w_\ell\rangle$ expresses that a bidder would like to get at most one of the bundles she specifies, for the associated price Nisan (2006). In LL, this idea can be captured via the additive conjunction (&).

Definition 3. *An XOR-bid is a formula of the form*

$$(B_1 \multimap w_1) \,\&\, \ldots \,\&\, (B_\ell \multimap w_\ell),$$

where each B_i is a tensor product of atoms in $\mathcal{A}$ and each w_i is a weight atom from $\mathcal{W}$.

Definition 2 provides the semantics for XOR-bids by fixing the valuation functions they generate.

Example 4. *Given an XOR-bid $(p \multimap u) \,\&\, (q \multimap w) \,\&\, (p \otimes q \otimes r \multimap z)$, suppose the auctioneer provides $\{p, p, q, r, s\}$. Thus, it is possible to satisfy each of the atomic bids in the XOR-bid. For example, the auctioneer can satisfy the bid producing z:*

$$\dfrac{\dfrac{\dfrac{\ldots}{p, q, r, p \otimes q \otimes r \multimap z \vdash z}}{p, p, q, r, s, p \otimes q \otimes r \multimap z \vdash z}\, W}{p, p, q, r, s, (p \multimap u) \,\&\, (q \multimap w) \,\&\, (p \otimes q \otimes r \multimap z) \vdash z}\, \&L$$

However, we have to choose which atomic bid to satisfy, according to the meaning of &.

Example 5. *We define two classes of valuation functions, adapting their definitions from Nisan (2006) to the multi-unit case. The simple additive valuation, $v(X) = |X|$ for $X \subseteq \mathcal{M}$, can be expressed via the following formula, which is exponential in size in the number of items in $\mathcal{M}$ (we slightly abuse the notation identifying the multiset B with the corresponding tensor formula): $\&_{B \subseteq \mathcal{M}}(B \multimap u^{|B|})$. The simple unit demand valuation, $v(X) = 1$ for $X \neq \emptyset$ and $v(\emptyset) = 0$, can be expressed in the XOR language via: $(p_1 \multimap u) \,\&\, \cdots \,\&\, (p_m \multimap u)$*

We say that a valuation $v : \mathcal{P}(M) \to \mathbb{N}$ is *monotonic* if and only if for all $X_1, X_2 \subseteq \mathcal{M}$, if $X_1 \subseteq X_2$, then $v(X_1) \leq v(X_2)$. Recall that we can model both free disposal or the lack thereof simply by using $\vdash$ with and without weakening (W), respectively. Following Nisan (2006) and Cerquides et al. (2007) we can easily prove that, also in our framework, the XOR-language without free disposal can express all valuations and the XOR language with free disposal is fully expressive over the space of monotonic valuations.

Proposition 6. *The following hold:*

1. *Every valuation $v : \mathcal{P}(\mathcal{M}) \to \mathbb{N}$ is generated by some XOR-bid without free disposal.*
2. *XOR-bids with free disposal generate all monotonic valuations and only those.*

4.2 OR/⊗-bids

An OR-bid $\langle B_1, w_1\rangle \text{ OR} \cdots \text{OR} \langle B_\ell, w_\ell\rangle$ states that a bidder agrees to receive any number of disjoint bundles at the sum of their prices Nisan (2006). The appropriate LL connective for modelling this kind of semantics is the tensor ($\otimes$).

Definition 7. *An OR-bid is a formula of the form*

$$(B_1 \multimap w_1) \otimes \cdots \otimes (B_\ell \multimap w_\ell),$$

where each B_i is a tensor product of atoms in $\mathcal{A}$ and each w_i is a weight atom from $\mathcal{W}$.

The intended meaning of a tensor/OR-bid is that the bidder would pay the sum of the corresponding w_i for each bundle of goods B_i she gets. The formal semantics of OR-bids is again given by Definition 2.

The usual condition on OR-bids, namely that the required bundles of goods do not overlap, works well if goods are available in single unit: since we are here considering the multi-unit case, the condition of not overlapping is replaced by imposing that the right amount of goods is provided in order to satisfy the atomic bids in the OR-bid. For example, the OR-bid $\langle p, 1\rangle \text{ OR} \langle p, 1\rangle$ will be fully satisfied only if the auctioneer provides two p. This is the meaning of the provability of a sequent containing OR-bids in Definition 2.

Example 8. *Given an OR-bid $(p \otimes q \multimap v) \otimes (q \multimap w)$, suppose the auctioneer provides $\{p, q\}$. The OR-bid can be satisfied in two possible ways:*

$$\dfrac{\dfrac{\dfrac{\dfrac{\dots}{p, q, p \otimes q \multimap v \vdash v}}{p, q, p \otimes q \multimap v, (q \multimap w) \vdash v}\, W}{p, q, (p \otimes q \multimap v) \otimes (q \multimap w) \vdash v}\, \otimes L \qquad \dfrac{\dfrac{\dfrac{\dfrac{\dots}{q, q \multimap w \vdash w}}{p, q, q \multimap w \vdash w}\, W}{p, q, (p \otimes q \multimap v), q \multimap w \vdash w}\, W}{p, q, (p \otimes q \multimap v) \otimes (q \multimap w) \vdash w}\, \otimes L$$

The definition of the valuation generated by OR-bids then lets us take the maximum of w and v.

Observe that the OR-language is only attractive if we do assume free disposal (i.e., weakening); without it, it has the same expressive power as the simple language of atomic bids. For example, without free disposal, $(p \multimap u^k) \otimes (q \multimap u^{k'})$ and $p \otimes q \multimap u^{k+k'}$ generate the same valuation.

It is interesting to remark that the usual characterization of the expressivity of OR-language for single-unit CAs cannot straightforwardly be extended to multi-unit case. Concerning the expressive power of our OR/$\otimes$-language, the following proposition holds.

Proposition 9. *OR-bids generate valuations that satisfy $v(X + Y) \geq v(X) + v(Y)$ whenever $X \cap Y = \emptyset$.*

Note that the assumption on the empty intersection amounts to assuming that there are no restrictions, besides monotonicity, on how functions behave on overlapping multisets. In particular, any function that is monotonic on multisets containing just one type of good should be expressible by means of an OR-bid. However, the converse of Proposition 9 is not true. Take for example the following function v.

$$v : \{p\} \mapsto 2$$

$$v : \{p, p\} \mapsto 5$$

$$v : \{p, p, p\} \mapsto 6$$

Suppose there is an OR-bid φ that generates v. As $v(\{p\}) = 2$, φ has to include an atomic bid $p \multimap 2$. As $v(\{p, p\}) = 5$, we have two choices. We can add an atomic bid to φ that specifies the marginal value of having a second copy of p, or we can add a bid $(p \otimes p \multimap 5)$. As the marginal value of having a second p in this case is strictly greater than $v(\{p\})$ (i.e. $v(\{p, p\}) - v(\{p\}) > v(\{p\})$), we cannot build v in the first way, because this would contradict the value of $v(\{p\})$. Thus, we have to add a bid $(p \otimes p \multimap 5)$ to φ. The value on $\{p, p, p\}$ cannot be 6, as the maximal value we can prove by using $\{p, p, p\}$, $p \multimap 2$, and $(p \otimes p \multimap 5)$ has to be at least $5 + 2$.

Thus, there is an important difference with the single-unit case. The problem is connected with the interpretation of the marginal value that can be associated to various copies of a same item. Moreover, since we are dealing with multisets of finite multiplicity, the valuations generated by our languages cannot grow arbitrarily, so at a certain point the function generated by the OR expression will provide a constant value. We leave a proper investigation of the expressivity of our tensor language to a future work.

An important restriction of our or/⊗-language provides the class of additive functions: $\textsc{add} = \bigotimes_{i\in\{1,\dots,m\}}[\underbrace{(p_i \multimap u) \otimes \cdots \otimes (p_i \multimap u)}_{\mathcal{M}(p_i) \text{ times}}]$

Proposition 10. ADD *generates all additive valuations and only those.*

For a proof, we refer to Porello and Endriss (2010a).

4.3 Goalbase Languages

The idea of a goalbase is that an agent is willing to pay the price associated to the formulas satisfied by the allocated propositional atoms. Thus, an agent views the allocation of a good a as *reusable* to satisfy his own goals. Namely, the reasoning within a goalbase is classical. Thus, one way to cope with goalbase languages is by means of formulas $a \multimap !a$: they intuitively express the fact that copies of the good a do not matter for the agent preferences. However, across different goalbases, quantities of goods matters: we cannot allocate the same instance of a good a to two different agents. We will discuss this point in the next section when we present our modelling of allocations.

Definition 11. *A goalbase is a formula of the form, with a_i in some of the B_j:*

$$a_1 \multimap !a_1 \otimes \cdots \otimes a_m \multimap !a_m \otimes (B_1 \multimap w_1) \otimes \cdots \otimes (B_l \multimap w_l)$$

Note that the difference with our OR-language is precisely in the interpretation of the goods. The semantics of our goalbase bids is given by Definition 2 as well.

Example 12. *Given an goalbase $p \multimap !p \otimes q \multimap !q \otimes (p \otimes q \multimap v) \otimes (q \multimap w)$, suppose the auctioneer provides $\{p, q\}$. The following proof shows that $\{p, q\}$ is enough to satisfy both atomic bids in the goalbase.*

$$\dfrac{\dfrac{q, q \multimap\, !q \vdash !q \qquad p, p \multimap\, !p \vdash !p}{p, q, p \multimap\, !p, q \multimap\, !q \vdash !p \otimes !q}\otimes R \qquad \dfrac{\dfrac{\dfrac{\dfrac{p, q, p \otimes q \multimap v \vdash v}{!p, q, p \otimes q \multimap v \vdash v}\,!L}{!p, !q, p \otimes q \multimap v \vdash v}\,!L \qquad \dfrac{q, q \multimap w \vdash w}{!q, q \multimap w \vdash w}\,!L}{!p, !q, !q, p \otimes q \multimap v, q \multimap w \vdash v \otimes w}\otimes R}{!p, !q, p \otimes q \multimap v, q \multimap w \vdash v \otimes w}\,!C}{!p \otimes !q, p \otimes q \multimap v, q \multimap w \vdash v \otimes w}\otimes L}{p, q, p \multimap\, !p, q \multimap\, !q, p \otimes q \multimap v, q \multimap w \vdash v \otimes w}\,cut$$

$$\dfrac{p, q, p \multimap\, !p, q \multimap\, !q, p \otimes q \multimap v, q \multimap w \vdash v \otimes w}{p, q, p \multimap\, !p \otimes q \multimap\, !q \otimes (p \otimes q \multimap v) \otimes (q \multimap w) \vdash v \otimes w}\otimes L \text{ and } E$$

The right hand side of the proof shows that reusable resources can satisfy both atomic bids. The left hand side shows that within the goalbase the goods are interpreted in a set-theoretic way by means of the formulas $a_i \multimap\, !a_i$. The last step in the proof shows that $\{p, q\}$ are enough to satisfy the goalbase bid.

Note that we can also mix different types of bids, e.g., bids that do and do not consume goods (OR- and goalbase bids). We will discuss in more detail the relationship between different types of resources later.

Regarding the expressivity of our goalbase bids, it is possible to adapt the relevant results of Uckelman et al. (2009) to the case of multiple units and to our LL framework.

Remark 13 (Classical and linear reasoning)**.** *Intuitionistic (and classical) logic can be translated into LL. cf. Girard (1995). Define the translation $(\cdot)^*$ as follows: $p^* = p$, $(A \wedge B)^* = A^* \,\&\, B^*$, $A \to B = !(A^*) \multimap B^*$, $(A \vee B)^* = A^* \oplus B^*$. We have that: $\Gamma \vdash_{IL} A$ if and only if $!\Gamma^* \vdash_{LL} A^*$. So we can translate any goalbase into a LL formula with the same logical behaviour, in the sense that they will be satisfied by the same sets of resources. However, the full power of exponentials makes LL with weakening, though decidable, cf. Kopylov (1995), exponential-space hard Urquhart (2000), while full LL is undecidable, cf. Lincoln et al. (1992).*

Note that if bidder demands are expressed by "classical" formulas, then the auctioneer has to provide any number of required goods. e.g., $!(p \,\&\, q) \multimap w$ can be satisfied only by providing any number of p and q, i.e., $!p$ and $!q$. Vice versa, if the auctioneer provides "sets" of goods, e.g. $!p, !q$, namely arbitrary quantities of each type of goods, the auctioneer would satisfy bidders' demands , e.g. $p \otimes q \multimap w$. While in principle one can model the interaction of bounded and unbounded resources (sets and multisets) in LL, the price to pay is complexity.

The use of exponentials for modelling goalbases sets us outside the Horn fragment (and MALL). Anyway, it is possible to define a bounded form of exponential as $!^{\ell}\varphi$, meaning that we can use φ at most ℓ times, cf. Girard et al. (1992). The upper bound for l in our model would be the maximal demand of a given type of good a. Thus, in order to provide a version of the goalbase language that lies within the Horn fragment, we can modify our definition as follows.

$$a_1 \multimap a_1^{h_1} \otimes \cdots \otimes a_m \multimap a_m^{h_m} \otimes (B_1 \multimap w_1) \otimes \cdots \otimes (B_l \multimap w_l)$$

Here $a_i^{h_i}$ denotes the tensor of a_i h_i-times, where h_i is the number of a_i that the agent's atomic bids actually demand.

5 Modelling Allocations

In this section, we formulate the problem of computing an allocation producing a certain amount of revenue as the problem of finding a proof for a LL sequent. This allows us, at least in principle, to model the winner determination problem as a series of calls to a LL theorem prover.

Let $\mathcal{M}$ again be a multiset of goods owned by the auctioneer, and let $\mathcal{N} = \{1, ..., n\}$ be the set of bidders. We add to the set of atoms $\mathcal{A} = \{p_1, \ldots, p_m\}$ all atoms p_i^j to express that the good p_i is allocated to the individual j. From now on, we will assume that bids are defined using these indexed names of goods, i.e., bidder $j \in \mathcal{N}$ must express her bid using the set of atoms $\{p_1^j, \ldots, p_m^j\}$.

In order to express that each (copy of) a good may be allocated to any of the bidders (but not to more than one), we shall use the following formula: [4]

$$\textsc{map} := \bigotimes_{p \in \mathcal{A}} [\&_{j \in \mathcal{N}}(p \multimap p^j)]^{\mathcal{M}(p)} \tag{1}$$

We now define the concept of *allocation sequent*, which is intended to capture the problem, faced by the auctioneer, of finding a feasible allocation returning

[4] Formula (1) is required in order for our approach to work with goalbase languages, since here we have to model that, on the one hand, goods are *reusable* within the bid of a single bidder and, on the other, goods are not *shareable* across the bids of distinct bidders.

a particular revenue. We restrict ourselves to the case of $\mathcal{W} = \{u\}$. We take $\mathcal{M}$ and $\mathcal{N}$ to be fixed, and MAP to be defined accordingly.

Definition 14. *The allocation sequent for revenue k and bids* $\text{BID}_1, \ldots, \text{BID}_n$ *is defined as the following LL sequent:*

$$\mathcal{M}, \text{MAP}, \text{BID}_1, \ldots, \text{BID}_n \vdash u^k$$

Example 15. *Supposes the auctioneer sells* $\{p, p, q, r\}$*. Suppose that three bidders express their preferences by means of an XOR bid, a goalbase bid and an or-bid.*

$$b_1 : (p_1 \otimes q_1 \multimap 4) \,\&\, (p_1 \otimes r_1 \multimap 3)$$

$$b_2 : p_1 \multimap !p_1 \otimes q_1 \multimap !q_1 \otimes (p \multimap 2) \otimes (p \otimes q \multimap 5) \otimes (q \multimap 2)$$

$$b_3 : (q \otimes r \multimap 5) \otimes (r \multimap 2)$$

The optimal allocation gives p and q to agent 2, thus satisfying the three atomic bids in b_2*, and the remaining p and r to bidder 1, thus satisfying* $(p_1 \otimes r_1 \multimap 3)$*. Agent 3 gets nothing. The proof in LL is presented in Table 2.*

We are now ready to state the relationship between proofs and actual allocations.

Theorem 16. *Given n bids in any of the bidding languages introduced (XOR, OR, Goalbase), every allocation* α *with revenue k provides a proof* π *of an allocation sequent for k, and vice versa, every proof* π *of an allocation sequent for k provides an allocation* α *with revenue k.*

For the proof, we refer to Porello and Endriss (2010b). The proof shows how it is possible to retrieve the goods allocated to the bidders from the sequent calculus proof as $A_j = \{a_j \mid a_j \vdash a_j \text{ is an axiom in } \pi\}$, those are the goods actually used to satisfy the bids in π.

The reason why we choose to model allocation in the intuitionistic fragment is that any intuitionistic sequent has just one conclusion, thus we can always pinpoint the revenue formula in the allocation sequent.

Our result shows that we could import known algorithms for winner determination for CAs into our framework. On the other hand, given a proof π in

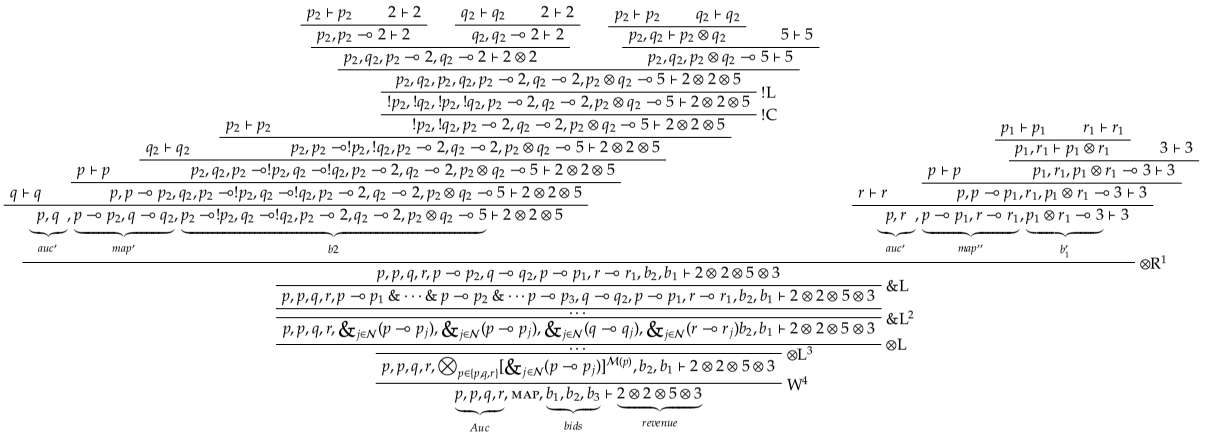


Table 2: Allocation sequent

1: The two branches of the proof satisfy b_2 and b_1 respectively. *auc*′ and *auc*″ are submultisets of the goods owned by the auctioneer, *map*′ and *map*′′ are subformulas of the MAP formula, b'_1 and b'_2 are subformulas of the bids.

2: The applications of &L are building the MAP formula.

3: The applications of ⊗L are building the MAP formula.

4: The application of weakening introduces the unused bid b_3.

the fragments we saw, we can transform it into a cut-free proof in polynomial time Girard et al. (1992). In a cut-free proof, each connective is visited exactly once, so given a proof of the allocation sequent, we can retrieve an allocation in polynomial time.

For the three languages presented, allocation sequents belong to HLL, so the complexity of checking whether revenue k is attainable is in NP Kanovich (1994), meaning that our form of modelling the problem does not increase complexity with respect to the standard approach Cramton et al. (2006). Of course, Theorem 16 only provides a method for solving the *decision variant* of the WDP. In practice, we will want to find the maximal revenue k such that u^k is provable. This can be achieved by using binary search over possible values of k and checking the corresponding allocation sequents in turn.

6 Modelling Multilateral Negotiation

Given an allocation α, by Theorem 16, we can associate a proof π that satisfies (some of) the agents' preferences. Let A be the tensor formula corresponding to the multiset of goods $\mathcal{M}$. Define $A_i \subseteq A$ to be the multiset of atoms allocated to agent i. We can state the definitions of social welfare within our framework as follows. Let VAL_i be a formula, in one of the language we have introduced, that expresses a valuation function.

Definition 17 (Utilitarian social welfare)**.**

$$sw_u(A) = \max\{k \mid A, \text{VAL}_1, \ldots, \text{VAL}_n \vdash u^k\}$$

We can consider a particular proof π of an allocation sequent and define the value of the allocation in that proof as $sw_u^\pi(A) = u^k$, where $A, \text{VAL}_1, \ldots, \text{VAL}_n \vdash u^k$ is in π. The value of the allocation for a certain agent i is given by: $u_i^\pi(A) = w_i$, where $A, \text{VAL}_i \vdash w_i$ is in π. So, for example, the utilitarian social welfare of a given allocation sequent is given by the sum of the individual utilities:

$$u_1^\pi(A) \otimes \cdots \otimes u_n^\pi(A) = sw_u^\pi(A)$$

Slightly abusing the notation, we identify the value $sw_u(A')$ with the value k of the tensor formula u^k. Given two allocations A and A', since we are using LL with (W), we have that $sw_u(A) \leq sw_u(A')$ iff $sw_u(A') \vdash sw_u(A)$. In order to define a strict order, we put $sw_u(A) < sw_u(A')$ iff $sw_u(A') \vdash sw_u(A) \otimes u$.

We can present now the definition of Pareto optimality.

Definition 18 (Pareto optimality)**.** *An allocation A is Pareto optimal iff there is no allocation A' such that $sw(A') \vdash sw(A) \otimes u$ and for all i, $u_i(A) \vdash u_i(A')$.*

6.1 Modelling deals

In this section, we define a general language to express deals, then in the next section we will see what it means for an agent to be willing to accept a deal. The language we define will be more general than one would expect, since we consider any kind of formula to be a deal.

We will not put structural constraints on the formula expressing deals; rather, the condition we will put on the feasibility of the negotiation will provide the expected meaning of deals, namely that they transform an allocation A into and allocation A'.

Definition 19. *A deal is any formula of linear logic built over the indexed alphabet $\mathcal{A}^j$.*

So for example a single atom p^j means that p goes to agent j. The meaning of a deal of the form $p^1 \multimap q^3$ is simply the agent 1 loses p and the agent 3 gets q.

Definition 20. *We say that an allocation A' is obtained from A by a* DEAL *iff*

$$A, \text{DEAL} \vdash A'$$

The fact that we use provability to model the passage from a A to A' amounts to assuming that the deals are feasible in the sense that they concern the resources in A. For example, take $p^1 \multimap p^2$; if agent 1 does not own p in A, then such a deal will not be used.

Remark 21. *There are some situations we are excluding. The valuations we are considering are defined on multisets that are represented in our language by tensor formulas. We will not consider here valuations defined on other types of formulas, as options like $a \& b$ (agent has the choice) or $a \oplus b$ (agent doesn't have the choice): it would require a rather different definition of valuation functions. We leave such extensions to future work.*

We discuss some examples. Deals that simply move a single resource p from one agent to another (1-deals) can be modelled as implications of the form

$p^i \multimap p^j$. A *swap deal* Sandholm (1998) between individuals is defined by the following formula $(p^i \multimap p^j) \otimes (q^j \multimap q^i)$, which means that i gives p to j and j gives q to i. For example, let $A = \{p^1, q^2, r^3\}$, we can get $A' = \{p^2, q^1, r^3\}$ by the swap:

$$p^1, q^2, r^3, (p^1 \multimap p^2) \otimes (q^2 \multimap q^1) \vdash p^2 \otimes q^1 \otimes r^3$$

Note that, according to this definition, there might be deals that change nothing, e.g., $p^i \multimap p^i$. Moreover, we can also consider deals that simply provide a resource p to a certain agent i, p^i. In this way, we can for example model, as a form of negotiation, the passage form a partial allocation, in which some goods were not allocated, to a total one:

$$\underbrace{p^1, q^2}_{A}, \underbrace{p^1 \multimap p^2, q^3}_{\text{DEAL}} \vdash \underbrace{p^2 \otimes q^2 \otimes q^3}_{A'}$$

Cluster deals Sandholm (1998), where agents exchange more then one item, can be modelled using tensors: $p^i \otimes q^i \otimes r^i \multimap p^j \otimes q^j \otimes r^j$, meaning that i gives one p, one q and one r to j.

The language of LL allows for expressing deals that entail some forms of choice. Let us call them *optative deals*. So, for example, $(p^1 \multimap p^3) \,\&\, (p^2 \multimap p^3)$ means that 3 would get p from 1 or from 3 (but not from both), or $(p^1 \multimap p^2) \,\&\, (p^1 \multimap p^2)$ means that 1 would give p to 2 or to 3 (but not to both).

Using the distributivity law of LL, $A \parr (B \,\&\, C) \dashv\vdash (A \parr B) \,\&\, (A \parr C)$, we can write optative deals in the following forms. We can express deals like "someone gives p to i" as follows: $(p^1 \oplus \cdots \oplus p^n) \multimap p^i$. Symmetrically, we can express "i gives p to someone": $p^i \multimap (p^1 \,\&\, \cdots \,\&\, p^n)$. In an analogous way, we can consider "i gives something to j" and "i gets something from j".

Taking the language of deals in its full generality, we can also define *transformations* of deals, for example $(p^i \multimap p^j) \multimap (r^j \multimap r^i)$, the intuitive meaning of which is that j would give r to i if the deal $(p^i \multimap p^j)$ has been accepted in the negotiation.

Example 22. *Let A be $\{p^1, r^3, p^1, q^2\}$ and the deals $p^1 \multimap p^2$ and $q^2 \multimap q^3$, meaning that 1 gives on p to 2 and 2 gives one q to 3.*

The following proof shows that $A' = \{p^1, r^3, p^2, q^3\}$ is obtained from A:

$$\dfrac{p^1, r^3 \vdash p^1 \otimes r^3 \qquad \dfrac{p^1, p^1 \multimap p^2 \vdash p^2 \qquad q^2, q^2 \multimap q^3 \vdash q^3}{p^1, q^2, p^1 \multimap p^2, q^2 \multimap q^3 \vdash p^2 \otimes q^3} \otimes L}{p^1, r^3, p^1, q^2, p^1 \multimap p^2, q^2 \multimap q^3 \vdash p^1 \otimes r^3 \otimes p^2 \otimes q^3} \otimes L$$

We can prove that the language of deals is sufficiently powerful to express every transformation of allocations A, A'.

Proposition 23. *Let A and A' be two allocations. Then there exists a formula* DEAL *in the deal language such that*

$$A, \textsc{deal} \vdash A'$$

The proof is obvious in the sense that it is enough to consider the formula $A \multimap A'$ as a deal. We can define a general notion of negotiation as follows.

Definition 24. *A negotiation is a sequent $A, \textsc{deal}_1, \ldots, \textsc{deal}_l \vdash A'$ where* $\textsc{deal}_1$, ..., $\textsc{deal}_l$ *are accepted deals according to some criterion.*

We can also consider the feasibility of an allocation with respect a given multiset of resources as follows:

$$\mathcal{M}, \textsc{map} \vdash A \tag{2}$$

Here, MAP is the formula defined as in (1). The provability of (2) entails that, given the actual multiset of resources $\mathcal{M}$, A is a feasible way to assign goods.

6.2 Rationality of Deals

In this section, we present some conditions that specify when an agent would accept a deal. Basically, according to the relevant literature, cf. Endriss et al. (2006), we distinguish two cases, one with *side payments* and one without. A *payment function* is a function $p : \mathcal{N} \to \mathbb{Z}$ such that

$$\sum_{i \in \mathcal{N}} p(i) = 0$$

Using side payments, the notion of individual rationality can be defined as follows. A deal is individually rational iff whenever A' is obtained by A by

means of that deal, then there exist a payment function p such that for all $i \in \mathcal{N}$:

$$v_i(A') > p(i) + v_i(A)$$

We rephrase the notion of payment function considering formulas in our language as side payments. The requirement that the prices actually paid must sum up to zero is here interpreted as the provability of the sequent containing positive and negative payments. Intuitively, there should be a matching between who pays and who gets payments.

Definition 25. *A side payment is a sequent $X \dashv\vdash Y$, where X and Y are tensors of u, that is provable in LL. We call the formulas on the left negative payments and those on the right positive payments.*

We could also consider more general formulas as side payments. As an example of possible generalisation, we can consider an individual i who would accept to face a loss of three units of her utility for getting one q; it can be modelled using the formula $u^3 \multimap q^i$.

Using payment sequents we can rephrase the notion of individual rationality as follows.

Definition 26. *Given a deal* DEAL *such that A' is obtained by A by means of* DEAL *and a side payment $X \dashv\vdash Y$, we say that* DEAL *is individually rational iff for all i, $u_i(A'_i), X_i \vdash u_i(A_i) \otimes Y_i$ and there exists a j such that: $u_j(A'), X_j \vdash u_j(A_j) \otimes u \otimes Y_j$, where $X_1 \otimes \cdots \otimes X_n \cong X$ and $Y_1 \otimes \cdots \otimes Y_n \cong Y$.*

Note that, since we are working with integers, we do not require all agents to experience a (possibly infinitesimally small) improvement, but rather ask that no agent suffers a loss, and at least one of them gains one full unit u. We can derive the case without side payments, by taking the payment sequent to be $\mathbf{1} \vdash \mathbf{1}$, yielding the following definition of *cooperative rationality*,cf. Endriss et al. (2006):

Definition 27. *A deal formula* DEAL *such that A, DEAL $\vdash A'$ is cooperatively rational iff for all i, $u_i(A') \vdash u_i(A)$ and there exists a j such that $u_j(A') \vdash u_j(A) \otimes u$.*

In what follows, w.l.o.g., we will consider payments in which, for each i, (at least one of) X_i or Y_i is the tensor unit $\mathbf{1}$.

Example 28. *Suppose we want to determine whether a deal taking us from allocation A to A' is individually rational. Let $u_1(A') = u^{15}$, $u_2(A') = u^{10}$, $u_3(A') = u^5$ and $u_1(A) = u^2$, $u_2(A) = u^1$, $u_3(A) = u^6$. We can define X_i and Y_i as follows:*

$$u^{15} \vdash u^2 \otimes u^6 \qquad Y_1 = u^6$$
$$u^{10} \vdash u^1 \otimes u^2 \qquad Y_2 = u^2$$
$$u^5, u^8 \vdash u^6 \qquad X_3 = u^8$$

We have that positive and negative payments match: $u^8 \dashv\vdash u^6 \otimes u^2$.

We can now state the relationship between individual rationality and social welfare by means of the following theorems. The next result corresponds to (Endriss et al. 2006, Lemma 1), except that we get a more precise characterisation in the context of integer valuations: a deal is individually rational if and only if it increases social welfare by at least one unit.

Theorem 29 (Rational deals and social welfare)**.** *A deal formula* DEAL *with* $A, \text{DEAL} \vdash A'$ *is individually rational iff* $sw_u(A') \vdash sw_u(A) \otimes u$.

For a proof we refer to Porello and Endriss (2010a). In a similar way we can prove a result linking cooperative rationality and Pareto improvements Endriss et al. (2006).

The following result shows that allocations with maximal utilitarian social welfare can be reached from any (suboptimal) allocation A by means of individually rational deals.

Theorem 30. *Let* A^* *be an allocation with maximal social welfare. Then for any allocation* A *with lower social welfare there exists an individually rational deal* DEAL *such that* $A, \text{DEAL} \vdash A^*$.

The proof relies on the fact that there always exists a deal to reach A^* from A, by Proposition 23. Since social welfare improves, by Theorem 29 such a deal is individually rational.

It is interesting to remark that, since we are dealing with integer valuations, if we consider any set of rational deals, each of them must make social welfare increase by at least one unit. Thus, if k is the difference between the maximal social welfare and the social welfare of the initial allocation, then we will always reach an optimal allocation by means of any sequence of at most k individually rational deals.

7 Conclusion

We have introduced a model of MARA based on linear logic and we have applied it to model combinatorial auctions and negotiation over multisets of goods. In particular, we have stressed how linear logic allows for understanding MARA problems as reasoning tasks: we have discussed the relationship between types of goods and suitable reasoning rules. In particular, agents can reason by means of classical logic about *sets* of objects, while we have to drop contraction in order to reason about multisets of goods. Moreover, we have interpreted the free disposal assumption as a matter of accepted reasoning rules, namely weakening.

Linear logic allows also for describing different types of goods. The idea that LL may be useful in designing bidding languages that can distinguish shareable from nonsharable goods has already been hinted at by Boutilier and Hoos (2001). We can define the full availability of a good for the bidders as $!^{\ell}p$ where ℓ is big enough, so $!^{\ell}p$ can be shared by all bidders demanding it. In order to express the *reusability* of a good for a single bidder j, we can write $!^{\ell}p^{j}$, which will satisfy only bidder j's bids. In order to make explicit that j can reuse p as much as she likes, we can add the formula $p^{j} \multimap !^{\ell}p^{j}$ to j's bid formula as we did for modelling the goalbase languages. We presented here a treatment of MARA on sets and multisets of goods. We can easily extend our approach to *lists* of goods, namely bundles of goods such that the order in which they are sold matters. In order to do it, we need to drop (E). The logic we obtain is then the well known Lambek calculus, cf. Pentus (2003). In principle, all the languages we have developed can be adapted to the non-commutative case. Moreover, the interaction of multisets and lists of goods can be modelled by using the partially commutative linear logic, cf. de Groote (1996). Such extensions are left for future work.

Acknowledgements I would like to thank the participants of the LIRa seminar for the insightful comments and discussions.

References

C. Boutilier and H. H. Hoos. Bidding languages for combinatorial auctions. In *Proc. 17th Int'l Joint Conf. on Artif. Intell. (IJCAI-2001)*, 2001.

J. Cerquides, U. Endriss, A. Giovannucci, and J. A. Rodríguez-Aguilar. Bidding languages and winner determination for mixed multi-unit combinatorial auctions. In *Proc. 20th Int'l Joint Conf. on Artif. Intell. (IJCAI-2007)*, 2007.

Y. Chevaleyre, P. E. Dunne, U. Endriss, J. Lang, M. L. tre, N. Maudet, J. Padget, S. Phelps, J. A. R. guez Aguilar, and P. Sousa. Issues in multiagent resource allocation. *Informatica*, 30:3–31, 2006.

Y. Chevaleyre, U. Endriss, S. Estivie, and N. Maudet. Multiagent resource allocation in *k*-additive domains: Preference representation and complexity. *Annals of Operations Research*, 163(1):49–62, 2008.

Y. Chevaleyre, U. Endriss, and N. Maudet. Simple negotiation schemes for agents with simple preferences: Sufficiency, necessity and maximality. *Journal of Autonomous Agents and Multiagent Systems*, 20(2):234–259, 2010.

P. Cramton, Y. Shoham, and R. Steinberg, editors. *Combinatorial Auctions*. MIT Press, 2006.

P. de Groote. Partially commutative linear logic: Sequent calculus and phase semantics, 1996.

U. Endriss and E. Pacuit. Modal logics of negotiation and preference. In *Proc. 10th European Conference on Logics in Artificial Intelligence (JELIA-2006)*. Springer-Verlag, 2006.

U. Endriss, N. Maudet, F. Sadri, and F. Toni. Negotiating socially optimal allocations of resources. *Journal of Artificial Intelligence Research*, 25:315–348, 2006.

M. Fisher. Characterizing simple negotiation as distributed agent-based theorem-proving: A preliminary report. In *Proc. 4th International Conference on Multi-Agent Systems (ICMAS-2000)*, 2000.

J.-Y. Girard. Linear logic. *Theor. Comput. Sci.*, 50(1):1–101, 1987.

J.-Y. Girard. Linear logic: Its syntax and semantics. In *Advances in Linear Logic*. Cambridge University Press, 1995.

J.-Y. Girard, A. Scedrov, and P. J. Scott. Bounded linear logic: a modular approach to polynomial-time computability. *Theor. Comput. Sci.*, 97(1):1–66, 1992. ISSN 0304-3975. doi: http://dx.doi.org/10.1016/0304-3975(92)90386-T.

J. Harland and M. Winikoff. Agent negotiation as proof search in linear logic. In *Proc. 1st International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-2002)*, 2002.

M. I. Kanovich. The complexity of Horn fragments of linear logic. *Ann. Pure Appl. Logic*, 69(2-3):195–241, 1994.

A. P. Kopylov. Decidability of linear affine logic. In *Proc. 10th Annual IEEE Symposium on Logic in Computer Science (LICS-1995)*. IEEE Computer Society, 1995. ISBN 0-8186-7050-6.

P. Küngas and M. Matskin. Symbolic negotiation with linear logic. In *Proc. 4th International Workshop on Computational Logic in Multiagent Systems (CLIMA IV)*. Springer-Verlag, 2004.

J. Leite, J. J. Alferes, and B. Mito. Resource allocation with answer-set programming. In *Proc. 8th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-2009)*, 2009.

P. Lincoln. Deciding provability of linear logic formulas. In *Proc. Workshop on Advances in Linear Logic*. Cambridge University Press, 1995. ISBN 0-521-55961-8.

P. Lincoln, J. C. Mitchell, A. Scedrov, and N. Shankar. Decision problems for propositional linear logic. *Ann. Pure Appl. Logic*, 56(1–3):239–311, 1992.

N. Nisan. Bidding languages for combinatorial auctions. In *Combinatorial Auctions*. MIT Press, 2006.

M. Pentus. Lambek calculus is np-complete. *Theoretical Computer Science*, 357: 2006, 2003.

D. Porello and U. Endriss. Modelling multilateral negotiation in linear logic. In *Proceedings of the 19th European Conference on Artificial Intelligence (ECAI-2010)*, August 2010a.

D. Porello and U. Endriss. Modelling combinatorial auctions in linear logic. In *Proc. 12th International Conference on the Principles of Knowledge Representation and Reasoning (KR-2010)*, 2010b.

F. Sadri, F. Toni, and P. Torroni. An abductive logic programming architecture for negotiating agents. In *Proc. 8th European Conference on Logics in Artificial Intelligence (JELIA-2002)*. Springer-Verlag, 2002.

T. W. Sandholm. Contract types for satisficing task allocation: I Theoretical results. In *Proc. AAAI Spring Symposium: Satisficing Models*, 1998.

A. S. Troelstra. *Lectures on Linear Logic*. CSLI Publications, 1992.

J. Uckelman, Y. Chevaleyre, U. Endriss, and J. Lang. Representing utility functions via weighted goals. *Mathematical Logic Quarterly*, 55(4):341–361, 2009.

A. Urquhart. The complexity of linear logic with weakening. In Buss, Hájek, and Pudlák, editors, *Logic Colloquium '98*, number 13 in Lecture Notes in Logic, pages 57–90. AK Peters, 2000.

Consequentialist Deontic Logic for Decisions and Games

Xin Sun and Fenrong Liu

Department of philosophy , Tsinghua University
shinksun@gmail.com, fenronguva@gmail.com

Abstract

The paper presents a new consequentialist deontic logic in which the relation of preference over sets of possible worlds and the relation of conditional dominance are both transitive. This logic validates the principle that absolute ought can be derived from conditional ought in case the conditional statement is the agent's absolute ought. Ought about conditionals is not implied by conditional ought in this logic. Next, we show how this logic at work in the following two settings: first, we we look at a recent paradox that has been presented in Kolodny and MacFarlane (2010) and provide our solution. Second, we revisit Anderson's six deontic principles of conditional ought and show their validity or invalidity in our framework. In addition, we propose a more proper interpretation for the notion of permission after reviewing the latest work in multi-agent deontic action logic.

1 Introduction

Let us introduce an example from game theory right away:

Example 1 (matching pennies)**.** Ann is playing a matching pennies game with Bob. They choose, simultaneously, whether to show the head or the tail of a coin. If they show the same side, Ann receives one dollar from Bob. If they show different sides, Bob receives one dollar from Ann. The situation can be depicted in Figure 1:

		Bob: head	Bob: tail
Ann	head	w_1 1,-1	w_2 -1,1
	tail	w_3 -1,1 $^{\gamma}$	w_4 1,-1

Figure 1

Obviously, under the condition of Bob showing the head, Ann ought to see to it that she shows the head. The question we would ask here is the following, does Ann ought to see to it that if Bob shows the head then she shows the head? The answer seems to be positive at first sight. However, we claim here that the answer is negative. Here is how we reason: Denote the situation in which Ann shows the tail and Bob shows the head as γ. First note that to Ann, both showing the head and showing the tail are optimal, which implies that Ann is permitted to show the tail. Since Ann showing the tail may lead to the situation γ, Ann is permitted to act towards the situation γ. Next, assume the answer is positive, that is, Ann ought to see to it that if Bob shows the head, then Ann shows the head. This means Ann ought to prohibit the following outcome: Bob shows the head but Ann shows the tail. Hence Ann ought to prohibit the situation γ. Contradiction.

A decision theoretical example of the same sort can be found at the end of Chapter 5 of Horty (2001) and we re-state it here:

Example 2 (swerve or continue?)**.** Two drivers are traveling toward each other on a one-lane road, with no time to stop or communicate, and with a single moment at which each must choose, independently, either to swerve or to

continue along the road. There is only one direction in which the drivers might swerve, and so a collision can be avoided only if one of the drivers swerves and the other does not; if neither swerves, or both do, a collision occurs. We can depict this situation in Figure 2:

		Driver B: continue	Driver B: swerve
Driver A	swerve	w_1 1	w_2 0
	continue	w_3 0	w_4 1

Figure 2

Given the above situation, under the condition that Driver B swerves, Driver A ought to see to it that he continues. Similarly, we can ask the same question: does Driver A ought to see to it that if Driver B swerves then he continues? Again, our answer to this question is No. The reason is the following: If Driver A ought to see to it that if Driver B swerves, then he continues. Then he ought to prohibit the situation in which both Driver B and he himself swerve. In order to prohibit that situation, Driver A has to see to it that he continues. This means, Driver A ought not to choose to swerve. But this contradicts to our intuition.

In what follows, we will refer sentences of the form "under the condition of φ, agents ought to see to it that ψ" as *conditional ought*, and we use *ought about conditionals* to name sentences of the form "agents ought to see to it that if φ then ψ". Anderson (1959) suggested that a desired theory of conditional ought should include the principle that conditional ought implies ought about conditionals. In Aqvist (1994), this principle is presented as a valid axiom in the strongly normal system **G**. However, the above two examples have both suggested that this principle is doubtable. Similar scenarios abound in real life. In this paper, we are going to investigate around this issue and propose a consequentianlist deontic logic, in which ought about conditionals is not implied by conditional ought.

The structure of the paper is as follows: Section 2 is an introduction to our new consequentianlist deontic logic, including the language and semantics. We then show the system at work in the following two settings: moral dilemma

and deontic logic tradition. In Section 3 we study a recent widely-discussed paradox that has appeared in Kolodny and MacFarlane (2010) and provide our solution by using ideas from our new framework. In Section 4 we analyze the six deontic principles concerning conditional ought first proposed by Anderson and prove their validity and invalidity in our logic. Finally, in Section 5 we look at the related works, in particular, we propose a new notion of permission for the latest work on multi-agent deontic action logic presented in Tamminga (2011 Nov. 24). We end up with some conclusions and future work in Section 6.

2 A New Consequentianlist Deontic Logic

2.1 Language

Definition 2.1 (language)**.** The language of the consequentianlist deontic logic is built from a finite set A of agents and a countable set P of atomic propositions. We will use p and q as variables for atomic propositions in P, and use F and G, where $F, G \subseteq A$, as groups of agents. The consequentianlist deontic language $\mathfrak{L}$ is given by the following Backus-Naur Form:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Diamond\varphi \mid [G]\varphi \mid \bigodot{}^{F}_{G}\varphi \mid \bigodot{}^{F}_{G}(\varphi/\varphi)$$

Intuitively, $\Diamond\varphi$ can be read as " It is possible that φ". $[G]\varphi$ can be read as "Group G sees to it that φ". $\bigodot{}^{F}_{G}\varphi$ can be read as "In the interest of group F, group G ought to see to it that φ". $\bigodot{}^{F}_{G}(\varphi/\psi)$ can be read as "In the interest of group F, group G ought to see to it that φ under the condition of ψ". We use $\boldsymbol{P}^{F}_{G}\varphi$ as an abbreviation of $\neg \bigodot{}^{F}_{G} \neg\varphi$, which can be read as 'In the interest of group F, group G is permitted to lead to a situation in which φ is true."

2.2 Consequentialist Frame

The semantics of consequentianlist deontic logic is based on the consequentianlist frames. Similar to Kooi and Tamminga (2008a), our definition of consequentianlist frame is as follows:

Definition 2.2 (consequentianlist frame)**.** A *consequentianlist frame* $\boldsymbol{F}$ is a quadruple $\langle W, A, Choice, \{Value_F\}_{F\subseteq A}\rangle$, where W is a nonempty set of possible worlds, A is a finite set of agents, *Choice* is a choice function, and $Value_F$, represents the preference of some group of agents $F \subseteq A$ and a function from W to the set of real numbers $\boldsymbol{R}$. Formally, $Value_F : W \to \boldsymbol{R}$.

The choice function *Choice* is a function from the power set of A to the power set of the power set of W, *i.e.* $Choice : \wp(A) \mapsto \wp(\wp(W))$. *Choice* is built from the individual Choice function *IndChoice*: $A \mapsto \wp(\wp(W))$. *IndChoice* must satisfy the following three conditions:

(1) for each agent $i \in A$ it holds that $IndChoice(i)$ is a partition of W;

(2) for each selection function s that assigning to each agent $i \in A$ a set of possible worlds $s(i) \in IndChoice(i)$, it holds that $\bigcap_{i\in A} s(i)$ in nonempty;

(3) for each $i \in A$, the set $IndChoice(i)$ is finite.

Let *Select* be the set of all selection functions, then

$$Choice(G) = \{\bigcap_{i\in G} s(i) : s \in Select\}$$

if G is nonempty. Otherwise, $Choice(G) = \{W\}$. For any two world w and w', if there exists a $K \in Choice(G)$ such that $w \in K$ and $w' \in K$, we denote it as $w \sim_G w'$. Intuitively, $w \sim_G w'$ means the choice of group G cannot distinguish w from w'.

In order to better understand this definition, take the Prisoner's Dilemma from Osborne and Rubinstein (1994) as an example.

Example 3 (Prisoner's Dilemma)**.** The Prisoner's Dilemma can be represented by the following figure:

		player β: quiet	player β: fink
player α	quiet	w_1 3, 3	w_2 0, 4
	fink	w_3 4, 0	w_4 1, 1

Figure 3

In this example, A={α, β},$W = \{w_1, w_2, w_3, w_4\}$, *IndChoice* ($\alpha$)={{$w_1$, w_2},{w_3,w_4}}, *IndChoice*(β) ={{w_1, w_3},{w_2,w_4}}. Apparently, both *IndChoice*(α) and *IndChoice*(β) are partitions of W. And there are four selection functions, *Select* ={s_1, s_2, s_3, s_4}, where:

$$s_1(\alpha) = \{w_1, w_2\}, s_1(\beta) = \{w_1, w_3\}$$

$$s_2(\alpha) = \{w_1, w_2\}, s_2(\beta) = \{w_2, w_4\}$$

$$s_3(\alpha) = \{w_3, w_4\}, s_3(\beta) = \{w_1, w_3\}$$

$$s_4(\alpha) = \{w_3, w_4\}, s_4(\beta) = \{w_2, w_4\}$$

So we have for each $s \in$ *Select*, $\bigcap_{i\in A} s(i)$ is not empty. Therefore the two conditions of individual choice are both satisfied. Then we have *Choice*(A) = $\{\bigcap_{i\in A} s(i) : s \in \textit{Select}\}$= {{$w_1$}, {$w_2$}, {$w_3$}, {$w_4$}}.

Having defined consequentianlist frames, we are ready to define *preferences over sets of worlds*. [1]

Definition 2.3 (preferences over sets of worlds; $\leq_F$, $<_F$)**.** Let $X \subseteq W$, $Y \subseteq W$ be two sets of worlds, F a group of agents from a consequentialist frame. Then $X \leq_F Y$(Y is *weakly preferred* to X) if and only if
(1) for each $w \in X$, for each $w' \in Y$, $Value_F(w) \leq Value_F(w')$ and
(2) there exists some $v \in X$, $v' \in Y$, $Value_F(v) \leq Value_F(v')$. $X <_F Y$ (Y is *stongly preferred* to X) if and only if $X \leq_F Y$ and it is not the case that $Y \leq_F X$.

Given Definition 2.3, we can obtain some useful lemmas and propositions, listed below. Note that in what follows we will omit some obvious proofs and present the important ones only with details.

Lemma 1. *Let X and Y be two sets of worlds, F a group of agents from a consequentialist frame. Then $X \leq_F Y$ if and only if $Value(w) \leq Value(w')$ for each $w \in X$, for each $w' \in Y$ and $X \neq \emptyset$, $Y \neq \emptyset$.*

[1]The definition in Horty (2001) has some flaw, hence cannot ensure the transitivity of the preference relation, see more details in Sun (2011).

Lemma 2. *Let X and Y be two sets of worlds, F a group of agents from a consequentialist frame. Then $X <_F Y$ if and only if*
(1) $Value_F(w) \leq Value_F(w')$ for each $w \in X$, for each $w' \in Y$,
(2) $Value_F(w) < Value_F(w')$ for some $w \in X$, for some $w' \in Y$.

Proof. Left to right. Assume $X <_F Y$, then we have $X \leq_F Y$ and it is not the case that $Y \leq_F X$. $X \leq_F Y$ plus Lemma 1 implies (1) and $X \neq \emptyset$, $Y \neq \emptyset$. By it is not the case that $Y \leq_F X$, we have either $X = \emptyset$ or $Y = \emptyset$ or for some $w' \in Y$, for some $w \in X$, $Value_F(w') > Value_F(w)$. But we already have $X \neq \emptyset$ and $Y \neq \emptyset$. Hence for some $w' \in Y$, for some $w \in X$ $Value_F(w') > Value_F(w)$, then (2) is true.

Right to left. By (2) we know $X \neq \emptyset$ and $Y \neq \emptyset$. This plus (1) implies $X \leq_F Y$. So it's sufficient to prove the following: it is not the case that $Y \leq_F X$. Suppose $Y \leq_F X$, then for each $w' \in Y$ and for each $w \in X$, $Value_F(w') \leq Value_F(w)$. But according to (2), for some $w' \in Y$ and for some $w \in X$, $Value_F(w') > Value_F(w)$. Contradiction. □

Proposition 1. *Let X and Y be sets of worlds, F a group of agents from a consequentialist frame. Then:*

1. *If $X \leq_F Y$ and $Y \leq_F Z$, then $X \leq_F Z$.*
2. *If $X \leq_F Y$ and $Y <_F Z$, then $X <_F Z$.*
3. *If $X <_F Y$ and $Y \leq_F Z$, then $X <_F Z$.*
4. *If $X <_F Y$ and $Y <_F Z$, then $X <_F Z$.*

Proof. Here we just prove Clause 1. Assume $X \leq_F Y$ and $Y \leq_F Z$, then $X \neq \emptyset$, $Y \neq \emptyset$ and $Z \neq \emptyset$. Let w be an arbitrary history in X, w'' be an arbitrary history in Z. By $Y \neq \emptyset$ we have that there exists some $w' \in Y$. By $X \leq_F Y$ and $Y \leq_F Z$ we have $Value_F(w) \leq Value_F(w')$, $Value_F(w') \leq Value_F(w'')$, hence $Value_F(w) \leq Value_F(w'')$. Therefore $X \leq_F Z$. □

Proposition 1 states that the relation of preference over sets of worlds is transitive.[2] This property is what has been assumed in Horty (2001) and our semantics. And it is crucial in that only with this transitive relation, we can properly define the concept of dominance and optimal.

[2]Though we are aware of the long-standing discussions on whether transitivity should be taken as the property of preference. We take it in this context to conform to Horty's spirit.

Definition 2.4 (dominance relation; $\leq_G^F$)**.** Let F be a consequentialist frame. Let $F, G \subseteq A$ and $K, K' \in Choice(G)$. Then

$$K \leq_G^F K' \quad \text{iff} \quad \text{for all } S \in Choice(A - G), K \cap S \leq_F K' \cap S.$$

$K \leq_G^F K'$ can be read as "in the interest of group G, K' weakly dominates K". From a game theoretical perspective, $K \leq_G^F K'$ means that no matter how other agents act, the agent's payoff of choosing K' is no less than that of choosing K. We use $K <_G^F K'$ as an abbreviation of $K \leq_G^F K'$ but $K' \leq_G^F K$ does not hold. If $K <_G^F K'$, we then say K' strongly dominate K.

Lemma 3. *Let F, G be groups of agents from a consequentialist frame, and let K, $K' \in Choice(G)$. Then $K <_G^F K'$ if and only if*
(1) $K \cap S \leq_F K' \cap S$ for each $S \in Choice(A - G)$, and
(2) $K \cap S <_F K' \cap S$ for some $S \in Choice(A - G)$.

Proof. Left to right. Assuming $K <_G^F K'$, then $K \leq_G^F K'$ and it is not the case that $K' \leq_G^F K$. (1) directly follows from $K \leq_G^F K'$. By it is not the case that $K' \leq_G^F K$, we have for some state $S \in Choice(A - G)$, it is not the case that $K' \cap S \leq_F K \cap S$. But according to (1) we already have $K \cap S \leq_F K' \cap S$. Hence $K \cap S <_F K' \cap S$. Which means (2) is true.

Right to left. By (1) we know $K \leq_G^F K'$. So we only need to prove it is not that $K' \leq_G^F K$. That is, for some state $S \in Choice(A-G)$, it is not that $K' \cap S \leq_F K \cap S$. This is implied by (2), because (2) means for some $S \in Choice(A - G)$, $K \cap S \leq_F K' \cap S$ and it is not the case that $K' \cap S \leq_F K \cap S$. □

Proposition 2. *Let F, G be groups of agents from a consequentialist frame, and let K, K', $K'' \in Choice(G)$. Then:*

1. *If $K \leq_G^F K'$ and $K' \leq_G^F K''$, then $K \leq_G^F K''$.*
2. *If $K \leq_G^F K'$ and $K' <_G^F K''$, then $K <_G^F K''$.*
3. *If $K <_G^F K'$ and $K' \leq_G^F K''$, then $K <_G^F K''$.*
4. *If $K <_G^F K'$ and $K' <_G^F K''$, then $K <_G^F K''$.*

Proof. Similar to the proof of Proposition 4.7 in (Horty 2001). □

Proposition 2 states that the dominance relation is transitive. This transitivity is actually still true even if we delete the existential clause in Definition 2.3. Because the definition of choice function ensures that for any $K \in \mathit{Choice}(G)$, for any $S \in \mathit{Choice}(A - G)$, $K \cap S \neq \emptyset$. However, the transitivity of conditional dominance (Definition 2.6 below) does rely on the existential clause of Definition 2.3. The conditional dominance is defined on restricted choice sets, which we will give below.

Definition 2.5 (restricted choice sets)**.** Let F, G be groups of agents from a consequentialist frame, X a set of worlds in the frame. Then

$$\mathit{Choice}(G/X) = \{K : K \in \mathit{Choice}(G) \text{ and } K \cap X \neq \emptyset\}$$

Intuitively, $\mathit{Choice}(G/X)$ is the collection of group G's choice which is consistent with condition X. We can further define conditional dominance relation over agent's choice. The intuition is this: to compare whether the agent's choice K is dominated by K' under the condition X, we only need to consider other agents' choices which are consistent with the condition X and one of K and K'.

Definition 2.6 (conditional dominance; $\leq^F_{G/X}$)**.** Let F, G be groups of agents from a consequentialist frame, X a set of worlds in the frame. Let $K, K' \in \mathit{Choice}(G/X)$. Then

$$K \leq^F_{G/X} K' \quad \text{iff} \quad \text{for all } S \in \mathit{Choice}((A-G)/(X \cap (K \cup K'))), K \cap X \cap S \leq_F K' \cap X \cap S.$$

$K \leq^F_{G/X} K'$ can be read as "in the interest of group F, K' weakly dominates K under the condition of X". The intuition behind this definition is this: when we want to know whether our action K dominates action K' under some condition X, other agents' choices which are inconsistent with the condition X or inconsistent with neither K nor K' are treated as irrelevant.

And we will use $K <^F_{G/X} K'$, which reads as "in the interest of group F, K' strongly dominates K under the condition of X", to express $K \leq^F_{G/X} K'$ and it is not the case that $K' \leq^F_{G/X} K$.

Lemma 4. *Let F, G be groups of agents from a consequentialist frame, X a set of worlds in the frame. Let $K, K' \in \mathit{Choice}(G/X)$. Then $K <^F_{G/X} K'$ if and only if*

(1) $K \cap X \cap S \leq_F K' \cap X \cap S$ *for each* $S \in Choice((A-G)/(X \cap (K \cup K')))$*, and*

(2) $K \cap X \cap S <_F K' \cap X \cap S$ *for some* $S \in Choice((A-G)/(X \cap (K \cup K')))$*.*

Proof. Similar to the proof of Lemma 3. □

Proposition 3. *Let F, G be groups of agents from a consequentialist frame, X a set of worlds in the frame. Let* $K, K', K'' \in Choice(G/X)$*. Then the following holds,*

1. *If* $K \leq^F_{G/X} K'$ *and* $K' \leq^F_{G/X} K''$*, then* $K \leq^F_{G/X} K''$*.*
2. *If* $K \leq^F_{G/X} K'$ *and* $K' <^F_{G/X} K''$*, then* $K <^F_{G/X} K''$*.*
3. *If* $K <^F_{G/X} K'$ *and* $K' \leq^F_{G/X} K''$*, then* $K <^F_{G/X} K''$*.*
4. *If* $K <^F_{G/X} K'$ *and* $K' <^F_{G/X} K''$*, then* $K <^F_{G/X} K''$*.*

Proof. To prove Proposition 3, we first prove two lemmas:
Lemma A Let F, G be groups of agents from a consequentialist frame, X a set of worlds in the frame. Let $K, K' \in Choice(G/X)$. If $K \leq^F_{G/X} K'$, then $Choice((A-G)/(X \cap K))$=$Choice((A-G)/(X \cap K'))$.

Proof of Lemma A We are going to prove $Choice((A-G)/(X \cap K)) \subseteq Choice((A-G)/(X \cap K'))$ and $Choice((A-G)/(X \cap K)) \supseteq Choice((A-G)/(X \cap K'))$.

For $Choice((A-G)/(X \cap K)) \subseteq Choice((A-G)/(X \cap K'))$, assume there exists some $S \in Choice(A-G)$ such that $S \in Choice((A-G)/(X \cap K))$ but $S \notin Choice((A-G)/(X \cap K'))$. Then $S \cap X \cap K \neq \emptyset$ and $S \cap X \cap K' = \emptyset$. Therefore $S \cap X \cap (K \cup K') \neq \emptyset$ and $S \in Choice((A-G)/(X \cap (K \cup K')))$. Now by $K \leq^F_{G/X} K'$ we have $K \cap X \cap S \leq_F K' \cap X \cap S$. This plus Lemma 1 implies $K' \cap X \cap S \neq \emptyset$. Contradiction. Hence $Choice((A-G)/(X \cap K)) \subseteq Choice((A-G)/(X \cap K'))$.

The case for $Choice((A-G)/(X \cap K)) \supseteq Choice((A-G)/(X \cap K'))$ is similar.

Lemma B Let G be a group of agents, X and Y be sets of worlds from a consequentialist frame. If $Choice((A-G)/X) = Choice((A-G)/Y)$, then $Choice((A-G)/X) = Choice((A-G)/(X \cup Y))$.

Proof of Lemma B For $Choice((A-G)/X) \subseteq Choice((A-G)/(X \cup Y))$. If $S \in Choice((A-G)/X)$, then $S \in Choice((A-G)/Y)$ and $S \cap X \neq \emptyset$, $S \cap Y \neq \emptyset$. Hence $(S \cap X) \cup (S \cap Y) \neq \emptyset$, $S \cap (X \cup Y) \neq \emptyset$. So we have $S \in Choice((A-G)/(X \cup Y))$.

For $Choice((A-G)/X) \supseteq Choice((A-G)/(X \cup Y))$. If $S \in Choice((A-G)/(X \cup Y))$, then $S \cap (X \cup Y) \neq \emptyset$, $(S \cap X) \cup (S \cap Y) \neq \emptyset$. Now assume $S \notin Choice((A-G)/X)$, then $S \notin Choice((A-G)/Y)$. Hence $S \cap X = \emptyset$ and $S \cap Y = \emptyset$. So we have $(S \cap X) \cup (S \cap Y) = \emptyset$. Contradiction.

Proof of Proposition 3 Here we just prove clause 1. Other clauses are similar. Assume $K \leq^F_{G/X} K'$ and $K' \leq^F_{G/X} K''$. By Lemma A,

$Choice((A-G)/(X \cap K)) = Choice((A-G)/(X \cap K')) = Choice((A-G)/(X \cap K''))$.
By Lemma B, we now have

$Choice((A-G)/(X \cap K)) = Choice((A-G)/((X \cap K) \cup (X \cap K'))) = Choice((A-G)/(X \cap (K \cup K')))$,

$Choice((A-G)/(X \cap K)) = Choice((A-G)/((X \cap K) \cup (X \cap K''))) = Choice((A-G)/(X \cap (K \cup K'')))$,

$Choice((A-G)/(X \cap K')) = Choice((A-G)/((X \cap K') \cup (X \cap K''))) = Choice((A-G)/(X \cap (K' \cup K'')))$.
Hence for each $S \in Choice((A-G)/(X \cap (K \cup K'')))$, we have $S \in Choice((A-G)/(X \cap (K \cup K')))$ and $S \in Choice((A-G)/(X \cap (K' \cup K'')))$. Therefore by $K \leq^F_{G/X} K'$ we have $K \cap X \cap S \leq_F K' \cap X \cap S$, by $K' \leq^F_{G/X} K''$ we have $K' \cap X \cap S \leq_F K'' \cap X \cap S$. Since the relation $\leq_F$ is transitive, we have $K \cap X \cap S \leq_F K'' \cap X \cap S$. Therefore $K \leq^F_{G/X} K''$. □

Proposition 3 corresponds to Proposition 5.4 in Horty (2001). Notice that to make Proposition 3 true, the existential condition in Definition 3 is necessary. [3]

[3] In Horty (2001) $\leq_F$ is defined with the universal clause in Definition 3 only. $K \leq^F_{G/X} K'$ is defined as: for all $S \in Choice(A-G)$, $K \cap X \cap S \leq_F K' \cap X \cap S$. In that case we can construct the following counterexample to falsify the transitivity of $\leq^F_{G/X}$:

	S_1	S_2
K_1	w_1 (2)	w_2 (0) X
K_2	w_3 (3) X	w_4 (1)
K_3	w_5 (3)	w_6 (1) X

Figure 4

In Kooi and Tamminga (2008b), the definitions of restricted choice set and conditional dominance are different from ours. See the following:

Definition 2.7. Let F, G be groups of agents from a consequentialist frame, X a set of worlds in the frame. Then

$$Choice(G/X) = \{K \cap X : K \in Choice(G) \text{ and } K \cap X \neq \emptyset\}$$

Definition 2.8. Let F, G be groups of agents from a consequentialist frame, X a set of worlds in the frame. Let $k, k' \in Choice(G/X)$. Then

$$k \leq^F_{G/X} k' \quad \text{iff} \quad \begin{array}{l}\text{for all } S \in Choice(A - G) \text{ and for all } w \text{ and } w' \in W \text{ it holds} \\ \text{that if } w \in k \cap S \text{ and } w' \in k' \cap S, \text{ then } Value_F(w) \leq Value_F(w')\end{array}$$

Look at Figure 4 again, it's easy to verify that according to above definitions, $Choice(G/X) = \{\{w_2\}, \{w_3\}, \{w_6\}\}$ and $\{w_6\} \leq^F_{G/X} \{w_3\}$, $\{w_3\} \leq^A_{\{\alpha\}/X} \{w_2\}$, but $\{w_6\} \leq^A_{\{\alpha\}/X} \{w_2\}$ does not hold. In other words, this version of conditional dominance is not transitive.

2.3 Semantics

As in traditional modal logic, a model is a frame plus a valuation function.

Definition 2.9 (consequentialist model)**.** A consequentialist model M is an ordered pair $\langle F, V\rangle$ where F is a consequentialist frame and V a valuation function that assigns to each atomic proposition $p \in P$ a set of worlds $V(p) \subseteq W$.

In our semantics, we use the optimal choice and conditional optimal choice to interpret the deontic operators. The definition of optimal (Definition 2.10) and conditional optimal (Definition 2.11) is rather simple, we will show it very soon.

Definition 2.10 ($Optimal^F_G$)**.** Let F, G be groups of agents from a consequentialist frame,

Here $W = \{w_1, ..., w_6\}$, $A = \{\alpha, \beta\}$, $Choice(\{\alpha\}) = \{K_1, K_2, K_3\}$, $Choice(\{\beta\}) = \{S_1, S_2\}$, $K_1 = \{w_1, w_2\}$, $K_2 = \{w_3, w_4\}$, $K_3 = \{w_5, w_6\}$, $S_1 = \{w_1, w_3, w_5\}$, $S_2 = \{w_2, w_4, w_6\}$, $X = \{w_2, w_3, w_6\}$, and the number in the brackets represents the value of the world in the interest of group A. According to Horty's definition, we have $K_3 \leq^A_{\{\alpha\}/X} K_2$ and $K_2 \leq^A_{\{\alpha\}/X} K_1$ but we don't have $K_3 \leq^A_{\{\alpha\}/X} K_1$. Therefore, in Horty (2001) the $\leq^F_{G/X}$ relation is not transitive and Proposition 5.4 is mistaken.

$$Optimal^F_G = \{K \in Choice(G) : \text{there is no } K' \in Choice(G) \text{ such that } K <^F_G K'\}.$$

Proposition 4. *Let F, G be groups of agents from a consequentialist frame, then for each $K \in Choice(G) - Optimal^F_G$, there exists some $K' \in Optimal^F_G$ such that $K <^F_G K'$.*

Proof. Similar to the proof of Proposition 4.11 in (Horty 2001) □

Definition 2.11 ($Optimal^F_{G/X}$)**.** Let F, G be groups of agents from a consequentialist frame,

$$Optimal^F_{G/X} = \{K \in Choice(G/X) : \text{there's no } K' \in Choice(G/X) \text{ such that } K <^F_{G/X} K'\}.$$

Proposition 5. *Let F, G be groups of agents from a consequentialist frame, for each $K \in Choice(G/X) - Optimal^F_{G/X}$, there exists $K' \in Optimal^F_{G/X}$ such that $K <^F_{G/X} K'$.*

Proof. Similar to the proof of Proposition 5.7 in Horty (2001). □

Here again to ensure the truth of Proposition 5, the existential condition in Definition 2.3 is necessary. Otherwise we would have a counterexample as follows:

	S_1	S_2	S_3
K_1	w_1 (3) X	w_2 (0) X	w_3
K_2	w_4	w_5 (1) X	w_6 (4) X
K_3	w_7 (2) X	w_8	w_9 (5) X

Figure 5

In the above case, $W = \{w_1, ..., w_9\}$, $A = \{\alpha, \beta\}$, $Choice(\{\alpha\}) = \{K_1, K_2, K_3\}$, $Choice(\{\beta\}) = \{S_1, S_2, S_3\}$, $K_1 = \{w_1, w_2, w_3\}$, $K_2 = \{w_4, w_5, w_6\}$, $K_3 = \{w_7, w_8, w_9\}$, $S_1 = \{w_1, w_4, w_7\}$, $S_2 = \{w_2, w_5, w_8\}$, $S_3 = \{w_3, w_6, w_9\}$, and $X = \{w_1, w_2, w_5, w_6, w_7, w_9\}$, the numbers in the brackets represent the value of the world in the interest of A. If Definition 3 did not have the existential condition, we would have $K_1 <^A_{\{\alpha\}/X} K_2$, $K_2 <^A_{\{\alpha\}/X} K_3$, and $K_3 <^A_{\{\alpha\}/X} K_1$. Hence $Optimal^A_{\{\alpha\}/X} = \emptyset$, which contradicts to Proposition 5.

Definition 2.12 (truth conditions)**.** Let $M = \langle \boldsymbol{F}, V\rangle$ be consequentialist model. Let $w \in W$ rand let $\varphi, \psi \in \mathfrak{L}$. Then

(1) $M, w \models p$	iff	$w \in$V(p);
(2) $M, w \models \neg\varphi$	iff	it is not that $M, w \models \varphi$;
(3) $M, w \models \varphi \wedge \psi$	iff	$M, w \models \varphi$ and $M, w \models \psi$;
(4) $M, w \models \Diamond\varphi$	iff	there is a w' such that $M, w' \models \varphi$;
(5) $M, w \models [G]\varphi$	iff	for all w' with $w \sim_G w'$ it holds that $M, w' \models \varphi$;
(6) $M, w \models \bigodot^F_G \varphi$	iff	$K \subseteq \|\varphi\|$ for each $K \in Optimal^F_G$;
(7) $M, w \models \bigodot^F_G(\varphi/\psi)$	iff	$K \subseteq \|\varphi\|$ for each $K \in Optimal^F_{G/\psi}$.

Here $\|\varphi\| = \{w \in W : M, w \models \varphi\}$. $Optimal^F_{G/\psi}$ is a shorthand for $Optimal^F_{G/\|\psi\|}$.

We say φ is true in the world w of a consequentialist model M if $M, w \models \varphi$. Just like in the standard modal logic (for instance, P. Blackburn and Venema (2001)), we introduce the concept of validity as following: a formula φ is valid in a world w of a consequentialist frame $\boldsymbol{F}$ (notation: $\boldsymbol{F}, w \models \varphi$) if φ is true at w in every model $\langle \boldsymbol{F}, V\rangle$ based on $\boldsymbol{F}$; φ is valid in a consequentialist frame $\boldsymbol{F}$ (notation: $\boldsymbol{F} \models \varphi$) if it is valid at every world of $\boldsymbol{F}$; φ is valid (notation: $\models \varphi$) if it is valid in the class of all consequentialist frames.

Let us summarize what we have done so far. Motivated by the two scenarios that are prevalent in ordinary life, we have proposed a new logic characterizing the process of decision making, in which our new definition of preference over sets of worlds ensures its property of transitivity. We have explored its logical properties in details.

3 A Solution to the Ten Miners Paradox by "Sure-thing Reasoning"

Having defined our logic, now it's time to test it by dealing with puzzles and paradoxes. A well-circulated paradox involved conditional ought is presented

by Kolodny and MacFarlane (2010). It goes as follows:

Ten miners are trapped either in shaft A or in shaft B, but we don't know which. Flood waters threaten to flood the shafts. We have enough sandbags to block one shaft, but not both. If we block one shaft, all the water will go into the other shaft, killing any miners insider it. If we block neither shaft, both shafts will fill halfway with water, and just one miner, the lowest in the shaft, will be killed.

From a consequentianlist's perspective, we take it as obvious that

(1) We ought to block neither shaft.

However, the following two statements also seems plausible:

(2) If the miners are in shaft A, we ought to block shaft A.
(3) If the miners are in shaft B, we ought to block shaft B.

Since the miners must be either in shaft A or in shaft B, we can conclude either we ought to block shaft A or we ought to block shaft B. Furthermore,

(4) We ought to block at least one shaft.

Here (1) and (4) contradict to each other.

Kolodny and MacFarlane (2010) has extensively discussed some possible solutions to the paradox. Here we simply present another solution using ideas from our new logic. First, we reformulate the above statement (1)-(4) with our consequentianlist deontic language.

Let φ denote the miners are in shaft A, then $\neg\varphi$ denote the miners are in shaft B. Let ψ' denote we block shaft A, ψ'' denote we block shaft B. Let ψ denote we block at least one of the two shafts. Then
(2) can be formulated as $\bigodot^F_G(\psi'/\varphi)$.
(3) can be formulated as $\bigodot^F_G(\psi''/\neg\varphi)$.
(4) can be formulated as $\bigodot^F_G \psi$.

The reasoning from the premiss (2) and (3) to the conclusion (4) is the following:

From $\bigodot^F_G(\psi'/\varphi)$ we infer that $\bigodot^F_G(\psi/\varphi)$, from $\bigodot^F_G(\psi''/\neg\varphi)$ we infer that $\bigodot^F_G(\psi/\neg\varphi)$. From $\bigodot^F_G(\psi/\varphi)$ and $\bigodot^F_G(\psi/\neg\varphi)$ we infer that $\bigodot^F_G \psi$.

We are going to solve this paradox by proving that the deduction from $\bigodot^F_G(\psi/\varphi)$

and $\bigodot^F_G(\psi/\neg\varphi)$ to $\bigodot^F_G \psi$ is not valid. According to our semantics, it's not hard to verify that the formula $(\bigodot^F_G(\psi/\varphi) \wedge \bigodot^F_G(\psi/\neg\varphi)) \rightarrow \bigodot^F_G \psi$ is not valid. For illustration, we can easily construct the following model:

	block A	block B	block neither
In A	w_1 (10) φ, ψ	w_2 (0) φ, ψ	w_3 (9) φ
In B	w_4 (0) $\neg\varphi$, ψ	w_5 (10) $\neg\varphi$, ψ	w_6 (9) $\neg\varphi$

Figure 6

In Figure 6, we let group G have three choices: block shaft A, block shaft B or block neither shaft. And we view nature as an agent and let it have two choices: set miners in shaft A or set miners in shaft B. Let the numbers in the brackets represent group F's preference (no matter what Group F is).

According to the above model, $Optimal^F_{G/\varphi}$=$\{\{w_1, w_4\}\}$, $Optimal^F_{G/\neg\varphi}$ =$\{\{w_2, w_5\}\}$. Hence $\{w_1, w_4\} \subseteq \|\psi\|$,$\{w_2, w_5\} \subseteq \|\psi\|$. Therefore both $(\bigodot^F_G(\psi/\varphi)$ and $\bigodot^F_G(\psi/\neg\varphi))$ are true in this model. But we also have $\{w_3, w_6\} \in Optimal^F_G$ and $\{w_3, w_6\}$ is not a subset of $\|\psi\|$. So $\bigodot^F_G \psi$ is not true in this model. Therefore the formula $(\bigodot^F_G(\psi/\varphi) \wedge \bigodot^F_G(\psi/\neg\varphi)) \rightarrow \bigodot^F_G \psi$ is not valid.

In Horty (2001), deductions of the above form are called "sure-thing reasoning". Although sure-thing reasoning seems plausible at first sight, Horty (2001) pointed out that it is correct only if the conditional statement is *independent* from the action statement. Otherwise strange situation will arise. As we have seen, this condition is indeed crucial for us to underpin the real problem of the Ten Miners Paradox. To even better illustrate this point, we give one more example in terms of games.

		Agent 2		
		Guess H	Guess T	Refrain
Agent 1	Head	w_1 (10)	w_2 (-10) ψ	w_3 (0)
	Tail	w_4 (-10)	w_5 (10) ψ	w_6 (0)

Figure 7

The background of Figure 7 is a gambling game. Agent 1 flips a coin. If agent 2 refrain from guessing which side is the up side, then agent 2 receives nothing meanwhile loses nothing. If agent 2 guesses one side then she wins 10 dollars when her answer is right and loses 10 dollars otherwise. According to our intuitions, we can not blame agent 2 if she chooses to refrain. However, applying the principle of sure-thing reasoning we will arrive at the conclusion that agent 2 ought to guess. Here is the reasoning: If the up side is head, then agent 2 should guess head, which implies she should guess. If the up side is tail, then agent 2 should guess tail, which also implies she should guess. Now use the sure-thing reasoning, we know agent 2 ought to guess.

4 Revisit Anderson's Six Principles of Conditional Ought

Anderson (1959) suggested that the logic of commitment or conditional ought should satisfy six principles, and we state them in our language as follows:

(1) $(\psi \wedge \bigodot^F_G(\varphi/\psi)) \to \bigodot^F_G \varphi$.

(2) $(\bigodot^F_G \psi \wedge \bigodot^F_G(\varphi/\psi)) \to \bigodot^F_G \varphi$.

(3)$(\boldsymbol{P}^F_G\psi \wedge \bigodot^F_G(\varphi/\psi)) \to \boldsymbol{P}^F_G\varphi$.

(4) $(\bigodot^F_G(\varphi/\psi) \wedge \bigodot^F_G(\chi/\varphi)) \to \bigodot^F_G(\chi/\psi)$

(5) $\bigodot^F_G(\varphi/\psi) \to \bigodot^F_G(\psi \to \varphi)$

(6) $\bigodot^F_G(\varphi/\neg\varphi) \to \bigodot^F_G \varphi$

The strongly normal system ***G*** in Aqvist (1994) excludes the principle (1) and (4), but includes the rest. In our logic, however, only principle (2) and (6) are valid. The invalidity of principle (1) and (4) and the validity of principle (6) are easy to prove, here we skip it. As for the rest, the following hold:

Theorem 1. *The statement* $(\bigodot^F_G \psi \wedge \bigodot^F_G(\varphi/\psi)) \to \bigodot^F_G \varphi$ *is valid.*

Proof. Assume this formula is not valid, then there is a consequentialist frame $\boldsymbol{F}$ and a world w in $\boldsymbol{F}$ such that for some model M based on $\boldsymbol{F}$, $M, w \vDash \bigodot^F_G \psi \wedge \bigodot^F_G(\varphi/\psi)$ but $M, w \nvDash \bigodot^F_G \varphi$. Hence there must be some $K \in Optimal^F_G$ such that $K \nsubseteq \|\varphi\|$ but $K \subseteq \|\psi\|$.

As $K \subseteq \|\psi\|$, for arbitrary $S \in Choice(A-G)$, for arbitrary $K' \in Choice(G)$, we must have $S \cap (\|\psi\| \cap (K \cup K')) \neq \emptyset$ because $S \cap K \subseteq S \cap (\|\psi\| \cap (K \cup K'))$ and $S \cap K \neq \emptyset$ by the definition of choice function. Hence $S \in Choice((A - G)/(\|\psi\| \cap (K \cup K')))$ and $Choice(A - G) = Choice((A - G)/(\|\psi\| \cap (K \cup K')))$.

Obviously either $K \in Optimal^F_{G/\psi}$ or $K \notin Optimal^F_{G/\psi}$. If $K \in Optimal^F_{G/\psi}$, then by $M, w \vDash \bigodot^F_G(\varphi/\psi)$, $K \subseteq \|\varphi\|$. Contradiction. Therefore $K \notin Optimal^F_{G/\psi}$. Then by Proposition 5 there exists some $K' \in Choice(G/\psi)$ with $K <^F_{G/\psi} K'$. It must be either $K' \in Optimal^F_G$ or $K' \notin Optimal^F_G$. We can show both these two cases imply contradictions.

For the first case, assume $K' \in Optimal^F_G$. Then by $M, w \vDash \bigodot^F_G \psi$, $K' \subseteq \|\psi\|$. According to $K <^F_{G/\psi} K'$, for each $S \in Choice((A-G)/(\|\psi\| \cap (K \cup K')))$, $K \cap S \cap \|\psi\| \leq_F K' \cap S \cap \|\psi\|$. Since $K \subseteq \|\psi\|$ and $K' \subseteq \|\psi\|$, we know $K \cap S \cap \|\psi\| = K \cap S$, $K' \cap S \cap \|\psi\| = K' \cap S$. Therefore $K \cap S \leq_F K' \cap S$. Note we have already proved $Choice(A-G) = Choice((A-G)/(\|\psi\| \cap (K \cup K')))$, hence for each $S \in Choice(A-G)$, $K \cap S \leq_F K' \cap S$. That is, $K \leq^F_G K'$. By Lemma 4, $K <^F_{G/\psi} K'$ also implies that for some $S \in Choice((A-G)/(\|\psi\| \cap (K \cup K'))) = Choice(A-G)$, $K \cap S \cap \|\psi\| <_F K' \cap S \cap \|\psi\|$. Use $K \cap S \cap \|\psi\| = K \cap S$, $K' \cap S \cap \|\psi\| = K' \cap S$ one more time we have the

conclusion that for some $S \in Choice(A - G)$, $K \cap S <_F K' \cap S$. Now by Lemma 3 we know $K <^F_G K'$, which contradicts to $K \in Optimal^F_G$.

For the second case, assume $K' \notin Optimal^F_G$. Then there exists some $K'' \in Choice(G)$ with $K' <^F_G K''$. So we have for each $S \in Choice(A - G) = Choice((A - G)/(\|\psi\| \cap (K \cup K')))$, $K' \cap S \leq_F K'' \cap S$. By $K <^F_{G/\psi} K'$, $K \cap S \cap \|\psi\| \leq_F K' \cap S \cap \|\psi\|$. It then follows from Lemma 1 that $K' \cap S \cap \|\psi\| \neq \emptyset$. This plus $K' \cap S \leq_F K'' \cap S$ implies $K' \cap S \cap \|\psi\| \leq_F K'' \cap S$. Note that $K \cap S = K \cap S \cap \|\psi\|$ since $K \subseteq \|\psi\|$, therefore $K \cap S \leq_F K' \cap S \cap \|\psi\|$. Now by Proposition 1 we have $K \cap S \leq_F K'' \cap S$. Hence $K \leq^F_G K''$. By Lemma 4, $K <^F_{G/\psi} K'$ also implies for some $S \in Choice((A - G)/(\|\psi\| \cap (K \cup K'))) = Choice(A - G)$, $K \cap S \cap \|\psi\| <_F K' \cap S \cap \|\psi\|$. Use $K \cap S = K \cap S \cap \|\psi\|$ again we have $K \cap S <_F K' \cap S \cap \|\psi\|$. Since $K' \cap S \leq_F K'' \cap S$ and $K' \cap S \cap \|\psi\| \neq \emptyset$, we have $K' \cap S \cap \|\psi\| \leq_F K'' \cap S$. By Proposition 1 we now have $K \cap S <_F K'' \cap S$. It then follow from Lemma 3 that $K <^F_G K''$, contradict to $K \in Optimal^F_G$. □

Theorem 2. *The statement* $(\boldsymbol{P}^F_G\psi \wedge \bigodot^F_G(\varphi/\psi)) \rightarrow \boldsymbol{P}^F_G\varphi$ *is not valid.*

Proof. It's sufficient to construct a model M such that for some world w in M, $M, w \models \boldsymbol{P}^F_G\psi \wedge \bigodot^F_G(\varphi/\psi)$ but $M, w \not\models \boldsymbol{P}^F_G\varphi$.

As illustrated by Figure 5. Let $M = \langle W, A, Choice, \{Value_F\}_{F \subseteq A}, V\rangle$, $W = \{w_1, ..., w_6\}$, $A = \{\alpha, \beta\}$, $Choice(\{\alpha\}) = \{\{w_1, w_2\}, \{w_3, w_4\}, \{w_5, w_6\}\}$, $Choice\ (\{\beta\}) = \{\{w_1, w_3, w_5\}, \{w_2, w_4, w_6\}\}$, $Value_{\{\alpha\}}\ (w_1) = 100$, $Value_{\{\alpha\}}(w_2) = 0$, $Value_{\{\alpha\}}(w_3) = 20$, $Value_{\{\alpha\}}(w_4) = 30$, $Value_{\{\alpha\}}(w_5) = 50$, $Value_{\{\alpha\}}\ (w_6) = 50$. Let $F = \{\alpha\}$, $G = \{\alpha\}$, $\|\varphi\| = \{w_3, w_4\}$, $\|\psi\| = \{w_2, w_4\}$. Then $Optimal^F_G = \{\{w_1, w_2\}, \{w_5, w_6\}\}$, $Optimal^F_{G/\psi} = \{\{w_3, w_4\}\}$. Note that $M, w \models \boldsymbol{P}^F_G\varphi$ if and only if for some $K \in Optimal^F_G$, $K \cap \|\varphi\| \neq \emptyset$. Hence by the semantics we have $M, w_1 \models \boldsymbol{P}^F_G\psi$ and $M, w_1 \models \bigodot^F_G(\varphi/\psi)$. But as $\{w_1, w_2\} \cap \|\varphi\| = \emptyset$ and $\{w_3, w_4\} \cap \|\varphi\| = \emptyset$, we have $M, w_1 \not\models \boldsymbol{P}^F_G\varphi$. □

Theorem 3. *The statement* $\bigodot^F_G(\varphi/\psi) \rightarrow \bigodot^F_G(\psi \rightarrow \varphi)$ *is not valid.*

Proof. It's sufficient to construct a model M such that for some world w in M, $M, w \models \bigodot^F_G(\varphi/\psi)$ but $M, w \not\models \bigodot^F_G(\psi \rightarrow \varphi)$.

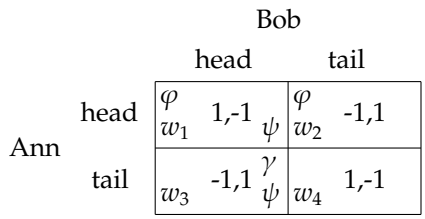


Figure 8

Look at Figure 8. Let $M = \langle W, A, Choice, \{Value_F\}_{F\subseteq A}, V\rangle$, $W = \{w_1, ..., w_4\}$, $A = \{\alpha, \beta\}$, $Choice$ $(\{\alpha\}) = \{\{w_1, w_2\}, \{w_3, w_4\}\}$, $Choice(\{\beta\}) = \{\{w_1, w_3\}, \{w_2, w_4\}\}$, $Value_{\{\alpha\}}(w_1) = 1$, $Value_{\{\alpha\}}$ $(w_2) = -1$, $Value_{\{\alpha\}}(w_3) = -1$, $Value_{\{\alpha\}}(w_4) = 1$. Let $F = \{\alpha\}$, $G = \{\alpha\}$, $\|\varphi\| = \{w_1, w_2\}$, $\|\psi\| = \{w_1, w_3\}$. In this situation, $Optimal^F_{G/\psi} = \{\{w_1, w_2\}\}$, hence $M, w_1 \models \bigodot^F_G(\varphi/\psi)$. As $Optimal^F_G = \{\{w_1, w_2\}, \{w_3, w_4\}\}$ but $\{w_3, w_4\} \not\subseteq \|\psi \to \varphi\|$, we have $M, w_1 \not\models \bigodot^F_G(\psi \to \varphi)$. □

The invalidity of principle (3) can be clearly illustrated by a variation of the matching pennies game.

Example 4 (matching pennies: a variation)**.** In this new game, Ann has three choices, showing the head, showing the tail and refraining from showing. If Ann refrains from showing, then no matter how Bob acts, Ann will receive 50 dollars. The payoff of Ann is indicated by numbers in brackets in Figure 9.

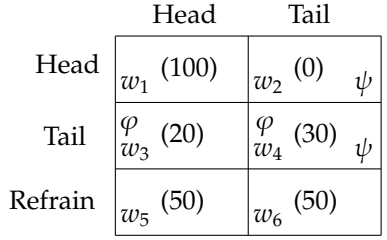


Figure 9

Denote the situation in which Ann shows one side of her penny and Bob shows the tail as situation ψ. Ann is permitted to lead to situation ψ, since showing the head is one of Ann's optimal actions and this action could lead to situation

ψ. Apparently, in situation ψ, Ann ought to see to it that she shows the tail. But Ann is not permitted to show the tail because this action is dominated by refraining.

5 Related Works and Comparison

In this section we will briefly compare our framework with other related works in the literature. Horty (2001) developed a new stit-based deontic logic and provided a uniformed view on many interesting issues from moral theory. Inspired by it, Kooi and Tamminga (2008a) and Kooi and Tamminga (2008b) proposed a consequentianlist deontic logic which can be used to analyze moral conflicts between different groups of agents with different moral codes. Our proposal is directly based on those works. We have shown in the previous sections that our new development fits well with the real-life examples and games, some potential technical problems of conditional ought that existed in those previous works can be solved in our new framework, and our logic can be best used to solve some new moral dilemma. In addition, we have also looked at how our framework is related to the original ideas on conditional ought in Anderson's work. In what follows we look at the latest work Tamminga (2011 Nov. 24) on consequentianlist deontic logic and provide some discussion from our point of view.

The logic presented in Tamminga (2011 Nov. 24) is called multi-agent deontic action logic. We will briefly review the main notions of that work and compare it with ours.

In that logic, actions of agents are treated as atomic and conditions are restricted to atomic action propositions. More exactly, the language of multi-agent deontic action logic is built from a finite set A of agents, a countable set $\mathfrak{P}$ of atomic propositions and another countable set $\mathfrak{A} = \{\alpha_G^n : G \subseteq A$ and n is a nature number $\}$ of atomic actions. Again, we use p and q as variables for atomic propositions in $\mathfrak{P}$, use F and G, where $F, G \subseteq A$, as groups of agents, use α_G, α_F as variables for atomic atomic propositions in $\mathfrak{A}$. The multi-agent deontic action language $\mathfrak{L}^a$ is given by the following Backus-Naur Form:

$$\varphi ::= p \mid \alpha_G \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Diamond\varphi \mid [G]\varphi \mid \bigodot_G^F \alpha_G \mid P_G^F\alpha_G \mid \bigodot_G^F(\alpha_G/\alpha_H) \mid P_G^F(\alpha_G/\alpha_H)$$

Here α_G can be read as "Group G performs action α_G". In the formula

$\bigodot_G^F(\alpha_G/\alpha_H)$, it is required that $G \cap H = \emptyset$.

The semantics of multi-agent deontic action logic is different from our consequentianlist deontic logic. In that logic, the author's intuition of a group's action K being obligatory is "it is the single best thing the group can do" and one action being permitted if and only if it is "among the best thing the group can do." This intuition makes sense to us. However, it seems that the formal definition of the permission operator causes some unwilling results. We will show it now.

In Tamminga (2011 Nov. 24), $P_G^F\alpha_G$ is true if and only if for every group G's action K which is different from α_G, $K \leq_G^F \alpha_G$ holds. This means that an action is permitted if and only if it weakly dominates every other action. But in fact, we may have more than one optimal choices. For example, in Figure 2 both swerve and continue are the driver' optimal choices. If we analyze this situation by using the semantics given by Tamminga (2011 Nov. 24). Driver 1's action swerve is not permitted because his action 'continue' is not weakly dominated by 'swerve'. For the same reason, 'continue' is also not permitted. Which contradicts to our intuition. So in our opinion, the formal definition of permission in multi-agent deontic action logic does not express its background intuition very well.

However, it turns out that if we restrict deontic operators to action statements, then our logic can be viewed as another version of multi-agent deontic action logic and it can express our intuition even better. In our semantics, $P_G^F\varphi$ is true if and only if for some $K \in Optimal_G^F$, $K \cap \|\varphi\| \neq \emptyset$. If we constrain ourselves to action statements, then $P_G^F\alpha_G$ is true if and only if $\alpha_G \in Optimal_G^F$. This means an action is permitted if and only if it is "among the best thing the group can do." Since there may be many optimal actions in a model, we can avoid the unpleasant result arisen from Figure 2.

6 Conclusion and Future Work

In this paper we started with two examples to motivate the idea that the principle that conditional ought implies ought about conditionals is not right. We then introduced a new consequentianlist deontic logic, with a new definition of preference over sets of worlds, the definition of conditional dominance relation was defined too, they both are transitive. We have shown that our logic

supports the principle that absolute ought can be derived from conditional ought whenever the conditional statement is the agent's absolute ought. And, according to our semantics, conditional ought does not imply ought about conditionals. To apply the ideas and results of our logic, we looked at a recent paradox and provided our solution. We then studied the six principles proposed by Anderson and proved that some of them are valid in our framework and some are not. Finally, we have compared our work with the very recent work on multi-agent deontic action logic and proposed a better interpretation for the notion of permission in that logic.

There are several issues we would like to explore in the future. First, we have proved many results in the context of semantic frames, and we want to see whether we can find a complete axiomatic system for the consequentianlist deontic logic. Next, introducing epistemic modality in the current framework to express the uncertainties seems to be a natural direction, one can see already from our above discussions. Related ideas can be found in E. Pacuit and Cogan (2006) and Loohuis (2009). Finally, we only looked at the $\forall\forall$-version of the preference relation in this paper, a question arises naturally here, are other options (say, $\forall\exists$-version) suitable, and how to make use of them game theoretically? Here again, notable references are van der Torre (1997), Halpern (1997), van Benthem et al. (2010), and Roy (2011), we would like to connect the current work to theirs.

Acknowledgements We thank Sonja Smets for her invitation to submit this paper to the ILLC LIRa yearbook. There have been several versions of this work, we want to thank Johan van Benthem, John Horty, Leon van der Torre, Xavier Parent, Allard Tamminga and the anonymous referees of the LORI conference for their useful comments.

References

A. Anderson. On the logic of commitment. *Philosophical Studies*, 10:23–27, 1959.

L. Aqvist. Deontic logic. *Handbook of Philosophical Logic*, 2:147–264, 1994.

R. P. E. Pacuit and E. Cogan. The logic of knowledge based obligation. *Synthese*, 149:311–341, 2006.

J. Halpern. Defining relative likelihood in partially-ordered preferential structures. *Journal of Artificial Intelligence Research*, 7:1–24, 1997.

J. Horty. *Agency and Deontic Logic*. Oxford University Press, New York, 2001.

N. Kolodny and J. MacFarlane. Ifs and oughts. *Journal of Philosophy*, 107: 115–143, 2010.

B. Kooi and A. Tamminga. Moral conflicts between groups of agents. *Journal of Philosophical Logic*, 37:1–21, 2008a.

B. Kooi and A. Tamminga. Conditioanl obligations in strategic situations. In M. S. G. Boella, G Pigozzi and H. Verhagen, editors, *3rd International Workshop on Normative Multiagent Systems(NorMAS 2008)*, pages 188–200, Luxembourg, July 2008b.

L. Loohuis. Obligations in a responsible world. In J. H. X. He and E. Pacuit, editors, *Logic, Rationality, and Interaction: Proceedings of the Second International Workshop (LORI)*, Chongqing, China, October 2009.

M. Osborne and A. Rubinstein. *A Course in Game Theory*. The MIT Press, Cambridge, Mass., 1994.

M. R. P. Blackburn and Y. Venema. *Modal Logic*. Cambridge University Press, Cambridge, 2001.

O. Roy. Deontic logic, neighborhood semantics and games. Technical Report at the Workshop on Logical Dynamics: New Trends and Interfaces, CWI, Amsterdam, 2011.

X. Sun. Conditional ought, a game theoretical perspective. In J. L. H. van Ditmarsch and S. Ju, editors, *Logic, Rationality, and Interaction: Proceedings of the Thire International Workshop*, pages 356–369, Guangzhou, China, October 2011.

A. Tamminga. Deontic logic for strategic games. *Erkenntnis*, pages 1–18, 2011 Nov. 24.

J. van Benthem, D. Grossi, and F. Liu. Deontics = betterness + priority. In G. Governatori and G. Sartor, editors, *Proceedings of Deontic Logic in Computer Science, 10th International Conference*, pages 50–65, Fiesole, Italy, July 2010.

L. van der Torre. *Reasoning about Obligations: Defeasibility in Preference-based Deontic Logic*. PhD thesis, Erasmus University Rotterdamm, Rotterdam, The Netherlands, 1997.

Dynamic Epistemic Logic for Implicit and Explicit Beliefs

Fernando R. Velázquez-Quesada

Facultad de Filosofía, Universidad de Sevilla
FRVelazquezQuesada@us.es

Abstract

Epistemic logic and its possible worlds semantics is a powerful and compact framework that allows us to represent an agent's information not only about propositional facts, but also about her own information. Nevertheless, agents represented in this framework are *logically omniscient*: their information is closed under logical consequence. This property, useful in some applications, is an unrealistic idealization in some others. And though most of the proposals to solve this problem focus on weakening the properties of the agent's information (usually by distinguishing between implicit and explicit information), some authors have argued that solutions of this kind are not completely adequate because they do not look at the heart of the matter: the actions that allow the agent to reach such omniscient state.

Recent works have explored how acts of observation, inference, consideration and forgetting affect an agent's implicit and explicit *knowledge*. The present work focusses on acts that affect the notions of implicit and explicit *beliefs*. We start by proposing a framework in which we can represent these two notions, and then we look into their dynamics, first by reviewing the existing notion of *belief revision*, and then by introducing a rich framework that allow us to represent diverse forms of inference that involve not only knowledge but also beliefs.

1 Introduction

Classical epistemic logic (*EL*; Hintikka (1962)) and its possible worlds semantics is a powerful framework that allows us to represent an agent's information not only about propositional facts but also about her own information.

Nevertheless, agents represented in this framework are *logically omniscient*: their information is closed under logical consequence. This property, useful in some applications, is an unrealistic idealization in some others. Many proposals to solve this issue are based on weakening the properties of the agent's information, most of them based on distinguishing between the agent's potential *implicit* information, describing what she can eventually get, and her actual *explicit* information, describing what she currently has (cf. Konolige (1984), Levesque (1984), Lakemeyer (1986), Vardi (1986), Fagin and Halpern (1988)).

But some authors (Drapkin and Perlis (1986), Duc (1995), van Benthem and Velázquez-Quesada (2010) among others) have argued that solutions of this kind are not completely adequate. First, an agent's information can be weakened in many ways, and there is no clear method to decide which restrictions produce reasonable agents and which ones make them too strong/weak. Second, and more important, these approaches do not look at the heart of the matter: they still describe the agent's information at a single (probably final) stage, without looking at how such state is reached. In other words, what is needed is not a representation of ideal or non-ideal agents: what is needed is a representation *of the actions* that allow an agent to change her information.

Recent works have combined a distinction between the agent's implicit and explicit information with a representation of the actions that change them. In a *propositional dynamic logic* style (*PDL*; Harel et al. (2000)), some of them have explored how the act of inference modifies an agent's explicit *knowledge* (Duc 2001, Jago 2006, Ågotnes and Alechina 2007). In a *dynamic epistemic logic* style (*DEL*; van Ditmarsch et al. (2007)), some others have explored how the acts of observation, inference, consideration and forgetting affect implicit and explicit *knowledge* (van Benthem 2008, Grossi and Velázquez-Quesada 2012, van Benthem and Velázquez-Quesada 2010, van Ditmarsch et al. 2009). But in our daily life we usually work with incomplete information, and therefore very few things are completely certain for us. Most of our behaviour is leaded not by what we *know*, but rather by what we *believe*.

The present work studies the notions of implicit and explicit *beliefs* and their

dynamics. We recall frameworks for representing non-ideal agents and beliefs in an *EL* style (Section 2) and then we combine them in a setting for representing *implicit* and *explicit beliefs* (Section 3). Then we look into the dynamics of these notions; first, by adapting the existing notion of *belief revision* in *DEL* to put it in harmony with our non-omniscient approach (Section 4), and then by introducing a new action our non-omniscient agents can perform: inferences involving not only knowledge but also *beliefs* (Section 5). We close with a list of further interesting questions that deserve additional investigation (Section 6). The extended version of this work (including proofs of Propositions) can be found in Velázquez-Quesada (2011).

2 Preliminaries

This section recalls two frameworks: one for representing implicit and explicit information and another for representing beliefs. By combining them, we will get our model for representing implicit and explicit beliefs. But before going into their details, we recall the framework on which all the others are based.

Epistemic logic The frameworks of the present work are based on that of *EL*. Given a set of atomic propositions P, the *EL* language extends the propositional one with formulas of the form $\Box\varphi$, read as *"the agent is informed about φ"*. The classical semantic model for *EL*-formulas is a *possible worlds model*, a tuple $M = \langle W, R, V\rangle$ with W a non-empty set of *possible worlds*, $R \subseteq (W \times W)$ an accessibility relation and $V : W \to \wp(P)$ an atomic valuation function.

Formulas are evaluated on *pointed models* (M, w) with M a possible worlds model and $w \in W$ (the evaluation point). Boolean connectives are interpreted as usual, and the key clause is the one for $\Box\varphi$: *the agent is informed about φ at w iff φ is true in all the worlds she considers possible from w*:

$$(M, w) \Vdash \Box\varphi \qquad \text{iff} \qquad \text{for all } u \in W, Rwu \text{ implies } (M, u) \Vdash \varphi$$

2.1 Modelling implicit and explicit information

The formula $\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$ is valid in Kripke models: the agent's information is closed under logical consequence. It has been argued that this

is not reasonable for 'real' agents since even computational ones, lacking the needed resources (space and/or time) to derive all the logical consequences of their information, may not have this property (Ågotnes and Alechina 2009).

One of the most influential solutions to this *omniscience problem* is *awareness logic* (Fagin and Halpern 1988), which follows the idea of distinguishing between *implicit* and *explicit* information. This approach makes crucial use of the concept of *awareness*: φ is explicit information iff it is implicit information and the agent is aware of it.

Syntactically, awareness logic extends the *EL* language with formulas of the form $\mathrm{A}\,\varphi$, read as *"the agent is aware of φ"*. Semantically, it extends possible worlds models with an *awareness function* A that indicates the set of formulas the agent is aware of at each possible world. The new formulas are evaluated in the following way:

$$(M, w) \Vdash \mathrm{A}\,\varphi \qquad \text{iff} \qquad \varphi \in \mathsf{A}(w)$$

Implicit information about φ is given by $\Box\varphi$, but explicit information of φ corresponds to $\Box\varphi \wedge \mathrm{A}\,\varphi$. Implicit information is still closed under logical consequence, but explicit information is not: the A-sets do not *need* to have any closure property ($\{\varphi \to \psi, \varphi\} \subseteq \mathsf{A}(w)$ does not imply $\psi \in \mathsf{A}(w)$).

This idea has produced models for representing implicit and explicit *knowledge* and their dynamics (e.g., Duc (1995), Jago (2006), van Benthem (2008), Velázquez-Quesada (2009a), van Ditmarsch et al. (2009), Grossi and Velázquez-Quesada (2012), van Benthem and Velázquez-Quesada (2010)). But the notion of *belief* is different, as we discuss below.

2.2 Plausibility models

Beliefs are different from knowledge. We do not believe something because it is true in all possible situations; we believe it because it is true in the ones we consider most likely to be the case (Grove 1988, Segerberg 2001). This suggests that the worlds an agent considers possible should be given not just by a plain set (what we get when we use equivalence relations for representing knowledge in *EL*); there should be also a *plausibility order* among them, indicating which worlds the agent considers more likely to be the case. This idea has led to the

development of variants of possible worlds models (Board 2004, van Benthem 2007) similar to those used for conditional logics (Lewis 1973, Veltman 1985, Lamarre 1991, Boutilier 1994). The models we will use, a small modification of the *plausibility models* of Baltag and Smets (2008), are introduced below.

A plausibility model is a possible worlds model in which the accessibility relation is interpreted as a plausibility relation, indicating the plausibility order of the posssible worlds from the agent's point of view. Different conditions can be imposed in this plausibility relation; in this work we will ask for it to be, using the terminology of Baltag and Smets (2008) a locally well-preorder or, as we will call it, a *locally-well preorder*.

Definition 2.1 (Locally-well preorder)**.** Let R be a binary relation over some non-empty domain W.

- For every $w \in W$, denote its *comparability class* by $C_R(w)$, that is, $C_R(w) := \{u \in W \mid Rwu \text{ or } Ruw\}$.
- For every $U \subseteq W$, denote its set of R-maximal elements as $\mathrm{Max}_R(U)$, that is, $\mathrm{Max}_R(U) := \{v \in U \mid \text{for every } u \in U, Ruv\}$.

The relation R is said to be a *locally-well preorder* iff it is a preorder such that, for each comparability class $C_R(w)$ and every non-empty $U \subseteq C_R(w)$, the set of R-maximal elements of U is non-empty, that is, $\mathrm{Max}_R(U) \neq \emptyset$. [1]

A locally-well preorder over a non-empty W partitions it into one or more comparability classes, each one of them being a connected preorder that has maximal elements (thus creating, inside each comparability class, one or more layers of equally-plausible elements with the layers themselves ordered according to their plausibility). Also that, because of local connectedness, the notion of maximal is global inside each comparability class: the maximal elements in each comparability class are the same from the perspective of any element belonging to it. In fact, a locally-well preorder is nothing but a locally connected[2] and conversely well-founded[3] preorder. This equivalence will be used when

[1] A *well preorder* is then a preorder in which there are maximal elements in every subset *of the whole domain* instead of just every subset of every comparability class.

[2] A binary relation R over W is *locally connected* iff, for every comparability class $C_R(w)$, every two elements $w_1, w_2 \in C_R(w)$ are R-comparable, that is, Rw_1w_2 or Rw_2w_1 or both.

[3] A binary relation R over W is *conversely well-founded* if there is no infinite $R^{|}$-ascending chain of elements in W, where $R^{|}$, the *strict* version of R, is given by $R^{|}wu$ iff Rwu and not Ruw.

arguing for certain properties, and for proving that certain model operations preserve locally-well preorders.

Now we can define formally what a plausibility model is.

Definition 2.2 (Plausibility model (Baltag and Smets 2008))**.** A *plausibility model* is a possible worlds model $M = \langle W, \leq, V \rangle$ in which the accessibility relation, denoted by $\leq$ and called the *plausibility relation*, is a locally-well preorder over W. If we have $w \leq u$ we will say that *"u is at least as plausible as w"*. We will also use the following abbreviations:

$$\begin{array}{rlll} u \text{ is more plausible than } w, & w < u & \text{iff} & w \leq u \text{ and } u \not\leq w \\ w \text{ is comparable to } u, & w \sim u & \text{iff} & w \leq u \text{ or } u \leq w \\ w \text{ is equally-plausible to } u, & w \simeq u & \text{iff} & w \leq u \text{ and } u \leq w \end{array}$$

For the language we use two modalities, one for the plausibility relation $\leq$, and another for the comparability relation $\sim$. Their interpretation is standard:

$$\begin{array}{ll} (M, w) \Vdash \langle \leq \rangle \varphi & \text{iff there is a } u \in W \text{ such that } w \leq u \text{ and } (M, u) \Vdash \varphi \\ (M, w) \Vdash \langle \sim \rangle \varphi & \text{iff there is a } u \in W \text{ such that } w \sim u \text{ and } (M, u) \Vdash \varphi \end{array}$$

Observe that comparability is actually epistemical indistinguishability: if a world is at least as plausible as another then, given one of them, the agent cannot rule out the other. Hence, we can use the relation $\sim$, an *equivalence* relation, to define the notion of knowledge: $[\sim]\varphi$ is read as *"the agent knows φ"*.

For the notion of belief, recall that each comparability class is in fact a connected preorder that has maximal elements. Then, as observed in Stalnaker (2006) and Baltag and Smets (2008), we can express that a given formula is true in the $\leq$-maximal (i.e., the most plausible) worlds of a given comparability class in the following way.

Fact 1. *Let (M, w) be a pointed plausibility model with $M = \langle W, \leq, V \rangle$. The formula φ is true in the $\leq$-maximal worlds from w iff w has a $\leq$-successor from which all $\leq$-successors satisfy φ. In symbols,*

$$(M, u) \Vdash \varphi \quad \textit{for every } u \in \mathrm{Max}_{\leq}(C_R(w)) \qquad \textit{iff} \qquad (M, w) \Vdash \langle \leq \rangle [\leq] \varphi$$

3 The plausibility acknowledgement framework

Our framework for representing implicit and explicit beliefs combines the described framework for representing implicit and explicit information with the described plausibility models for representing beliefs.

Formulas of our language are given by a propositional language extended, first, with formulas of the form $\mathrm{A}\,\varphi$ where φ is a formula, and second, with modalities $\langle\leq\rangle$ and $\langle\sim\rangle$. The formal definition is as follows.

Definition 3.1 (Language $\mathcal{L}$)**.** Given a set of atomic propositions P, formulas φ, ψ of the *plausibility acknowledgement* (*PA*) language $\mathcal{L}$ are given by

$$\varphi ::= p \mid \mathrm{A}\,\varphi \mid \neg\varphi \mid \varphi \vee \psi \mid \langle\leq\rangle\,\varphi \mid \langle\sim\rangle\,\varphi$$

where $p \in P$. Formulas of the form $\mathrm{A}\,\varphi$ are read as *"the agent has acknowledged φ as true"*. For the modalities, $\langle\leq\rangle\,\varphi$ is read as *"there is an at least as plausible world where φ holds"*, and $\langle\sim\rangle\,\varphi$ as *"there is an epistemically indistinguishable world where φ holds"*. Other boolean connectives as well as the universal modalities $[\leq]$ and $[\sim]$ are defined as usual ($[\leq]\,\varphi := \neg\langle\leq\rangle\,\neg\varphi$ and $[\sim]\,\varphi := \neg\langle\sim\rangle\,\neg\varphi$ for the latter).

Here is important to emphasize the interpretation of formulas of the form $\mathrm{A}\,\varphi$. Different from the awareness logic, they are not interpreted as *"the agent is aware of φ"* (i.e., conscious, but without any inclination about it); rather, they are interpreted as *"the agent has acknowledged φ as true"* (i.e., conscious *and* positive about it). In fact, the *awareness of* notion plays no role in this framework, as any agent represented with it is aware of every involved atomic propositions and, in general, aware of every formula of the language. An agent in this framework is not omniscient not because of lack of awareness, but because she does not need to acknowledge as true all the (infinite number of) formulas that are so in each possible world. The difference is important because it will justify our choice for the definitions of explicit knowledge/beliefs (Section 3.1).

Note how, as a consequence of this interpretation, an agent within this framework has in general uncertainty not only about which one is the real world, but also about what holds in each one of them. Even if she considers possible a single world where some φ holds, she may not recognize it as a φ-world because she may not have acknowledged that it indeed satisfies φ. From her perspective, there are not only worlds she identifies as φ-ones and worlds she

identifies as $\neg\varphi$-worlds: there are also worlds that are uncertain for her with respect to φ.

For the semantic model, a *plausibility acknowledgement* model extends the plausibility model with a function indicating the formulas the agent has acknowledged as true in each world. Such function is technically identical to the *awareness function* of *awareness logic*, but their contents are interpreted not as what the agent is aware *of* (i.e., what she entertains) but rather what she is aware *that* (i.e., what she acknowledges as true).

Definition 3.2 (Plausibility acknowledgement model)**.** Let P be a set of atomic propositions. A *plausibility acknowledgement (*PA*) model* is a tuple $M = \langle W, \leq, V, \mathsf{A}\rangle$ where $\langle W, \leq, V\rangle$ is a plausibility model over P (Definition 2.2) and

- $\mathsf{A} : W \to \wp(\mathcal{L})$ is the *acknowledgement set function*, indicating the formulas the agent has acknowledged as true at each possible world.

The *epistemic indistinguishability* relation, $\sim$, is defined as before: $w \sim u$ iff $w \leq u$ or $u \leq w$ (or, equivalently, $w \sim u$ iff $u \in C_R(w)$). A *pointed* PA *model* (M, w) is a *PA* model with a distinguished world $w \in W$.

For the semantic interpretation, formulas of the form $\mathrm{A}\,\varphi$ are interpreted with the acknowledgement set function, and the modalities $\langle\leq\rangle$ and $\langle\sim\rangle$ are interpreted via their corresponding relation.

Definition 3.3 (Semantic interpretation)**.** Let (M, w) be a pointed *PA* model with $M = \langle W, \leq, V, \mathsf{A}\rangle$.

$(M, w) \Vdash \mathrm{A}\,\varphi$	iff	$\varphi \in \mathsf{A}(w)$
$(M, w) \Vdash \langle\leq\rangle\varphi$	iff	there is a $u \in W$ such that $w \leq u$ and $(M, u) \Vdash \varphi$
$(M, w) \Vdash \langle\sim\rangle\varphi$	iff	there is a $u \in W$ such that $w \sim u$ and $(M, u) \Vdash \varphi$

For an axiom system we use Theorem 2.5 of Baltag and Smets (2008): the axiom system of Table 1 is sound and weakly complete for formulas of $\mathcal{L}$ with respect to *PA* models.

3.1 Implicit and explicit beliefs, and their basic properties

Now we will provide the definitions of implicit and explicit knowledge/beliefs.

Prop	$\vdash \varphi$ for φ a propositional tautology	*MP*	If $\vdash \varphi \to \psi$ and $\vdash \varphi$, then $\vdash \psi$
$K_{\leq}$	$\vdash [\leq](\varphi \to \psi) \to ([\leq]\varphi \to [\leq]\psi)$	$K_{\sim}$	$\vdash [\sim](\varphi \to \psi) \to ([\sim]\varphi \to [\sim]\psi)$
$Dual_{\leq}$	$\vdash \langle\leq\rangle\varphi \leftrightarrow \neg[\leq]\neg\varphi$	$Dual_{\sim}$	$\vdash \langle\sim\rangle\varphi \leftrightarrow \neg[\sim]\neg\varphi$
$Nec_{\leq}$	If $\vdash \varphi$, then $\vdash [\leq]\varphi$	$Nec_{\sim}$	If $\vdash \varphi$, then $\vdash [\sim]\varphi$
$T_{\leq}$	$\vdash [\leq]\varphi \to \varphi$	$T_{\sim}$	$\vdash [\sim]\varphi \to \varphi$
$4_{\leq}$	$\vdash [\leq]\varphi \to [\leq][\leq]\varphi$	$4_{\sim}$	$\vdash [\sim]\varphi \to [\sim][\sim]\varphi$
		$B_{\sim}$	$\vdash \varphi \to [\sim]\langle\sim\rangle\varphi$
LC	$(\langle\sim\rangle\varphi \wedge \langle\sim\rangle\psi) \to \big(\langle\sim\rangle(\varphi \wedge \langle\leq\rangle\psi) \vee \langle\sim\rangle(\psi \wedge \langle\leq\rangle\varphi)\big)$		
Inc	$\langle\leq\rangle\varphi \to \langle\sim\rangle\varphi$		

Table 1: Axiom system for $\mathcal{L}$ with respect to *PA* models.

Defining the 'implicit' notions is simple. Most works about the logical omniscient problem agree that what the classical epistemic logic framework gives us is actually the agent's *implicit* information, the best that she can do. Then, we will define *implicit knowledge* as what is true in all the worlds the agent considers epistemically possible, $[\sim]\varphi$, and *implicit beliefs* as what is true in the maximal worlds according to the agent's plausibility relation (i.e., the *most plausible* worlds), $\langle\leq\rangle[\leq]\varphi$.

For the 'explicit' notions there is no total consensus, even in similar frameworks. In the case of explicit knowledge, Duc (1995), Jago (2006), van Benthem (2008) and Velázquez-Quesada (2009a) define it directly as $\mathsf{A}\,\varphi$; some other approaches combine A with the modality for the epistemic relation, like the $[\sim]\varphi \wedge \mathsf{A}\,\varphi$ of Fagin and Halpern (1988), van Ditmarsch and French (2009) or the $[\sim]\mathsf{A}\,\varphi$ of Velázquez-Quesada (2009b).

In our case, we interpret the set $\mathsf{A}(w)$ as the formulas the agent has acknowledged as true in w. Each possible world provides an infinite amount of information (the infinite number of formulas that are true at it), so establishing a direct relation between what is true in a given world and what the agent identifies as true in it gives us an omniscient agent. But a more 'real' agent might fail to recognize as true in w a formula that indeed holds at it. The A function gives us the information the agent has acknowledged at each world, so we can define explicit information as what is true *and the agent has acknowledged as true*

in such set of worlds (cf. van Benthem and Velázquez-Quesada (2010), Grossi and Velázquez-Quesada (2012)).

Definition 3.4 (Implicit and explicit knowledge/beliefs)**.** The notions of *implicit* and *explicit knowledge/beliefs* are defined in Table 2.

Implicit knowledge: $K_{\mathrm{Im}}\varphi := [\sim]\varphi$	*Implicit* belief: $B_{\mathrm{Im}}\varphi := \langle\leq\rangle[\leq]\varphi$
Explicit knowledge: $K_{\mathrm{Ex}}\varphi := [\sim](\varphi \wedge \mathrm{A}\varphi)$	*Explicit* belief: $B_{\mathrm{Ex}}\varphi := \langle\leq\rangle[\leq](\varphi \wedge \mathrm{A}\varphi)$

Table 2: Implicit and explicit knowledge/beliefs.

The agent knows φ *implicitly* iff φ is true in all the epistemically indistinguishable worlds, and she knows φ *explicitly* if, in addition, she acknowledges it as true in all these worlds. Similarly, the agent believes φ *implicitly* iff φ is true in the most plausible worlds, and she believes φ *explicitly* if, in addition, she acknowledges it as true in these 'best' worlds.

We can also define the duals of these notions. The agent considers φ possible *implicitly*, $\widehat{K}_{\mathrm{Im}}\varphi$, iff there is an epistemically possible φ-world, $\langle\sim\rangle\varphi$; she considers φ possible *explicitly*, $\widehat{K}_{\mathrm{Ex}}\varphi$, iff there is an epistemically possible φ-world that she has identified, $\langle\sim\rangle(\varphi \wedge \mathrm{A}\varphi)$. For beliefs, she considers a φ-situation very likely *implicitly*, $\widehat{B}_{\mathrm{Im}}\varphi$, iff among the most plausible worlds there is a φ-one, $[\leq]\langle\leq\rangle\varphi$; she considers a φ-situation very likely *explicitly*, $\widehat{B}_{\mathrm{Ex}}\varphi$, iff among the most plausible worlds there is a φ-one that she has identified, $[\leq]\langle\leq\rangle(\varphi \wedge \mathrm{A}\varphi)$.

Implicit and explicit knowledge imply implicit and explicit beliefs.

Proposition 1. *$K_{\mathrm{Im}}\varphi \to B_{\mathrm{Im}}\varphi$ and $K_{\mathrm{Ex}}\varphi \to B_{\mathrm{Ex}}\varphi$ are valid in* PA *models.*

Properties of implicit and explicit knowledge under similar definitions have been already studied (Grossi and Velázquez-Quesada 2012). Here we will present some properties of the notions of implicit and explicit belief.

The notions are global Note how the notions of implicit and explicit beliefs are *global* in each comparability class.

Proposition 2. *$B_{\mathrm{Im}}\varphi \to [\sim]B_{\mathrm{Im}}\varphi$ and $B_{\mathrm{Ex}}\varphi \to [\sim]B_{\mathrm{Ex}}\varphi$ are valid* PA *models.*

Basic properties Explicit beliefs are obviously implicit beliefs.

Proposition 3. $B_{\mathrm{Ex}}\varphi \to B_{\mathrm{Im}}\varphi$ *is valid in* PA *models.*

Neither implicit nor explicit beliefs have to be true (e.g., $B_{\mathrm{Ex}}p \wedge \neg p$ is satisfiable in *PA* models) because the real world does not need to be among the most plausible ones. Nevertheless, implicit (and hence explicit) beliefs are consistent because any comparability class has always maximal elements.

Proposition 4. $\neg B_{\mathrm{Im}}\bot$ *is valid in* PA *models.*

Omniscience Implicit beliefs are omniscient.

Proposition 5. $B_{\mathrm{Im}}(\varphi \to \psi) \to (B_{\mathrm{Im}}\varphi \to B_{\mathrm{Im}}\psi)$ *is valid, and if* φ *is valid so is* $B_{\mathrm{Im}}\varphi$.

But explicit beliefs do not need to have these properties because the A-sets do not need to have any closure property.

Introspection Now let us review the introspection properties. First, implicit beliefs are positively and negatively introspective.

Proposition 6. *In* PA *models, implicit beliefs have the* positive *and the* negative *introspection property, that is,* $B_{\mathrm{Im}}\varphi \to B_{\mathrm{Im}}B_{\mathrm{Im}}\varphi$ *and* $\neg B_{\mathrm{Im}}\varphi \to B_{\mathrm{Im}}\neg B_{\mathrm{Im}}\varphi$ *are valid.*

Explicit beliefs do not have these properties in the general case, again because the A-sets do not need to have any closure property. Nevertheless, we can get introspection by asking for additional requirements.

For positive introspection, we need that if the agent has acknowledged that φ is true, then she has also acknowledged that she believes explicitly in it.

Proposition 7. *In* PA *models in which* $\mathrm{A}\,\varphi \to \mathrm{A}\,B_{\mathrm{Ex}}\varphi$ *is valid, explicit beliefs have the positive introspection property, that is,* $B_{\mathrm{Ex}}\varphi \to B_{\mathrm{Ex}}B_{\mathrm{Ex}}\varphi$ *is valid.*

For negative introspection, suppose the agent does not believe explicitly a given φ. This may be because φ is not even implicitly believed, but also because while φ is implicitly believed the agent has not acknowledged φ in the most plausible worlds. In the second case, φ implicitly but not explicitly believed, the agent is negatively introspective about this lack of explicit belief if she acknowledges $\neg B_{\mathrm{Ex}}\varphi$ in all the best worlds every time she does not acknowledge φ in all of them.

Proposition 8. *In* PA *models in which* $\neg B_{\mathrm{Im}} \mathrm{A}\, \varphi \rightarrow B_{\mathrm{Im}} \mathrm{A}\, \neg B_{\mathrm{Ex}} \varphi$ *is valid, the formula* $(\neg B_{\mathrm{Ex}} \varphi \wedge B_{\mathrm{Im}} \varphi) \rightarrow B_{\mathrm{Ex}} \neg B_{\mathrm{Ex}} \varphi$ *is also valid.*

Though in the general case explicit beliefs do not have neither positive nor negative introspection, they do have them in a weak form.

Proposition 9. *The formulas* $B_{\mathrm{Ex}} \varphi \rightarrow B_{\mathrm{Im}} B_{\mathrm{Ex}} \varphi$ *and* $\neg B_{\mathrm{Ex}} \varphi \rightarrow B_{\mathrm{Im}} \neg B_{\mathrm{Ex}} \varphi$ *are valid.*

After defining a framework for representing implicit and explicit forms of beliefs, we will now turn our attention to processes that transform them.

4 Belief revision

We start with *belief revision*, the act of changing beliefs in order to incorporate new external information in a consistent way (Gärdenfors 1992, Gärdenfors and Rott 1995, Williams and Rott 2001, Rott 2001). The study of this process and its properties can be traced back to the early 1980s, with *AGM* theory of Alchourrón et al. (1985) considered to mark the birth of the field.

Traditionally, there have been two approaches to study belief revision. The *postulational* approach analyses belief change without committing to any fixed mechanism, proposing instead abstract general principles that a "rational" belief revision process should satisfy. Most of the initial work on the field follows this approach, with the *AGM* theory being the most representative one.

Other works have approached belief revision in a more constructive way, presenting concrete mechanisms that change an agent's beliefs. An early 'syntactically flavoured' example is the epistemic entrenchment functions of Gärdenfors and Makinson (1988), based on an ordering among formulas. On the other side there are the approaches that represent beliefs in a different way, like Grove (1988) which uses a structure called *a system of spheres* (based on the earlier work of Lewis (1973)) to construct revision functions. Like an epistemic entrenchment, a system of spheres is essentially a preorder, but now the ordered objects are no longer formulas, but complete theories.

On its most basic form, belief revision involves an agent with her beliefs, and study the way these beliefs change when new information appears. Then, it is very natural to look for a belief revision approach within the *DEL* framework. Here we briefly review the main idea behind the most relevant proposals.

4.1 The *DEL* approach

The main idea behind plausibility models is that the set of worlds the agent considers possible has an order indicating the plausibility of each world; then, an agent believes what is true in the most plausible worlds (those she considers more likely to be the case). If beliefs are represented by a plausibility order, then changes in beliefs can be represented by changes in this order (van Ditmarsch 2005, van Benthem 2007, Baltag and Smets 2008). In particular, the act of *revising* beliefs in order to accept a given χ can be seen as an operation that puts χ-worlds at the top of the order. Of course, such a new order can be defined in several ways, but each one of them can be seen as a different policy for revising beliefs. The definition given below is just one of the many possibilities.

Definition 4.1 (Upgrade operation)**.** Let $M = \langle W, \leq, V, \mathsf{A}\rangle$ be a *PA* model and let χ be a formula in $\mathcal{L}$. The *upgrade* operation produces the *PA* model $M_{\chi\Uparrow} = \langle W, \leq', V, \mathsf{A}\rangle$, differing from M just in the plausibility order, given now by

$$\leq' := \underbrace{(\leq; \chi?)}_{(1)} \cup \underbrace{(\neg\chi?; \leq)}_{(2)} \cup \underbrace{(\neg\chi?; \sim; \chi?)}_{(3)}$$

The new plausibility relation is given in a *PDL* style. It states that, after an upgrade with χ, *"all χ-worlds become more plausible than all $\neg\chi$-worlds, and within the two zones, the old ordering remains"* (van Benthem 2007). More precisely, in $M_{\chi\Uparrow}$ we will have $w \leq' u$ iff in M (1) $w \leq u$ and u is a χ-world, or (2) w is a $\neg\chi$-world and $w \leq u$, or (3) $w \sim u$, w is a $\neg\chi$-world and u is a χ-world.

There are two important observations to make here. First, as said before, there are many definitions for a new plausibility relation that put χ-worlds at the top (see, e.g., (van Benthem 2007, van Eijck and Wang 2008)). But not all relation-changing operations are technically adequate. We are interested in those that preserve the required model properties, and therefore keep us in the relevant class of models. In other words, we are interested in operations that preserve locally-well preorders.

Proposition 10. *If M is a* PA *model, so is $M_{\chi\Uparrow}$.*

The second observation is related to the worlds that should be lifted by this operation. The idea behind an upgrade with χ is that χ will become more plausible than $\neg\chi$ *for the agent*; the lifted worlds should be those *the agent recognizes*

as χ-worlds. In classical *DEL* this boils down to lifting worlds satisfying χ, but in our approach our agents may not have direct access to all the information each possible world provides. Hence, for our agent, an upgrade with χ should not move to the top worlds satisfying χ, but rather those she identifies as satisfying χ, that is, those satisfying $\chi \wedge \mathrm{A}\,\chi$.

Now for the operation's effect. It puts on top worlds that satisfy $\chi \wedge \mathrm{A}\,\chi$ *in the original model* M (if there are none, the plausibility order will stay the same), but these worlds do not need to satisfy $\chi \wedge \mathrm{A}\,\chi$ *in the resulting model* $M_{(\chi \wedge \mathrm{A}\,\chi)\Uparrow}$. Then, an upgrade with χ does not necessarily make the agent believe in χ, even *implicitly*. This is a phenomena inherited from the 'omniscient' plausibility models, analogous to the well-known Moore-like sentences (*"p is the case and the agent does not know it"*) in *public announcement logic* (Plaza 1989, Gerbrandy 1999) that become false after being announced, and therefore cannot be known (Holliday and Icard 2010). For a simple example in the omniscient setting, assume that an upgrade with χ lifts worlds satisfying χ, and consider a *PA* model M with two worlds, w_1 in which an atom p is true, and the strictly more plausible w_2 in which p fails. Suppose we perform an upgrade with "*p is the case but the agent does not believe it (implicitly)*", $p \wedge \neg B_{\mathrm{Im}} p$. Since w_1 satisfies the formula but w_2 does not, after the upgrade w_1 will be strictly more plausible than w_2. But in the new model the agent does not believe (implicitly) the upgraded statement $p \wedge \neg B_{\mathrm{Im}} p$; it fails at the unique most plausible world w_1 because the second conjunct is false: now the agent *does believe* (implicitly) p!

Consider now the cases in which the upgraded χ is just a propositional formula. An upgrade with it moves to the top of the ordering those worlds satisfying χ in M. Since χ is propositional, every world satisfying it in M will also satisfy it in $M_{\chi\Uparrow}$, so after an upgrade with χ the agent will believe χ. This also extends to our non-omniscient setting. An upgrade with χ moves to the top the worlds satisfying $\chi \wedge \mathrm{A}\,\chi$ in M. Because χ is propositional and the operation does not affect A-sets, these worlds will still satisfy $\chi \wedge \mathrm{A}\,\chi$ in the new model. Hence, after an upgrade with χ, our non-omniscient agent will believe χ not only implicitly but also *explicitly*.

For the language we add the existential modality $\langle \chi \Uparrow \rangle$ (its universal version $[\chi \Uparrow]$ defined in the standard way). Formulas of the form $\langle \chi \Uparrow \rangle \varphi$ are read as *"it is possible for the agent to upgrade her beliefs with χ in such a way that after doing it φ is the case"*; their semantic interpretation is as follows.

Definition 4.2 (Semantic interpretation)**.** Let $M = \langle W, \leq, V, \mathsf{A} \rangle$ be a *PA* model

$\vdash \langle \chi \Uparrow \rangle p \leftrightarrow p$	$\vdash \langle \chi \Uparrow \rangle \mathrm{A}\, \varphi \leftrightarrow \mathrm{A}\, \varphi$
$\vdash \langle \chi \Uparrow \rangle \neg \varphi \leftrightarrow \neg \langle \chi \Uparrow \rangle \varphi$	$\vdash \langle \chi \Uparrow \rangle (\varphi \vee \psi) \leftrightarrow \langle \chi \Uparrow \rangle \varphi \vee \langle \chi \Uparrow \rangle \psi$
$\vdash \langle \chi \Uparrow \rangle \langle \trianglelefteq \rangle \varphi \leftrightarrow \langle \trianglelefteq \rangle \big(\chi \wedge \langle \chi \Uparrow \rangle \varphi\big) \vee \big(\neg \chi \wedge \langle \trianglelefteq \rangle \langle \chi \Uparrow \rangle \varphi\big) \vee \big(\neg \chi \wedge \langle \sim \rangle (\chi \wedge \langle \chi \Uparrow \rangle \varphi)\big)$	
$\vdash \langle \chi \Uparrow \rangle \langle \sim \rangle \varphi \leftrightarrow \langle \sim \rangle \langle \chi \Uparrow \rangle \varphi$	If $\vdash \varphi$, then $\vdash [\chi \Uparrow] \varphi$

Table 3: Axioms and rule for the upgrade modality.

and χ a formula in $\mathcal{L}$. Then,

$$(M, w) \Vdash \langle \chi \Uparrow \rangle \varphi \qquad \text{iff} \qquad (M_{\chi \Uparrow}, w) \Vdash \varphi$$

The upgrade operation is a total function: it can always be executed (there is no precondition) and it always yields one and only one model. It can be also argued that for the agent to upgrade her beliefs with χ she needs to consider χ possible. For this, we just need to indicate this requirement as a precondition for the operation. For an omniscient agent, this is given by the formula $\langle \sim \rangle \chi$; in our non-omniscient case, this is given by $\langle \sim \rangle (\chi \wedge \mathrm{A}\, \chi)$.

For an axiom system for the language with the new modality, we present *reduction axioms*: valid formulas that indicate how to translate a formula with the new modality $\langle \chi \Uparrow \rangle$ into a provably equivalent one without them. Soundness follows from the validity of the these new axioms; completeness follows from the completeness of the basic system. We refer to Chapter 7 of van Ditmarsch et al. (2007) for an extensive explanation of this technique.

Theorem 1 (Reduction axioms for the upgrade modality)**.** *The valid formulas of the language $\mathcal{L}$ plus the upgrade modality in* PA *models are exactly those provable by the axioms and rules for the static base language (Table 1) plus the reduction axioms and modal inference rules listed in Table 3.*

Atomic valuation and acknowledgement sets are not affected by an upgrade, and the reduction axioms for p and $\mathrm{A}\, \varphi$ reflect this. Reduction axioms for $\neg$ and $\vee$ are standard, and that for the indistinguishability modality $\langle \sim \rangle$ indicates that the operation just changes the order *within* each comparability class: after an upgrade there will be a comparable φ-world iff before the upgrade there is a comparable world that will satisfy φ after the upgrade.

The interesting axiom is the one for the plausibility modality $\langle\leq\rangle$. It simply translates the three-cases *PDL* definition of the new plausibility relation: after an upgrade with χ there will be a $\leq$-reachable world where φ holds iff before the operation (1) there is a $\leq$-reachable χ-world that will satisfy φ after the upgrade, or (2) the current world satisfies $\neg\chi$ and can $\leq$-reach a world that will satisfy φ after the operation, or (3) the current world satisfies $\neg\chi$ and can $\sim$-reach one that satisfies χ and will satisfy φ after the upgrade.

5 Belief-based inference

The act of *revision* has been borrowed from standard *DEL*. But our non-omniscient agent can perform more actions that change her information; in particular, she can perform *inference*. Before going into inferences that involve *beliefs*, let us review the main ideas behind inferences that involve only *knowledge*.

The intuition behind an action of *knowledge-based inference* is that, if the agent *knows explicitly* an implication and its antecedent, then a *modus ponens* step will make her *know explicitly* the implication's consequent. In our setting, this action can be semantically defined as an operation adds the implication's consequent to the A-set of those worlds in which the agent *knows explicitly the implication and its antecedent* (Grossi and Velázquez-Quesada 2012).

But take a closer look at the operation. What it actually does is discard those worlds in which the agent knows explicitly the implication and its antecedent, and replace them with copies that are almost identical, the only difference being that their A-sets now contain the consequent of the applied implication. And this is reasonable because, under the assumption that knowledge is true information, knowledge-based inference is simply *deduction*: the antecedent is true and the implication preserves truth, so the implication's consequent *must* be true. Moreover, situations where the implication and its antecedent are true but the consequent is not *are not possible*.

The case is different when the inference involve beliefs. For example, if the antecedent of an implication is explicitly known but the implication is only explicitly believed, then, though it is reasonable to consider very likely a situation in which the implication, its antecedent and its consequent hold, the agent should not discard a situation where the antecedent holds but the consequent (and hence the implication) fails. An operation representing such action

should *split* the current possibilities into two: one of them, the most plausible one, standing for the case in which the implication's consequent holds; the other, the less plausible one, standing for the case in which the implication's consequent (and hence the implication) fails.

More generally, an inference involving beliefs creates new possibilities, and an operation representing it should be faithful to this. Our proposal will be based in the plausibility version (Baltag and Smets 2008) of the *action models* and *product update* of Baltag et al. (1999).

5.1 Plausibility acknowledgement action models

The intuition behind the *action models* of Baltag et al. (1999) is that, just as the agent can be uncertain about which one is the real world, she can also be uncertain about which event has taken place. In such situations, her uncertainty about the event can be represented with a model similar to that used for representing her uncertainty about the situation. *Action models* are possible-worlds-like structures in which the agent considers different events as possible; then her uncertainty *after the action* is an even combination of her uncertainty about the situation *before the action* and her uncertainty *about the action*.

This idea has been extended in order to match richer structures that indicate not only possible worlds but also a plausibility order among them. A first approach was made in van Ditmarsch (2005) using orders based on plausibility ordinals for expressing degrees of belief. In contrast, the *plausibility action models* of Baltag and Smets (2008) are purely *qualitative*, and therefore provide a more natural extension to be used with the matching plausibility models.

In this section we will extend these plausibility action models in order to deal with our acknowledgement set function. Here is the formal definition.

Definition 5.1 (Plausibility acknowledgement action model)**.** A *plausibility acknowledgement (*PA*) action model* is a tuple $O = \langle E, \leqslant, \mathsf{Pre}, \mathsf{Pos}_{\mathsf{A}} \rangle$ where

- $\langle E, \leqslant, \mathsf{Pre} \rangle$ is a plausibility action model (Baltag and Smets 2008) with E a finite non-empty set of *events*, $\leqslant$ a *plausibility order* on E (with the same requirements as those for a plausibility order in *PA* models) and $\mathsf{Pre} : E \to \mathcal{L}$ a *precondition* function indicating the requirement each event should satisfy in order to take place. This requirement is given in terms of a formula in our language $\mathcal{L}$.

- $\mathsf{Pos_A} : (E \times \wp(\mathcal{L})) \to \wp(\mathcal{L})$ is the *new acknowledgement set* function, indicating the set of formulas the agent will acknowledge after the action, based on what she acknowledged before it and the event that has taken place.

Again, we define three new relations: *strict plausibility*, $\prec := \preccurlyeq \cap \overline{\succcurlyeq}$, *epistemic indistinguishability* (i.e., comparability), $\approx := \preccurlyeq \cup \succcurlyeq$, and *equal plausibility*, $\cong := \preccurlyeq \cap \succcurlyeq$. A pointed *PA* action model (O, e) has a distinguished event $e \in E$.

The effect of a *PA* action model over a *PA* model is given by the product update. Both the static and the action model are preorders with further properties, so there are two natural ways of building a preorder of their combination: to give priority to either the preorder of the static model, or else to that of the action model. The second option is closer to the intended spirit in which it is the *action* the one that will modify the agent's static plausibility order. The formal definition of this case is as follows.

Definition 5.2 (Product update)**.** Let $M = \langle W, \leq, V, \mathsf{A}\rangle$ be a *PA* model and $O = \langle E, \preccurlyeq, \mathsf{Pre}, \mathsf{Pos_A}\rangle$ be a *PA* action model. The *product update* operation $\otimes$ yields the *PA* model $M \otimes O = \langle W', \leq', V', \mathsf{A}'\rangle$ given by

- $W' := \Big\{(w, e) \in (W \times E) \mid (M, w) \models \mathsf{Pre}(e)\Big\}$
- $(w_1, e_1) \leq' (w_2, e_2)$ iff $\underbrace{\Big(e_1 \prec e_2 \text{ and } w_1 \sim w_2\Big)}_{(1)}$ or $\underbrace{\Big(e_1 \cong e_2 \text{ and } w_1 \leq w_2\Big)}_{(2)}$

and, for every $(w, e) \in W'$,

$$\bullet\ V'(w, e) := V(w) \qquad \bullet\ \mathsf{A}'(w, e) := \mathsf{Pos_A}(e, \mathsf{A}(w))$$

The domain of the new *PA* model is the restricted cartesian product of W and E: a pair (w, e) is a world in the new model iff event e can be executed at world w. For the atomic valuation, a world in the new model inherits the atomic valuation of its static component: an atom p holds at (w, e) iff p holds at w. For our syntactic component, the agent's acknowledgement set at world (w, e) is given by the function $\mathsf{Pos_A}$ with the event e and her original acknowledgement set at w as parameters (the use of the old acknowledgement set allows us to deal with dependence of the new set with respect to the old one).

The plausibility order of the new model is built following the 'action-priority' rule: (w_2, e_2) is more plausible than (w_1, e_1) iff either (1) e_2 is *strictly* more plausible than e_1 and w_1, w_2 are already epistemically indistinguishable, or else (2) e_1, e_2 are *equally plausible* and w_2 is more plausible than w_1.

The product update operation preserves *PA* models.

Proposition 11. *If M is a* PA *model and O a* PA *action model, $M \otimes O$ is a* PA *model.*

Our *PA* action models can mimic pure plausibility action models (just define $\mathsf{Pos}_{\mathsf{A}}(e, X) := X$ for every event $e \in E$), but they can do more. Thanks to the new acknowledgement set function we can truly represent acts that change not only the situations the agent considers possible, but also what she has acknowledged as true in each one of them, as we will see in Subsection 5.2.

Syntactically, we extend our language with modalities for each pointed *PA* action model (O, e) in order to build formulas of the form $\langle O, e\rangle \varphi$. Their semantic interpretation is given below.

Definition 5.3 (Semantic interpretation)**.** Let (M, w) be a pointed *PA* model and let (O, e) be a pointed *PA* action model with Pre its precondition function.

$$(M, w) \Vdash \langle O, e\rangle \varphi \qquad \text{iff} \qquad (M, w) \Vdash \mathsf{Pre}(e) \text{ and } (M \otimes O, (w, e)) \Vdash \varphi$$

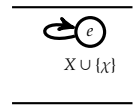
Now we will introduce inferences that can be portrayed with *PA* action models.

5.2 Examples of *PA* action models

Just like a public announcement corresponds to a single-event action model of Baltag et al. (1999), the action of knowledge-based inference corresponds to a single-event *PA* action model.

Definition 5.4 (Inference with known implication and known antecedent)**.** Let $\eta \to \chi$ be an implication. The action of *inference with known implication and known antecedent* is given by the *PA* action model $O^{\eta\to\chi}_{KK}$ whose definition (left) and diagram showing events, plausibility relation and the way acknowledgement sets are affected (right) appear below.

- $E := \{e\}$
- $\preccurlyeq := \{(e, e)\}$
- $\mathsf{Pre}(e) := K_{\mathrm{Ex}}(\eta \to \chi) \wedge K_{\mathrm{Ex}}\eta$
- $\mathsf{Pos}_{\mathsf{A}}(e, X) := X \cup \{\chi\}$



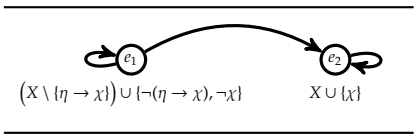
This action model has a single event, and its precondition is for the agent to know explicitly $\eta \to \chi$ and η. In the resulting model, the agent will acknowledge χ in all worlds satisfying the precondition. Note how, since both $\eta \to \chi$ and η are true in all epistemically indistinguishable worlds, χ must be true in all of them *in* M. But since the action only adds χ to the A-sets, only the truth-value of formulas containing $\mathrm{A}\,\chi$ is affected; hence χ itself cannot be affected and will still be true in all epistemically indistinguishable worlds in $M \otimes O_{KK}^{\eta \to \chi}$. Hence, the agent will know explicitly χ. We can look at this from the perspective of the truth-table of an implication: $\eta \to \chi$ and η are true so χ must be the case.

But our *PA* action models can represent more. Following our previous discussion, here is an action model for an inference in which the antecedent is known but the implication is just believed.

Definition 5.5 (Inference with believed implication and known antecedent)**.** Let $\eta \to \chi$ be an implication. The action of *inference with believed implication and known antecedent* is given by the *PA* action model $O_{BK}^{\eta \to \chi}$, defined as

- $E := \{e_1, e_2\}$ • $\preccurlyeq := \{(e_1, e_1), (e_1, e_2), (e_2, e_2)\}$ • $\mathsf{Pre}(e_i) := B_{\mathrm{Ex}}(\eta \to \chi) \wedge K_{\mathrm{Ex}}\eta$
- $\begin{cases} \mathsf{Pos}_{\mathsf{A}}(e_1, X) := \big(X \setminus \{\eta \to \chi\}\big) \cup \{\neg(\eta \to \chi), \neg\chi\} \\ \mathsf{Pos}_{\mathsf{A}}(e_2, X) := X \cup \{\chi\} \end{cases}$

The diagram below shows this two-event action model. The event on the right, the most plausible one, corresponds to the case in which the implication holds (and therefore so its consequent); hence the agent acknowledges χ. The one on the left, the less plausible one, corresponds to the case in which the implication fails (and therefore so its consequent, since the antecedent is true); hence the agent discards $\eta \to \chi$, acknowledging its negation and the negation of its consequent. In both events the precondition is the same: the agent should believe explicitly the implication and know explicitly its antecedent.



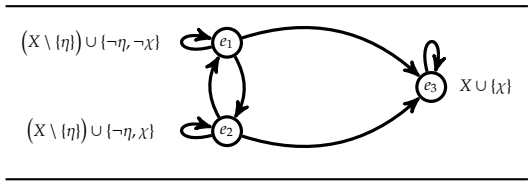
From the perspective of the truth-table of an implication, this case corresponds to the situations in which the antecedent η is true. There are two possibilities: either the implication $\eta \to \chi$ (and hence its consequent χ) are the case, or else the implication (and hence its consequent) fails.

We can also represent an analogous situation in which the implication is known but the antecedent is only believed. In the best scenario the believed antecedent is true, but the situations in which it fails should be also considered.

Definition 5.6 (Inference with known implication and believed antecedent)**.** Let $\eta \to \chi$ be an implication. The action of *inference with known implication and believed antecedent* is given by the *PA* action model $O^{\eta\to\chi}_{KB}$, defined as

- $E := \{e_1, e_2, e_3\}$
- $\mathsf{Pre}(e_i) := K_{\mathrm{Ex}}(\eta \to \chi) \wedge B_{\mathrm{Ex}}\eta$
- $\preccurlyeq := (E \times E) \setminus \{(e_3, e_1), (e_3, e_2)\}$
- $\begin{cases} \mathsf{Pos}_{\mathsf{A}}(e_1, X) := \big(X \setminus \{\eta\}\big) \cup \{\neg\eta, \neg\chi\} \\ \mathsf{Pos}_{\mathsf{A}}(e_2, X) := \big(X \setminus \{\eta\}\big) \cup \{\neg\eta, \chi\} \\ \mathsf{Pos}_{\mathsf{A}}(e_3, X) := X \cup \{\chi\} \end{cases}$

The diagram below shows this three-event action model. The most plausible event, e_3, corresponds the case in which the believed antecedent η is indeed the case, and hence the agent acknowledges the consequent χ. But since $\eta \to \chi$ holds (it is *known*), if η fails there are still two possibilities for χ: one in which it fails (event e_1), and another in which it does not (event e_2). Here we have defined these two situations as equally plausible, but other orderings are possible.



Again, from the perspective of a truth-table for implication, this corresponds to the three cases that are left when the implication is assumed as true.

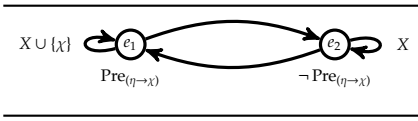
We can even represent a fourth scenario in which both the implication and the antecedent are just believed. In this case there is yet another possibility: the antecedent is indeed true, but the implication and the consequent fails.

The previous *PA* action models act 'globally', extending the agent's explicit information based on what she has *in a set of worlds* (the epistemically indistinguishable ones for knowledge; the most plausible ones for beliefs). But we can also represent 'local' inferences in which the agent extends what she acknowledges in some world based only on the information she has about it.

Definition 5.7 (Weak local inference)**.** Let $\eta \to \chi$ be an implication. Define the abbreviation $\mathrm{Pre}_{(\eta\to\chi)} := \mathrm{A}\,(\eta \to \chi) \wedge \mathrm{A}\,\eta$, stating that the agent has acknowledged it and its antecedent. The action of *weak local inference* is given by the following *PA* action model $O_W^{\eta\to\chi}$:

- $E := \{e_1, e_2\}$
- $\left\{ \begin{array}{l} \mathsf{Pre}(e_1) := \mathrm{Pre}_{(\eta\to\chi)} \\ \mathsf{Pre}(e_2) := \neg\,\mathrm{Pre}_{(\eta\to\chi)} \end{array} \right.$
- $\preccurlyeq := E \times E$
- $\left\{ \begin{array}{l} \mathsf{Pos}_{\mathsf{A}}(e_1, X) := X \cup \{\chi\} \\ \mathsf{Pos}_{\mathsf{A}}(e_2, X) := X \end{array} \right.$

With this action the agent works *locally*. Any given world satisfies either the precondition of e_1 or the precondition of e_2, but not both, so after the operation we will get a model that differs from the original static one only in that the agent will have acknowledged the implication's consequent exactly in those worlds in which she already acknowledged the implication and its antecedent. The diagram of this *PA* action model appears below.



A stronger form of local inference can be obtained by strengthening the precondition of e_1 into $\mathrm{Pre}_{(\eta\to\chi)} := (\eta \wedge \mathrm{A}\,\eta) \wedge ((\eta \to \chi) \wedge \mathrm{A}\,(\eta \to \chi))$, requiring now not only for the agent to acknowledge the implication and its antecedent, but also

for them to be true. If the precondition of e_2 is defined as $\neg \mathrm{Pre}_{(\eta\to\chi)}$, then after the operation the acknowledgement sets of worlds satisfying $\mathrm{Pre}_{(\eta\to\chi)}$will be extended with the implication's consequent, and those of the rest of the worlds will remain the same.

5.3 Completeness

Let us now turn to the syntactic characterization of validities involving the *PA* action model modalities. Just like in the upgrade case, we will provide *reduction axioms*, and here is our strategy. First, we will extend our static language to deal with what we will call *set expressions*, providing their semantic interpretation and their corresponding axioms. These expressions will allow us to look for formulas not only in A-sets, but also at more complex ones. Then, with their help we will provide reduction axioms for the class of *PA* action models in which the $\mathsf{Pos}_{\mathsf{A}}$ functions are definable by means of these expressions.

Definition 5.8 (Extended $\mathcal{L}$)**.** Given a set of atomic propositions P, formulas φ, ψ and *set expressions over formulas* Φ, Ψ of the *extended* PA *language* $\mathcal{L}$ are given, respectively, by

$$\varphi ::= p \mid \lfloor\Phi\rfloor\varphi \mid \neg\varphi \mid \varphi \vee \psi \mid \langle\sim\rangle\varphi \mid \langle\leq\rangle\varphi$$
$$\Phi ::= \mathrm{A} \mid \{\varphi\} \mid \overline{\Phi} \mid \Phi \cup \Psi$$

with p an atomic proposition in P.

Formulas of the form $\mathrm{A}\,\varphi$ have disappeared, leaving their place to formulas of the form $\lfloor\Phi\rfloor\varphi$ where Φ is what we call a *set expression*. While the $\mathrm{A}\,\varphi$ formulas allowed us to look only at the content of the A-sets, formulas of the form $\lfloor\Phi\rfloor\varphi$ allow us to look at the content of more complex sets Φ that are built from A and singletons $\{\varphi\}$ by means of complement and union. Even though our syntax for set expressions may suggest some strong semantic content, they are just a way of making syntactic comparisons between formulas, as it is fixed by their semantic interpretation.

Definition 5.9 (Semantic interpretation)**.** Let (M, w) be a pointed *PA* model with A its acknowledgement sets function. The semantic interpretation for the new formulas is given by

$(M, w) \Vdash \lfloor \mathrm{A} \rfloor \varphi \quad \text{iff} \quad \varphi \in \mathsf{A}(w)$ $\qquad$ $(M, w) \Vdash \lfloor \{\psi\} \rfloor \varphi \quad \text{iff} \quad \varphi \text{ is } \psi$

$(M, w) \Vdash \lfloor \overline{\Phi} \rfloor \varphi \quad \text{iff} \quad \varphi \in \overline{\Phi}$ $\qquad$ $(M, w) \Vdash \lfloor \Phi \cup \Psi \rfloor \varphi \quad \text{iff} \quad \varphi \in (\Phi \cup \Psi)$

So $\lfloor \mathrm{A} \rfloor \varphi$ is equivalent to the earlier $\mathrm{A}\,\varphi$. Also, we can even look at the contents of sets built with the intersection and difference operations following the standard definitions:

$$\Phi \cap \Psi := \overline{\overline{\Phi} \cup \overline{\Psi}} \qquad \Phi \setminus \Psi := \Phi \cap \overline{\Psi}$$

The earlier 'static' axiom system is not enough anymore: set expressions have special behaviour, characterized by the following extra axioms.

Theorem 2 (Extra axioms for extended $\mathcal{L}$ w.r.t. *PA* models)**.** *The axiom system of Table 1, together with the axioms from Table 4 is sound and (weakly) complete for the extended language $\mathcal{L}$ with respect to* PA *models.*

$SE^{\{\}}_{\mathsf{A}} \ \vdash \lfloor \{\psi\} \rfloor \psi$	$SE^{\{\}}_{\mathsf{A}} \ \vdash \neg \lfloor \{\psi\} \rfloor \varphi \quad \text{for } \varphi \neq \psi$
$SE^{-}_{\mathsf{A}} \ \vdash \lfloor \overline{\Phi} \rfloor \varphi \ \leftrightarrow \ \neg \lfloor \Phi \rfloor \varphi$	$SE^{\cup}_{\mathsf{A}} \ \vdash \lfloor \Phi \cup \Psi \rfloor \varphi \ \leftrightarrow \ \big(\lfloor \Phi \rfloor \varphi \vee \lfloor \Psi \rfloor \varphi \big)$

Table 4: Axiom system for extended $\mathcal{L}$ w.r.t. *PA* models.

The new axioms reflect the behaviour of these sets operations. Axioms $SE^{\{\}}_{\mathsf{A}}$ indicate that ψ and only ψ is an element of $\{\psi\}$. Axiom SE^{-}_{A} says that φ is in the complement of a set iff it is not in the set; axiom $SE^{\cup}_{\mathsf{A}}$ says that φ is in the union of two sets iff it is in at least one of them.

Moreover, the axioms for complement and union tell us that $\lfloor \overline{\Phi} \rfloor \varphi$ and $\lfloor \Phi \cup \Psi \rfloor \varphi$ are not really needed: they can be defined as $\neg \lfloor \Phi \rfloor \varphi$ and $\lfloor \Phi \rfloor \varphi \vee \lfloor \Psi \rfloor \varphi$, respectively. All we really need are expressions to verify syntactic identity between formulas, like formulas of the form $\lfloor \{\psi\} \rfloor \varphi$ do. With such extension, our original *PA* language $\mathcal{L}$ is enough for defining these new expressions. Nevertheless, we will keep this 'syntactic sugar' in order to simplify the reduction axioms that will be provided.

Let us define formally the class of *PA* action models whose new acknowledgement set in terms of set expressions.

Definition 5.10 (SE-definable *PA* action model)**.** A *set expression (*SE*) definable* PA *action model* is a *PA* action model in which, for each event e, the new acknowledgement set function $\mathsf{Pos}_{\mathsf{A}}(e)$ is given by a set expression over formulas. Note how all the *PA* action models presented in Subsection 5.2 are *SE* definable.

Now we can provide reduction axioms for the modalities that involve SE-definable *PA* action models.

Theorem 3. *The axiom system built from Tables 1, 4 and Table 5 (with* $\top$ *and* $\bot$ *the always true and always false formula, respectively) provide a sound and (weakly) complete axiom system for formulas in the extended language* $\mathcal{L}$ *plus modalities for action models with respect to* PA *models and* SE*-definable* PA *action models.*

$\vdash \langle O,e\rangle p \leftrightarrow \mathsf{Pre}(e) \wedge p$
$\vdash \langle O,e\rangle \neg\varphi \leftrightarrow \mathsf{Pre}(e) \wedge \neg\langle O,e\rangle\varphi$
$\vdash \langle O,e\rangle(\varphi \vee \psi) \leftrightarrow \big(\langle O,e\rangle\varphi \vee \langle O,e\rangle\psi\big)$
$\vdash \langle O,e\rangle\langle\leq\rangle\, \varphi \leftrightarrow \Big(\mathsf{Pre}(e) \wedge \big(\bigvee_{e\prec e'}\langle\sim\rangle\langle O,e'\rangle\varphi \vee \bigvee_{e\cong e''}\langle\leq\rangle\langle O,e''\rangle\varphi\big)\Big)$
$\vdash \langle O,e\rangle\langle\sim\rangle\, \varphi \leftrightarrow \Big(\mathsf{Pre}(e) \wedge \bigvee_{e\simeq e'}\langle\sim\rangle\langle O,e'\rangle\varphi\Big)$
If $\vdash \varphi$, then $\vdash [O,e]\varphi$
$\vdash \langle O,e\rangle\lfloor \mathrm{A}\rfloor\, \varphi \leftrightarrow \mathsf{Pre}(e) \wedge \lfloor \mathsf{Pos}_{\mathsf{A}}(e)\rfloor\, \varphi$
$\vdash \langle O,e\rangle\lfloor \{\psi\}\rfloor\, \psi \leftrightarrow \mathsf{Pre}(e) \wedge \top$
$\vdash \langle O,e\rangle\lfloor \{\psi\}\rfloor\, \varphi \leftrightarrow \mathsf{Pre}(e) \wedge \bot \quad$ for $\varphi \neq \psi$
$\vdash \langle O,e\rangle\lfloor \overline{\Psi}\rfloor\, \varphi \leftrightarrow \langle O,e\rangle\neg\lfloor \Psi\rfloor\, \varphi$
$\vdash \langle O,e\rangle\lfloor \Psi \cup \Phi\rfloor\, \varphi \leftrightarrow \langle O,e\rangle(\lfloor \Phi\rfloor\, \varphi \vee \lfloor \Psi\rfloor\, \varphi)$

Table 5: Axioms and rules for *SE*-definable action models.

On the first block, the first three axioms are standard: (O,e) does not affect atomic valuations, commute with negations (modulo the precondition) and distributes over disjunctions. The fourth, inherited from Baltag and Smets (2008), states that an (O,e) product update after which there is a more plausible φ-world can be performed iff the evaluation point satisfies e's precondition and in the original model there is an *epistemically indistinguishable* world that will satisfy φ after a product update with a *strictly more plausible* e', or there is a more plausible world that will satisfy φ after a product update with an *equally plausible*

e''. Finally, the fifth reduction axiom indicates that the comparability class does not change: an (O, e) product update after which there is an epistemically indistinguishable φ-world can be performed iff the evaluation point satisfies e's precondition and there is an epistemically indistinguishable world that will satisfy φ after a product update with an *indistinguishable* e'.

The second block contains the axioms for set expressions over formulas, with the first one being the key: after an (O, e) product update, φ will be in the agent's acknowledgement set iff e's precondition is satisfied an φ is in the set expression that defines the new acknowledgement set at event e:

$$\langle O, e\rangle\lfloor \mathsf{A}\rfloor \varphi \;\leftrightarrow\; \mathsf{Pre}(e) \wedge \lfloor \mathsf{Pos}_{\mathsf{A}}(e)\rfloor \varphi$$

The simplicity of the axiom takes advantage of the fact that our extended $\mathcal{L}$ language can deal with set expressions. As mentioned before, the original language $\mathcal{L}$ plus expressions for syntactic identity is powerful enough to express the membership of a given formula in a set defined from A-sets and singletons by means of complement and union. Then, reduction axioms without set expressions can be provided, but we would need an inductive translation from $\mathsf{Pos}_{\mathsf{A}}(e)$ to the formula that express the membership of φ in it. The remaining axioms for set expressions over formulas simply unfold the static axioms for the remaining set-expressions over formulas.

6 Conclusions and Further Work

We have provided a representation for implicit and explicit beliefs by combining ideas for representing non-omniscient agents with ideas for representing beliefs in a possible worlds setting. Then, we have reviewed the *DEL* approach for the act of *belief revision*, discussing what we need to take into account to implement it in our non-omniscient setting faithfully, and we have also defined *PA* action models, a powerful tool that allows us to represent different forms of inference that involve not only knowledge but also beliefs.

Like most research works, ours has provided some answers, but has also raised interesting questions. Here are the ones that we consider most appealing. (1) Long-term behaviour. We have defined the effect of a single execution of informational actions, but the result of their iterative application is also important. More precisely, fixed-point operators would allow us express the effect of iterative application of the defined actions, analogous to the Kleene star operator in

propositional dynamic logic. (2) Multi-agent notions of information. We have dealt with finer single-agent notions of information. But agents are usually not isolated, and in such settings, group notions of information, like group knowledge/beliefs and, more interestingly, common knowledge/beliefs, become important. In our fine-notions-of-information setting, this amounts to the study of implicit and explicit forms of group and common knowledge/beliefs.

Acknowledgements The author thanks the organizers, the anonymous referees and the audiences of the *Workshop on Theories of Information Dynamics and Interaction and their Application to Dialogue (TIDIAD'09)* and the *Third Workshop on Logics for Resource-Bounded Agents (LRBA-3)*; their comments and observations have greatly improved this paper. The author also thanks Johan van Benthem for the illuminating ideas that started this project, and Hans van Ditmarsch for pointing out some flaws in old versions and for the many suggestions that have helped to make this work better.

References

T. Ågotnes and N. Alechina. The dynamics of syntactic knowledge. *Journal of Logic and Computation*, 17(1):83–116, 2007.

T. Ågotnes and N. Alechina, editors. *Special issue on Logics for Resource Bounded Agents*, 2009. *Journal of Logic, Language and Information*, 18(1).

C. E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50(2):510–530, 1985.

A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. In G. Bonanno, W. van der Hoek, and M. Wooldridge, editors, *Logic and the Foundations of Game and Decision Theory (LOFT7)*, pages 13–60. Amsterdam University Press, 2008.

A. Baltag, L. Moss, and S. Solecki. The logic of public announcements, common knowledge and private suspicions. Technical Report SEN-R9922, CWI, Amsterdam, 1999.

O. Board. Dynamic interactive epistemology. *Games and Economic Behavior*, 49 (1):49–80, 2004.

C. Boutilier. Conditional logics of normality: A modal approach. *Artificial Intelligence*, 68(1):87–154, 1994.

J. Drapkin and D. Perlis. Step-logics: An alternative approach to limited reasoning. In *Proc. European Conf. on AI*, pages 160–163, 1986.

H. N. Duc. Logical omniscience vs. logical ignorance. on a dilemma of epistemic logic. In C. Pinto-Ferreira and N. Mamede, editors, *EPIA 1995*, volume 990 of *LNCS*, pages 237–248. Springer, 1995.

H. N. Duc. *Resource-Bounded Reasoning about Knowledge*. PhD thesis, Institut für Informatik, Universität Leipzig, Leipzig, Germany, 2001.

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

P. Gärdenfors, editor. *Belief Revision*. Cambridge University Press, 1992.

P. Gärdenfors and D. Makinson. Revisions of knowledge systems using epistemic entrenchment. In M. Vardi, editor, *TARK II*, pages 83–95. Morgan Kaufmann, 1988.

P. Gärdenfors and H. Rott. Belief revision. In D. Gabbay, C. Hoger, and J. Robinson, editors, *Epistemic and Temporal Reasoning*. Oxford University Press, 1995.

J. Gerbrandy. *Bisimulations on Planet Kripke*. PhD thesis, Institute for Logic, Language and Computation (ILLC), Universiteit van Amsterdam (UvA), Amsterdam, The Netherlands, 1999. ILLC Dissertation Series DS-1999-01.

D. Grossi and F. R. Velázquez-Quesada. A dynamic study of awareness, implicit and explicit knowledge. Working paper, 2012.

A. Grove. Two modellings for theory change. *Journal of Philosophical Logic*, 17 (2):157–170, 1988.

J. Halpern, editor. *Proc. TARK '86*, 1986. Morgan Kaufmann Publishers Inc.

D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic*. MIT Press, 2000.

J. Hintikka. *Knowledge and Belief*. Cornell University Press, 1962.

W. H. Holliday and T. F. Icard. Moorean phenomena in epistemic logic. In L. Beklemishev, V. Goranko, and V. Shehtman, editors, *Advances in Modal Logic*, pages 178–199. College Publications, 2010.

M. Jago. Rule-based and resource-bounded: A new look at epistemic logic. In T. Ågotnes and N. Alechina, editors, *Proc. Workshop on Logics for Resource-Bounded Agents, (ESSLLI 2006)*, pages 63–77, 2006.

K. Konolige. Belief and incompleteness. Technical Report 319, SRI International, 1984.

G. Lakemeyer. Steps towards a first-order logic of explicit and implicit belief. In Halpern (1986), pages 325–340.

P. Lamarre. S4 as the conditional logic of nonmonotonicity. In J. F. Allen, R. Fikes, and E. Sandewall, editors, *KR 91*, pages 357–367. Morgan Kaufmann, 1991.

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

D. K. Lewis. *Counterfactuals*. Blackwell, 1973.

J. A. Plaza. Logics of public communications. In M. L. Emrich, M. S. Pfeifer, M. Hadzikadic, and Z. W. Ras, editors, *Proc. Symposium on Methodologies for Intelligent Systems*, pages 201–216, 1989.

H. Rott. *Change, Choice and Inference: a Study of Belief Revision and Nonmonotonic Reasoning*. Oxford Science Publications, 2001.

K. Segerberg. The basic dynamic doxastic logic of AGM. In Williams and Rott (2001), pages 57–84.

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

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

J. van Benthem. Merging observation and access in dynamic logic. *Journal of Logic Studies*, 1(1):1–17, 2008.

J. van Benthem and F. R. Velázquez-Quesada. The dynamics of awareness. *Synthese (Knowledge, Rationality and Action)*, 177(Supplement 1):5–27, 2010.

H. van Ditmarsch. Prolegomena to dynamic logic for belief revision. *Synthese*, 147(2):229–275, 2005.

H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*. Springer, 2007.

H. van Ditmarsch, A. Herzig, J. Lang, and P. Marquis. Introspective forgetting. *Synthese (Knowledge, Rationality and Action)*, 169(2):405–423, 2009.

H. P. van Ditmarsch and T. French. Awareness and forgetting of facts and agents. In *Web Intelligence/IAT Workshops*, pages 478–483. IEEE, 2009.

J. van Eijck and Y. Wang. Propositional dynamic logic as a logic of belief revision. In W. Hodges and R. de Queiroz, editors, *WoLLIC*, volume 5110 of *LNCS*, pages 136–148. Springer, 2008.

M. Vardi. On epistemic logic and logical omniscience. In Halpern (1986), pages 293–305.

F. R. Velázquez-Quesada. Inference and update. *Synthese (Knowledge, Rationality and Action)*, 169(2):283–300, 2009a.

F. R. Velázquez-Quesada. Dynamic logics for explicit and implicit information. In X. He, J. F. Horty, and E. Pacuit, editors, *LORI-II*, volume 5834 of *LNCS*, pages 325–326. Springer, 2009b.

F. R. Velázquez-Quesada. Dynamic epistemic logic for implicit and explicit beliefs. To appear in JoLLI, 2011.

F. Veltman. *Logics for Conditionals*. PhD thesis, Universiteit van Amsterdam, 1985.

M.-A. Williams and H. Rott, editors. *Frontiers in Belief Revision*, 2001. Kluwer Academic Publishers.

Acts of Requesting in Dynamic Logic of Knowledge and Obligation

Tomoyuki Yamada

Graduate School of Letters, Hokkaido University
yamada@let.hokudai.ac.jp

Abstract

Although it seems intuitively clear that acts of requesting are different from acts of commanding, it is not very easy to sate their differences precisely in dynamic terms. In this paper we show that it becomes possible to characterize, at least partially, the effects of acts of requesting and compare them with the effects of acts of commanding by combining dynamified deontic logic with epistemic logic. One interesting result is the following: each act of requesting is appropriately differentiated from an act of commanding with the same content, but for each act of requesting, there is another act of commanding with much more complex content which updates models in exactly the same way as it does. We will also consider an application of our characterization of acts of requesting to acts of asking yes-no questions. It yields a straightforward formalization of the view of acts of asking questions as requests for information.
[Keywords] request, command, yes-no question, dynamified deontic logic, epistemic logic

1 Introduction

Acts of requesting seem undoubtedly different from acts of commanding. As Searle and Vanderveken have clearly stated, a request "allows for the possibility of refusal" (Searle and Vanderveken 1985, p. 199), but a command "commits the speaker to not giving him [= the commandee (the present author's clarification)] the option of refusal" (ibid., p. 201). Of course this does not mean that it is impossible to refuse to obey a command; but "when one refuses to obey an order or command, one cannot say that one refuses the order or command but rather that one refuses to *obey* it" because "[s]trictly speaking, one can only accept or refuse a speech act that allows for the option of acceptance or refusal" (ibid., p. 195). Thus "one can say literally 'I refused the offer' or 'I refused the invitation' "(ibid.), but one cannot say "I refused the command."

But what does this difference amount to in dynamic terms? In what way is the situation after an act of requesting different from the situation after an act of commanding? And what effects does an act of requesting bring about if it does not exclude the possibility of refusal? The purpose of this paper is to answer these and other related questions concerning the distinction between requesting and commanding by developing a dynamic logic in which effects of acts of requesting and commanding can be compared. For this purpose we will extend DMDL$^+$III ("Dynamified" Multi-agent Deontic Logic + alethic modality III) developed in Yamada (2008a), by adding epistemic operators to it. Since an act of requesting allows for the possibility of refusal, an agent who makes a request will be in need of knowing whether it will be granted or refused, and an appropriate response to a request should address this question. One interesting result is the following: each act of requesting is appropriately differentiated from an act of commanding with the same content, but for each act of requesting, there is another act of commanding with much more complex content which updates models in exactly the same way as it does. We will also consider an application of our analysis of acts of requesting to acts of asking yes-no questions. It yields a straightforward formalization of the view of acts of asking questions as requests for information.

The structure of this paper is as follows. In Section 2, we review the development of dynamified deontic logics that leads to DMDL$^+$III closely, and show how acts of commanding and acts of promising are modeled in DMDL$^+$III. In Section 3, we add epistemic operators to DMDL$^+$III, and briefly examine what more can be said about acts of commanding and promising with their help.

We then show how the workings of acts of requesting can be captured in the extended logic DMEDL (Dynamified Multi-agent Epistemic Deontic Logic) in Section 4. In Section 5, we first compare acts of requesting with acts of commanding further in order to show how each act of requesting is differentiated from an act of commanding with the same content (this illustrates the first part of the above mentioned result), and then show how our analysis of acts of requesting can be applied to the formalization of the notion of questions as requests for information by modeling acts of asking yes-no questions. In Section 6, we prove the second part of the above mentioned result: for each act of requesting, there is another act of commanding with much more complex content which updates models in exactly the same way as it does. Then we conclude with a brief discussion of the implications of this result and further research possibilities in Section 7.

Before proceeding to the next section, we would like to make a disclaimer here in order to make our goal clear. When we talk about acts of requesting and commanding in this paper, we have acts of commanding and requesting performed in a natural language in mind. But we will not deal directly with the semantics of natural language sentences used in performing these acts, but rather with the dynamic nature of the performed acts themselves. We will try to characterize what acts of commanding and acts of requesting are in terms of the effects they bring about. In doing so, we will not aim to capture the pragmatic mechanisms that explain, for example, how the utterance of one and the same sentence of natural language counts as the performance of an act of commanding in one context and that of an act of requesting in another, either. We will rather aim to capture what the act of commanding and the act of requesting accomplish when they are performed in the respective contexts, and retrospectively elucidate each act as the kind of act that accomplishes those kinds of things.

2 Acts of commanding and acts of promising in DMDL$^+$III

DMDL$^+$III is one of the "dynamified" logics inspired by the development of systems of DEL (Dynamic Epistemic Logic). In this section, we first give a brief look at PAL (Public Announcement Logic), the simplest system that falls under DEL, and illustrate how it dynamifies static epistemic logic. Then we closely

review the development of ECL (Eliminative Command Logic), the simplest logic that deals with acts of commanding. It dynamifies static deontic logic just like DEL dynamifies static epistemic logic. As DMDL$^+$III is a refinement of ECL, most of the concepts necessary for understanding DMDL$^+$III can be explained in simpler forms in reviewing the development of ECL. After that, we will show how DMDL$^+$III refines ECL in two steps.

2.1 A brief look at PAL

The development of PAL is illustrated in Figure 1 on Page 431 in the form of a diagram. As the upward arrow in Figure 1 indicates, PAL is obtained by adding dynamic modalities, which represent public announcements, to EL. EL is a multi-agent variant of the standard epistemic logic, and the formula of the form $K_i\varphi$ means that the agent i knows that φ. The formula of the form $[\varphi!]\xi$ of PAL means that ξ holds after every truthful public announcement that φ, and thus the formula of the form $[\varphi!]K_i\psi$ means that the agent i knows that ψ after every truthful public announcement that φ.

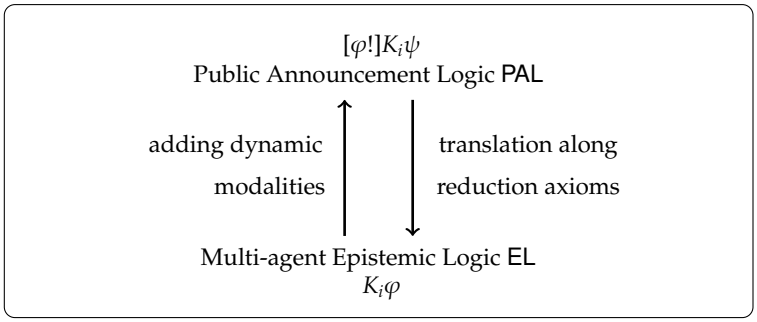


Figure 1: The development of PAL

Given a model M for EL and a world w of M, the public announcement modality $[\varphi!]$ is interpreted by the following clause in the truth definition for the language of PAL:

$$M, w \models_{\mathsf{PAL}} [\varphi!]\xi \text{ iff } M, w \models_{\mathsf{PAL}} \varphi \text{ implies } M_{\varphi!}, w \models_{\mathsf{PAL}} \xi \ ,$$

where $M_{\varphi!}$ is the "updated" model for EL obtained from M by replacing the epistemic accessibility relation R_i for each agent i with its subset $R_i - \{\langle x, y\rangle \in$

$R_i \mid M, x \models_{\mathsf{PAL}} \varphi$ and $M, y \models_{\mathsf{PAL}} \neg\varphi\} - \{\langle x, y\rangle \in R_i \mid M, x \models_{\mathsf{PAL}} \neg\varphi$ and $M, y \models_{\mathsf{PAL}} \varphi\}$.[1] Note that the truth of the formula of the form $[\varphi!]\xi$ at w in M is defined in terms of the truth of the content φ of the announcement $\varphi!$ at w in M and the truth of its subformula ξ at w in the updated model $M_{\varphi!}$. Thus the public announcement of the form $\varphi!$ is interpreted as the type of the events that change the situation (M, w) into $(M_{\varphi!}, w)$. If φ is a formula of EL and no operator of the form K_i occurs in φ, the formula of the form $\varphi \to [\varphi!]K_i\varphi$ is valid. This means that if φ is a non-modal formula, everyone comes to know that φ after every truthful public announcement that φ.[2] An interesting counterexample to the unqualified version of this principle is an announcement of the so-called "Moore formula" $(\varphi \wedge \neg K_i\varphi)$.

PAL is axiomatized by adding a set of so called "reduction axioms" and the necessitation rule for each announcement modality to the proof system of EL. As the downward arrow in Figure 1 indicates, the reduction axioms enable us to define translation function t that takes any formula φ from PAL and yields a formula $t(\varphi)$ of EL that is provably equivalent to φ. This translation in turn enables us to derive the completeness of PAL from the completeness of EL.

2.2 Acts of commanding in ECL

Inspired by the development of PAL and other dynamic epistemic logics, a series of dynamified deontic logics including DMDL$^+$III are developed. Eliminative Command Logic ECL developed in Yamada (2007a) is the simplest one in the series. Figure 2 on Page 433 shows the diagram of the development of ECL. Just like PAL, ECL is obtained by adding dynamic modalities, which represent types of acts of commanding, to the static base logic MDL$^+$ (Multi-agent Deontic Logic + alethic modality) and is axiomatized by adding a set of reduction axioms and

[1] This way of updating is usually called "link-cutting". Another way of updating, called "world elimination", eliminates every non-φ worlds from the domain of the model and restricts R_i to the new domain. For more on PAL and DEL, see van Ditmarsch, van der Hoek, and Kooi (2007). It gives a detailed state-of-the-art textbook exposition of the major systems of DEL as well as a useful historical overview of their development.

[2] Although there can be various artificial agents to which this applies, it seems too strong to be true of agents like us, as there is a possibility of disbelief on the side of the audience. There may be people who are so sceptical that they do not always believe public announcements. This gap, which is a gap between an illocutionary act (announcing that φ) and a perlocutionary act (getting addressees to know that φ, or convincing them that φ), can be avoided if we reinterpret $\varphi!$ as a type of event in which agents simultaneously and publicly learn that φ. Then we can be said to have a theory of (group) learnability in the form of DEL. For more on the gap, see Yamada (2008b).

necessitation rules for command modalities to the proof system of MDL^+.

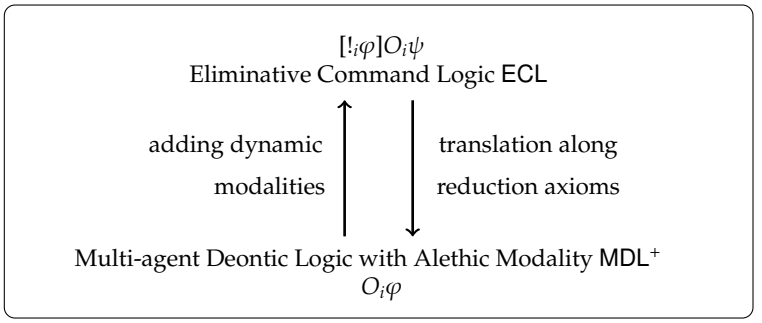


Figure 2: The development of ECL

The formula of the form $O_i\varphi$ means that it is obligatory upon agent i to see to it that φ. Although indexing of deontic operators with a set of agents is not standard in deontic logic, we need to be able to distinguish agents to whom commands are given from other agents if we are to use deontic logic to reason about how acts of commanding change situations. For this purpose, the language of MDL^+ has a separate deontic operator O_i for each agent i. The language and the models of MDL^+ are defined as follows.[3]

Definition 2.1. Take a countably infinite set Aprop of proposition letters and a finite set I of agents, with p ranging over Aprop and i over I. The multi-agent deontic language $\mathcal{L}_{\mathsf{MDL}^+}$ is given by:

$$\varphi ::= \top \mid p \mid \neg\varphi \mid (\varphi \wedge \psi) \mid \Box\varphi \mid O_i\varphi \ .$$

We use standard abbreviations $\vee$, $\rightarrow$, $\leftrightarrow$, and $\Diamond$. In addition, we abbreviate $O_i\neg\varphi$ as $F_i\varphi$, and $\neg O_i\neg\varphi$ as $P_i\varphi$.

Definition 2.2. By an $\mathcal{L}_{\mathsf{MDL}^+}$-model, we mean a tuple $M = \langle W^M, A^M, \{D_i^M \mid i \in I\}, V^M\rangle$ where:

1. W^M is a non-empty set (heuristically, of 'possible worlds' or 'states')
2. $A^M \subseteq W^M \times W^M$

[3]The definition of the models in this paper is slightly different from that of Yamada (2007a), but there is no substantial difference.

3. $D_i^M \subseteq A^M$ for each agent $i \in I$
4. V^M is a function that assigns a subset $V^M(p)$ of W^M to each proposition letter $p \in$ Aprop.

Based on these definitions, the truth definition for the formulas of $\mathcal{L}_{\mathsf{MDL}^+}$ is given in a completely standard way by associating alethic modality $\Box$ with A^M and each deontic modality O_i with D_i^M. Thus the formula of the form $O_i\varphi$, for example, is interpreted by the following clause:

$$M, w \models_{\mathsf{MDL}^+} O_i\varphi \text{ iff for any } v \text{ such that } \langle w, v\rangle \in D_i^M,\ M, v \models_{\mathsf{MDL}^+} \varphi \ .$$

Note that the following axiom, called "Mix" is shown to be valid according to these definitions:

$$P_i\varphi \to \Diamond\varphi \ .$$

This means that what is permitted is possible. The so-called axiom D of the following form, however, is not valid:

$$O_i\varphi \to P_i\varphi \ .$$

Since we may receive conflicting commands from different authorities, we cannot assume D axiom to be valid, as we will see later.

Note also that no restrictions are imposed upon the alethic accessibility A^M. Since further restrictions do not affect the discussion in this paper, we will not bother to add them. Thus,

Definition 2.3. The proof system for MDL^+ contains the following axioms and rules:

(Taut) all instantiations of propositional tautologies over the present language

($\Box$-Dist) $\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$ ($\Box$-distribution)

(O_i-Dist) $O_i(\varphi \to \psi) \to (O_i\varphi \to O_i\psi)$ (O_i-distribution)

(Mix) $P_i\varphi \to \Diamond\varphi$ (Mix Axiom)

(MP) $\dfrac{\varphi \quad \varphi \to \psi}{\psi}$ (Modus Ponens)

($\Box$-Nec) If φ is proved, infer $\Box\varphi$ ($\Box$-necessitation)

(O_i-Nec) If φ is proved, infer $O_i\varphi$. (O_i-necessitation)

The soundness and the completeness of this proof system can be proved in an entirely standard way.

Now let's move on to ECL. As we have seen in Figure 2, the language of ECL is obtained by adding dynamic modalities, which represent types of acts of commanding, to the language of the static base logic MDL^+. Thus,

Definition 2.4. Take the same countably infinite set Aprop of proposition letters and the same finite set I of agents as before, with p ranging over Aprop and i over I. The language of eliminative command logic $\mathcal{L}_{\mathsf{ECL}}$ is given by:

$$\varphi ::= \top \mid p \mid \neg\varphi \mid (\varphi \wedge \psi) \mid \Box\varphi \mid O_i\varphi \mid [\pi]\varphi$$
$$\pi ::= !_i\varphi \ .$$

An expression of the form $!_i\varphi$, which we will call a command term, represents the type of acts of commanding given to a commandee i to the effect that i should see to it that φ, and the formula of the form $[!_i\varphi]\psi$ means that ψ holds after i is commanded to see to it that φ. Note that command terms are not formulas.

The truth definition for this language is given with reference to an $\mathcal{L}_{\mathsf{MDL}^+}$-model by extending the truth definition for $\mathcal{L}_{\mathsf{MDL}^+}$ mutatis mutandis with the following clause for the new formulas:

$$M, w \models_{\mathsf{ECL}} [!_i\varphi]\psi \text{ iff } M_{!_i\varphi}, w \models_{\mathsf{ECL}} \psi \ ,$$

where $M_{!_i\varphi}$ is the updated $\mathcal{L}_{\mathsf{MDL}^+}$-model obtained from M by replacing only the deontic accessibility relation D_i^M for the agent i with its subset $D_i^{M_{!_i\varphi}} = \{\langle x, y\rangle \in D_i^M \mid M, y \models_{\mathsf{ECL}} \varphi\}$. Thus, the update by the act of commanding of the type $!_i\varphi$ only cuts the arrows of deontic accessibility for the agent i which arrive in non-φ-worlds in M; it does not cut any arrows of deontic accessibility for other agents.

Note that the truth of the formula of the form $[!_i\varphi]\psi$ at w in M is defined in terms of the truth of its subformula ψ at w in the updated model $M_{!_i\varphi}$. This fits the intended meaning of $[!_i\varphi]\psi$, namely that ψ holds after i is commanded to see to it that φ. Note also that $D_i^{M_{!_i\varphi}} \subseteq D_i^M$. This guarantees that updated models will always be $\mathcal{L}_{\mathsf{MDL}^+}$-models.

Figure 3 on Page 436 gives an image of how an act of commanding works in an example taken from Yamada (2007a). Imagine the following situation. You

are working in an office shared with your boss and a few other colleagues on a hot day in summer. There is a window, but it is closed now. There is an air conditioner, but it is not running now. The temperature is rising and it now is at 30 degrees Celsius. The model-state pair (M, s_0) represents this situation.

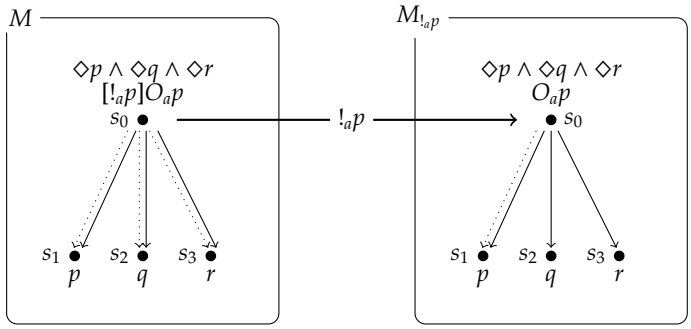


Figure 3: Your boss's command

Let p stand for the proposition that the window is open, q for the proposition that the air conditioner is running, and r for the proposition that the temperature is above 30 degree Celsius. The presence of formulas near the states indicates that they hold in these states, and the absence of proposition letters near the states (but not the absence of non-atomic formulas) indicates that they do not hold in these states. The solid arrows represent the alethic accessibility, and the dotted arrows represent the deontic accessibility for you, here represented by a.[4] Thus you can open the window, or turn on the air conditioner, or even ignore the heat by concentrating on your work. All these alternatives are possible and permitted for you in (M, s_0).

But now you hear your boss's voice. She commanded you to open the window. The pair $(M_{!_a p}, s_0)$ represents the situation you are in after your boss's act commanding. All the alternatives that were possible in (M, s_0) are still possible in $(M_{!_a p}, s_0)$. But in order to obey your boss's command, you have to open the window. It becomes the only permissible alternative to you now. This effect of her command is modeled by cutting the arrows of deontic accessibility for you

[4]For the sake of simplicity, arrows of deontic accessibility relations for other people and the reflexive arrows for alethic accessibility are omitted.

that arrive in non-p-states in M. Thus we have $M_{!_a p}, s_0 \models_{\mathsf{ECL}} O_a p$, and this in turn means that we have $M, s_0 \models_{\mathsf{ECL}} [!_a p] O_a p$. The thick arrow from (M, s_0) to $(M_{!_a p}, s_0)$ is an imaginary arrow in the sense that it is neither in M nor in $M_{!_a p}$, but it helps us to understand your boss's command of the form $!_a p$ as the event that takes you from (M, s_0) to $(M_{!_a p}, s_0)$.

Note that the treatment of acts of commanding in ECL is based on a simplifying assumption that command issuing agents have suitable authority over commandees. We follow this treatment in this paper.[5] With the help of this simplifying assumption, the following result is obtained:

Proposition 1 (The CUGO Principle)**.** *If φ is a formula of* MDL^+ *and is free of modal operators of the form O_i, $[!_i\varphi]O_i\varphi$ is valid.*

This principle means that, though not without exceptions, commands usually generate obligations (hence "CUGO"). It partially characterizes the effects of acts of commanding.[6] As we have said in Section 1, the purpose of this paper is to give a similar kind of characterization to acts of requesting by extending a refinement $\mathsf{DMDL}^+\mathsf{III}$ of ECL.

Note that the unqualified version of the CUGO principle is not valid. A bit of terminology is of some help here. Let O be some modal operator. We call the pair $\langle x, y\rangle$ of worlds an O-arrow and say y is O-accessible from x if $\langle x, y\rangle$ is in the accessibility relation R that interprets O. Then we can say that the update by an act of commanding of the type $!_i\varphi$ cuts every O_i-arrow that arrives in $\neg\varphi$-worlds in M. This guarantees that every O_i-arrow that remains after this update arrives in a world in which φ holds in M. But this does not guarantee that φ holds there in the updated model $M_{!_i\varphi}$. If O_i occurs in φ, φ might be false at some world O_i-accessible from w in $M_{!_i\varphi}$.

ECL is axiomatized by adding a set of so-called "reduction axioms" and the necessitation rule for each command operator to the proof system of MDL^+. Thus,

Definition 2.5. The proof system for ECL contains all the axioms and all the rules of the proof system for MDL^+, and in addition the following reduction axioms and rules:

[5] The standard method used in order to treat preconditions for action like this is to introduce a function *pre* that assigns to each event term e its precondition $pre(e)$. For more on this, see Baltag, Moss, and Solecki (1998) or van Ditmarsch, van der Hoek, and Kooi (2007).

[6] This characterization is partial because acts of commanding involve other effects as well. For more on this, see Sections 6 and 7.

(RA1) $[!_i\varphi]p \leftrightarrow p$ where $p \in Aprop$

(RA2) $[!_i\varphi]\top \leftrightarrow \top$

(RA3) $[!_i\varphi]\neg\psi \leftrightarrow \neg[!_i\varphi]\psi$

(RA4) $[!_i\varphi](\psi \wedge \chi) \leftrightarrow ([!_i\varphi]\psi \wedge [!_i\varphi]\chi)$

(RA5) $[!_i\varphi]\Box\psi \leftrightarrow \Box[!_i\varphi]\psi$

(RA6) $[!_i\varphi]O_j\psi \leftrightarrow O_j[!_i\varphi]\psi$ where $i \neq j$

(RA7) $[!_i\varphi]O_i\psi \leftrightarrow O_i(\varphi \rightarrow [!_i\varphi]\psi)$

($[!_i\varphi]$-Nec) If ψ is proved, infer $[!_i\varphi]\psi$.

The crucial axiom here is RA7. The formula on the left hand side, $[!_i\varphi]O_i\psi$, states that $O_i\psi$ holds after the update. The formula on the right hand side specifies the necessary and sufficient condition for this in terms of what holds before the update. Take an arbitrary $\mathcal{L}_{\mathsf{MDL}^+}$-model M and a world w of M. In order for $O_i\psi$ to hold in w in the updated model $M_{!_i\varphi}$, ψ must hold in every world O_i-accessible from w in $M_{!_i\varphi}$. But those worlds are exactly the φ-worlds in M that are O_i-accessible from w in M. In order for ψ to hold in those worlds after the update, $[!_i\varphi]\psi$ has to hold in those world before the update. Thus $O_i(\varphi \rightarrow [!_i\varphi]\psi)$ has to hold in w in M.

Note that the first two axioms enable us to eliminate command operators prefixed to proposition letters and $\top$. The remaining axioms enable us to reduce the length of the subformula to which command operators are prefixed step by step. Thus these axioms enable us to define translation function that takes any formula of ECL and yields a formula of MDL^+ which is provably equivalent to the original formula. This translation in turn enables us to derive the completeness of ECL from that of MDL^+.

2.3 Conflicting commands in ECL and ECLII

Now we can move on to refinements. The following results about ECL are reported in Yamada (2007a):

Proposition 2 (The Dead End Principle)**.** $[!_i(\varphi \wedge \neg\varphi)]O_i\xi$ *is valid.*

Proposition 3 (The Restricted Sequential Conjunction Principle)**.** *If φ and ψ are formulas of* MDL^+ *and free of modal operators of the form O_i,* $[!_i\varphi][!_i\psi]\xi \leftrightarrow [!_i(\varphi \wedge \psi)]\xi$ *is valid.*

The dead end principle means that if an agent receives a command with contradictory content, everything comes to be obligatory upon him. The situation of this kind is usually called "deontic explosion". Since the updated by $!_i(\varphi \wedge \neg\varphi)$ cuts every O_i-arrow that arrives in a $\neg(\varphi \wedge \neg\varphi)$-world, it cuts every O_i-arrow, and so $D_i^{M^{!_i(\varphi\wedge\neg\varphi)}}$ becomes empty. Thus $O_i\xi$ becomes vacuously true in every world after the update by $!_i(\varphi \wedge \neg\varphi)$.

If we put $\neg\varphi$ in the place of ψ in the restricted sequential conjunction principle, we get deontic explosion again. Situations of this kind can arise in real life as an agent might receive such a pair of commands from different command issuing authorities.[7]

$\mathsf{MDL}^+\mathsf{II}$ and ECLII refine MDL^+ and ECL respectively in order to deal with conflicting commands in a more satisfactory way by indexing deontic operators and deontic accessibility relations by the set $I \times I$ of pairs of agents (Yamada 2007b). Thus the formula of the form $O_{(i,j)}\varphi$ from the static base logic $\mathsf{MDL}^+\mathsf{II}$ means that it is obligatory upon an agent i with respect to the authority j to see to it that φ, and the formula of the form $[!_{(i,j)}\varphi]\psi$ of the dynamified logic ECLII means that ψ holds after an authority j's act of commanding an agent i to see to it that φ. The definitions of the languages, the models, the relations of truth in models, and the proof systems for $\mathsf{MDL}^+\mathsf{II}$ and ECLII are given in the same way as those for MDL^+ and ECL except for the indexing by $I \times I$.

Since I is a finite set, indexing by $I \times I$ is just an instance of indexing by a finite set. Thus $\mathsf{MDL}^+\mathsf{II}$ and ECLII are just another instantiations of MDL^+ and ECL respectively, hence all the results obtained for MDL^+ and ECL apply to $\mathsf{MDL}^+\mathsf{II}$ and ECLII mutatis mutandis. In particular, the CUGO principle now reads:

Proposition 4 (The CUGO Principle)**.** *If φ is a formula of $\mathsf{MDL}^+\mathsf{II}$ and is free of modal operators of the form $O_{(i,j)}$, $[!_{(i,j)}\varphi]O_{(i,j)}\varphi$ is valid.*

With the help of this principle, we now have:

$$(M_{!_{(a,b)}p})_{!_{(a,c)}\neg p}, w \models_{\mathsf{ECLII}} (O_{(a,b)}p \wedge O_{(a,c)}\neg p) \ .$$

This is the situation a will be in after a receives a command from an authority c to the effect that a should see to it that $\neg p$ after a receives a command from another authority b to the effect that a should see to it that p. Since it is not possible to obey both commands in this example, a has to decide which command to obey.

[7] Thus D Axiom cannot be included in the proof system of MDL^+.

Note that this combination of incompatible commands does not generally produce deontic explosion. In $(M_{!_{(a,b)}p})_{!_{(a,c)}\neg p}$, p-worlds that are $O_{(a,b)}$-accessible in M (if any) and $\neg p$-worlds that are $O_{(a,c)}$-accessible in M (if any) will remain $O_{(a,b)}$-accessible and $O_{(a,c)}$-accessible respectively, since the update by $!_{(a,b)}p$ only cuts $O_{(a,b)}$-arrows arriving in $\neg p$-worlds in M and the update by $!_{(a,c)}\neg p$ only cuts $O_{(a,c)}$-arrows arriving in p-worlds in M. Deontic explosions occur only when incompatible commands are given to one and the same agent by one and the same authority or an authority issues a command having contradictory content. If the command issuing authority is rational, such a situation will be avoided; otherwise, obedience could not be expected.

Note also that similar conflicts can arise between requests as well as between a request and a command. So, we will follow this treatment in developing our system later.

2.4 DMDL$^+$III

DMDL$^+$III refines ECLII further in order to model acts of promising along with acts of commanding (Yamada 2008a). In the case of acts of commanding, obligations commandees owe are created by command issuing authorities (commanders, for short).But in the case of acts of promising, obligations owed by agents who give promises (promisers, for short) are created by promisers themselves. Moreover, agents to whom promises are given (promisees, for short) will be entitled to rely on promisers to do what they have promised to do. In order to deal with this complexities, deontic operators and their corresponding accessibility relations are indexed by the set $I \times I \times I$ in the static base logic MDL$^+$III. As before, indexing by $I \times I \times I$ is just indexing by a finite set, and thus MDL$^+$III is yet another instantiation of MDL$^+$. But this time DMDL$^+$III includes more than ECL does. It deals not only with acts of commanding but also with acts of promising.

In MDL$^+$III and in DMDL$^+$III, the formula of the form $O_{(i,j,k)}\varphi$ means that it is obligatory upon agent i with respect to j in the name of k to see to it that φ. The agent i here is the agent who owes the obligation (sometimes called an obligor), j is the agent to whom the obligation is owed (sometimes called an obligee), and k is the creator of the obligation. As we will see shortly, they need not be distinct.

In DMDL$^+$III, the formula of the form $[Com_{(i,j)}\varphi]\psi$ means that ψ holds after an

agent i commands an agent j to see to it that φ, and the formula of the form $[Prom_{(i,j)}\varphi]\psi$ means that ψ holds after an agent i promises an agent j that she (i) will see to it that φ. Note that the order of the parameters in the command term $Com_{(i,j)}\varphi$ is changed from that of the term $!_{(i,j)}\varphi$ of ECLII. In $Com_{(i,j)}\varphi$, i is the commander and j is the commandee.

In the truth definition for the language of $\mathsf{DMDL}^{+}\mathsf{III}$, the added dynamic formulas are interpreted by the following clauses:

$$M, w \models_{\mathsf{DMDL}^{+}\mathsf{III}} [Com_{(i,j)}\varphi]\xi \text{ iff } M_{Com_{(i,j)}\varphi}, w \models_{\mathsf{DMDL}^{+}\mathsf{III}} \xi$$
$$M, w \models_{\mathsf{DMDL}^{+}\mathsf{III}} [Prom_{(i,j)}\varphi]\xi \text{ iff } M_{Prom_{(i,j)}\varphi}, w \models_{\mathsf{DMDL}^{+}\mathsf{III}} \xi,$$

where

1. $M_{Com_{(i,j)}\varphi}$ is the $\mathcal{L}_{\mathsf{MDL}^{+}\mathsf{III}}$-model obtained from M by replacing $D^{M}_{(j,i,i)}$ with its subset $\{\langle x, y\rangle \in D^{M}_{(j,i,i)} \mid M, y \models_{\mathsf{DMDL}^{+}\mathsf{III}} \varphi\}$, and
2. $M_{Prom_{(i,j)}\varphi}$ is the $\mathcal{L}_{\mathsf{MDL}^{+}\mathsf{III}}$-model obtained from M by replacing $D^{M}_{(i,j,i)}$ with its subset $\{\langle x, y\rangle \in D^{M}_{(i,j,i)} \mid M, y \models_{\mathsf{DMDL}^{+}\mathsf{III}} \varphi\}$.

Thus the update by $Com_{(i,j)}\varphi$ only cuts $O_{(j,i,i)}$-arrows arriving in $\neg\varphi$-worlds in the original model M, and the update by $Prom_{(i,j)}\varphi$ only cuts $O_{(i,j,i)}$-arrows arriving in $\neg\varphi$-world in the original model M.

Again, we have:

Proposition 5. The CUGO Principle: *If φ is a formula of* $\mathsf{MDL}^{+}\mathsf{III}$ *and is free of modal operators of the form $O_{(j,i,i)}$, $[Com_{(i,j)}\varphi]O_{(j,i,i)}\varphi$ is valid.*

And in addition to this, we have:

Proposition 6. The PUGO Principle: *If φ is a formula of* $\mathsf{MDL}^{+}\mathsf{III}$ *and is free of modal operators of the form $O_{(i,j,i)}$, $[Prom_{(i,j)}\varphi]O_{(i,j,i)}\varphi$ is valid.*

These principles partially capture how acts of commanding and promising work.

Note the differences between the obligations generated. In the case of the obligation generated by an act of commanding the commandee j owes the obligation created by the commander i, but in the case of the obligation generated by an act of promising the promiser i owes the obligation created by

the promiser i herself, and the promisee j is the agent whom the obligation is owed. This difference enables us to consider the obligations created by acts of promising as representing the commitments of the promisers. This will be of some importance when we analyze acts of requesting. [8]

We are now in a position to extend $\mathsf{MDL}^+\mathsf{III}$ and $\mathsf{DMDL}^+\mathsf{III}$.

3 The securing of uptake in DMEDL

We add epistemic operators to $\mathsf{MDL}^+\mathsf{III}$. For the sake of simplicity, we ignore alethic modality. Thus we define:

Definition 3.1. Take a countably infinite set **Aprop** of proposition letters, and a finite set I of agents, with p ranging over **Aprop**, and i, j, k over I. The language $\mathcal{L}_{\mathsf{MEDL}}$ of the Multi-agent Epistemic Deontic Logic MEDL is given by:

$$\varphi ::= \top \mid p \mid \neg\varphi \mid (\varphi \wedge \psi) \mid K_i\varphi \mid O_{(i,j,k)}\varphi$$

Definition 3.2. By an $\mathcal{L}_{\mathsf{MEDL}}$-model, we mean a tuple $M = \langle W^M, \{E_i^M \mid i \in I\}, \{D^M_{(i,j,k)} \mid i, j, k \in I\}, V^M\rangle$ where:

1. W^M is a non-empty set (heuristically, of 'possible worlds' or 'states')
2. E_i^M is an equivalence relation such that $E_i^M \subseteq W^M \times W^M$
3. $D^M_{(i,j,k)} \subseteq W^M \times W^M$
4. V^M is a function that assigns a subset $V^M(p)$ of W^M to each proposition letter $p \in$ Aprop.

The truth definition (defining the relation $\models_{\mathsf{MEDL}}$) and the definition of the proof system (defining the relation $\vdash_{\mathsf{MEDL}}$) can be given in an entirely standard way. Since MEDL is a simple fusion of the deontic fragment of $\mathsf{MDL}^+\mathsf{III}$ and the multi-agent variant of the standard epistemic logic, there is a complete axiomatization of it.

We dynamify MEDL into DMEDL (Dynamified MEDL) by adding the dynamic modalities indexed by action terms of the forms $Com_{(i,j)}\varphi$, $Prom_{(i,j)}\varphi$ and $Req_{(i,j)}\varphi$.

[8] There may be room for disagreement over whether the notion of the agent whom the obligation is owed make sense with respect to the obligation created by an act of commanding. But we will not pursue this point here.

The formula of the form $[Com_{(i,j)}\varphi]\psi$ and the formula of the form $[Prom_{(i,j)}\varphi]\psi$ are interpreted in exactly the same way as in $\mathsf{DMDL}^{+}\mathsf{III}$, except with reference to $\mathcal{L}_{\mathsf{MEDL}}$-models and with $\models_{\mathsf{DMDL}^{+}\mathsf{III}}$ replaced with $\models_{\mathsf{DMEDL}}$. Thus we again have the DMEDL versions of the CUGO principle and the PUGO principle.

Before moving on to the analysis of acts of requesting, we take a brief look at what more we can say about acts of commanding and promising in DMEDL. One immediate consequence of having epistemic operators is the fact that we can now talk about the knowledge agents have about effects of speech acts. When a command is successfully given, for example, the commandee must know what command she has been given. Unless the force and the content is understood, no illocutionary act can be successfully performed, since the effects of illocutionary acts depend on the agreement on (and so the understanding of) what has been performed. Surprisingly, the following principles are valid in DMEDL:

Proposition 7. The CUGU Principle: *If φ is a formula of* DMEDL *and is free of modal operators of the form $O_{(j,i,i)}$, $[Com_{(i,j)}\varphi]K_jO_{(j,i,i)}\varphi$ is valid.*

Proposition 8. The PUGU Principle: *If φ is a formula of* DMEDL *and is free of modal operators of the form $O_{(i,j,i)}$, $[Prom_{(i,j)}\varphi]K_jO_{(i,j,i)}\varphi$ is valid.*

These principles state that acts of commanding and acts of promising usually generate knowledge of the effects captured in the CUGO principle and the PUGO principle respectively on the side of addressees.

We call these principles "CUGU" and "PUGU" because we believe that these principles characterize what Austin calls "the securing of uptake". According to Austin, "the securing of uptake" means "bringing about the understanding of the meaning and of the force of the locution". It is the "effect" that "must be achieved on the audience if the illocutionary act is to be carried out." And so,"the performance of an illocutionary act involves the securing of uptake"(Austin 1975, pp. 117-118). In the case of an act of commanding, the understanding of the force means the understanding of the commander's locution as an act of commanding and the understanding of the meaning of her locution includes the understanding of what is commanded. The CUGU principle partially characterizes what these understanding amount to. The same thing can be said of the PUGU principle as well.

We said "surprisingly" above because no epistemic update operation is required for these results. Take any model M, any world w of M, and any proposition letter p for example. After an act of commanding of the form $Com_{(i,j)}p$ is performed

in a situation (M, w), $O_{(j,i,i)}p$ holds in any world v in the updated model $M_{Com_{(i,j)}p}$ as every $O_{(j,i,i)}$-arrow arriving in a non-p-world of M is eliminated in $M_{Com_{(i,j)}p}$, and every p-world of M remains to be a p-world in $M_{Com_{(i,j)}p}$. But if $O_{(j,i,i)}p$ holds in any world in the updated model $M_{Com_{(i,j)}p}$, it holds in any world K_j-accessible from any world of M. Thus $K_jO_{(j,i,i)}p$ holds in any world in $M_{Com_{(i,j)}p}$.

We need to note, however, that the same thing holds for any agent $i \in I$. It means that everyone comes to know that $O_{(j,i,i)}p$ in w in $M_{Com_{(i,j)}p}$. This is natural when we consider a small everyday situation like the situation of the shared office on the hot summer day we considered earlier. But even in a small everyday situations like this, there are many ways in which only some of the agents come to know what speech act is performed by a particular person at a particular time.

Here it is important to understand how everyone comes to know $O_{(j,i,i)}p$ in w in $M_{Com_{(i,j)}p}$. Although the update by $Com_{(i,j)}p$ does not affect any epistemic accessibility relations, it makes $O_{(j,i,i)}p$ true in any worlds epistemically accessible for each agent. In that sense, the independence of each agent's epistemic accessibility relation from that of others does not fully model the privacy of knowledge in the context of dynamified modal logics. The standard way to model the distinction between agents who know what happens and those who do not is to introduce the so-called "event models", in which (un)certainty of each agent as regards what has happened is modeled, and define the update operation called "product update". If we do this for MEDL, our current update by $Com_{(i,j)}\varphi$ will be modeled as the special case of product update by the event model in which every agent knows that an event of the type $Com_{(i,j)}\varphi$ happens. Although it is possible to extend MEDL by introducing event models and product update, we will not pursue this possibility here as there are many things yet to be done before making life more complicated.[9]

4 Acts of requesting in DMEDL

Now we move on to the analysis of acts of requesting. As we have seen in Section 1, an act of requesting allows the possibility of refusal. As a consequence of this, the following principle is not valid even if no operators of the form $O_{(j,i,i)}$

[9] I have benefitted from a discussion with Johan van Benthem on this point. For more on the product update, see Baltag, Moss, and Solecki (1998) or van Ditmarsch, van der Hoek, and Kooi (2007).

occur in φ:

$$[Req_{(i,j)}\varphi]O_{(j,i,i)}\varphi \quad .$$

In this respect, acts of requesting stand in sharp contrast to acts of commanding, for which we have the CUGO principle. But it is also clear that it would not be without any problems if an agent who has been requested to do something (a requestee, for short) gives no response. Although it is not obligatory upon the requestee to do what is requested, it is obligatory upon her to decide whether she should do what is requested. Moreover, she has to let the agent who has made the request (the requester, for short) know her decision.

If the requestee j decides that she should do what is requested, and the requested action is not the kind of thing to be done on the spot, she can promise the requester i that she (j) will do what is requested. As the PUGU principle indicates, the requester i will know that $O_{(j,i,j)}\varphi$. If the requestee j decides that she (j) should reject the request, she (j) should let the requester i know that $\neg O_{(j,i,j)}\varphi$.

Now what about the case in which what is requested can be done on the spot. If the requestee j decides that she should do what is requested, she might do it on the spot without saying anything. Whether we should count this as the third alternative way of responding to an act of requesting, or consider it as skipping to the sequel of an implicit promise might be a matter of opinion. We take the formulation with the three options.[10] Thus the clause for the formula of the form $[Req_{(i,j)}\varphi]\xi$ reads:

$$M, w \models_{\mathsf{DMEDL}} [Req_{(i,j)}\varphi]\xi \text{ iff } M_{Req_{(i,j)}\varphi}, w \models_{\mathsf{DMEDL}} \xi \quad ,$$

where $M_{Req_{(i,j)}\varphi}$ is the $\mathcal{L}_{\mathsf{MEDL}}$-model obtained from M by replacing $D^M_{(j,i,i)}$ with its subset $\{\langle x, y\rangle \in D^M_{(j,i,i)} \mid M, y \models_{\mathsf{DMEDL}} (\varphi \vee K_i O_{(j,i,j)}\varphi \vee K_i \neg O_{(j,i,j)}\varphi)\}$.[11]

This interpretation supports the following principles.

Proposition 9. The RUGO Principle: *If φ is a formula of* MEDL *and is free of modal operators of the form $O_{(j,i,i)}$, $[Req_{(i,j)}\varphi]O_{(j,i,i)}(\varphi \vee K_i O_{(j,i,j)}\varphi \vee K_i \neg O_{(j,i,j)}\varphi)$ is valid.*

[10] Traum (1999, p. 195) also talks about similar obligations as effects of acts of requesting, but he includes only the options of accepting or refusing.

[11] The formulation with two options can be obtained by using $\{\langle x, y\rangle \in D^M_{(j,i,i)} \mid M, y \models_{\mathsf{DMEDL}} (K_i O_{(j,i,j)}\varphi \vee K_i \neg O_{(j,i,j)}\varphi)\}$ instead.

Proposition 10. The RUGU Principle: *If φ is a formula of* DMEDL *and is free of modal operators of the form $O_{(j,i,i)}$, $[Req_{(i,j)}\varphi]K_jO_{(j,i,i)}(\varphi \vee K_iO_{(j,i,j)}\varphi \vee K_i\neg O_{(j,i,j)}\varphi)$ is valid.*

We are now in a position to define the proof system for DMEDL. We first list three sets of reduction axioms.

Theorem 1 (Reduction axioms for acts of commanding)**.** *The following axioms are valid in* DMEDL.

(C1) $[Com_{(i,j)}\varphi]p \leftrightarrow p$

(C2) $[Com_{(i,j)}\varphi]\top \leftrightarrow \top$

(C3) $[Com_{(i,j)}\varphi]\neg\psi \leftrightarrow \neg[Com_{(i,j)}\varphi]\psi$

(C4) $[Com_{(i,j)}\varphi](\psi \wedge \chi) \leftrightarrow [Com_{(i,j)}\varphi]\psi \wedge [Com_{(i,j)}\varphi]\chi$

(C5) $[Com_{(i,j)}\varphi]K_l\psi \leftrightarrow K_l[Com_{(i,j)}\varphi]\psi$

(C6) $[Com_{(i,j)}\varphi]O_{(l,m,n)}\psi \leftrightarrow O_{(l,m,n)}[Com_{(i,j)}\varphi]\psi$ *if* $\langle l,m,n\rangle \neq \langle j,i,i\rangle$

(C7) $[Com_{(i,j)}\varphi]O_{(j,i,i)}\psi \leftrightarrow O_{(j,i,i)}(\varphi \rightarrow [Com_{(i,j)}\varphi]\psi)$

Theorem 2 (Reduction axioms for acts of Promising)**.** *The following axioms are valid in* DMEDL.

(P1) $[Prom_{(i,j)}\varphi]p \leftrightarrow p$

(P2) $[Prom_{(i,j)}\varphi]\bot \leftrightarrow \bot$

(P3) $[Prom_{(i,j)}\varphi]\neg\psi \leftrightarrow \neg[Prom_{(i,j)}\varphi]\psi$

(P4) $[Prom_{(i,j)}\varphi](\psi \wedge \chi) \leftrightarrow [Prom_{(i,j)}\varphi]\psi \wedge [Prom_{(i,j)}\varphi]\chi$

(P5) $[Prom_{(i,j)}\varphi]K_l\psi \leftrightarrow K_l[Prom_{(i,j)}\varphi]\psi$

(P6) $[Prom_{(i,j)}\varphi]O_{(l,m,n)}\psi \leftrightarrow O_{(l,m,n)}[Prom_{(i,j)}\varphi]\psi$ *if* $\langle l,m,n\rangle \neq \langle i,j,i\rangle$

(P7) $[Prom_{(i,j)}\varphi]O_{(i,j,i)}\psi \leftrightarrow O_{(i,j,i)}(\varphi \rightarrow [Prom_{(i,j)}\varphi]\psi)$

Theorem 3 (Reduction axioms for acts of Requesting)**.** *The following axioms are valid in* DMEDL.

(R1) $[Req_{(i,j)}\varphi]p \leftrightarrow p$

(R2) $[Req_{(i,j)}\varphi]\bot \leftrightarrow \bot$

(R3) $[Req_{(i,j)}\varphi]\neg\psi \leftrightarrow \neg[Req_{(i,j)}\varphi]\psi$

(R4) $[Req_{(i,j)}\varphi](\psi \wedge \chi) \leftrightarrow [Req_{(i,j)}\varphi]\psi \wedge [Req_{(i,j)}\varphi]\chi$

(R5) $[Req_{(i,j)}\varphi]K_l\psi \leftrightarrow K_l[Req_{(i,j)}\varphi]\psi$

(R6) $[Req_{(i,j)}\varphi]O_{(l,m,n)}\psi \leftrightarrow O_{(l,m,n)}[Req_{(i,j)}\varphi]\psi$ *if* $\langle l,m,n\rangle \neq \langle j,i,i\rangle$

(R7) $[Req_{(i,j)}\varphi]O_{(j,i,i)}\psi \leftrightarrow O_{(j,i,i)}((\varphi \vee K_iO_{(j,i,j)}\varphi \vee K_i\neg O_{(j,i,j)}\varphi) \rightarrow [Req_{(i,j)}\varphi]\psi)$

As before, the first two axioms of each group enable us to eliminate dynamic operators prefixed to proposition letters and $\top$. The remaining axioms enable us to reduce the length of the subformula to which dynamic operators are prefixed step by step.

Now we define:

Definition 4.1 (The proof system for DMEDL)**.** The proof system for DMEDL is comprised of

1. all the axioms and rules of the proof system for MEDL,
2. all the reduction axioms for acts of commanding,
3. all the reduction axioms for acts of promising,
4. all the reduction axioms for acts of requesting, and in addition,
5. the necessitation rules for the dynamic operators $[Com_{(i,j)}\varphi]$, $[Prom_{(i,j)}\varphi]$, and $[Req_{(i,j)}\varphi]$.

Since the above three sets of reduction axioms jointly enable us to define translation function that takes any formula from the language of DMEDL and yields the formula of MEDL that is provably equivalent to the original formula, we can derive the completeness of DMEDL from the completeness of MEDL. Thus we have:

Theorem 4 (The completeness of DMEDL)**.** *The proof system defined above completely axiomatizes* DMEDL*.*

5 Commanding, requesting, and asking questions in DMEDL

In this section, we first review the CUGO principle and the RUGO principle.

The CUGO Principle If φ is a formula of MEDL and is free of modal operators of the form $O_{(j,i,i)}$, $[Com_{(i,j)}\varphi]O_{(j,i,i)}\varphi$ is valid.

The RUGO Principle If φ is a formula of MEDL and is free of modal operators of the form $O_{(j,i,i)}$, $[Req_{(i,j)}\varphi]O_{(j,i,i)}(\varphi \vee K_iO_{(j,i,j)}\varphi \vee K_i\neg O_{(j,i,j)}\varphi)$ is valid.

In the following discussions, we assume that φ is a formula of MEDL and is free of modal operators of the form $O_{(j,i,i)}$, unless stated otherwise.

As we have seen, the CUGO principle is valid while $[Req_{(i,j)}\varphi]O_{(j,i,i)}\varphi$ is not. This fact enables us to understand clearly the sense in which acts of commanding do not allow for the the option of refusal. It becomes obligatory upon the agent j to see to it that φ after the act of commanding of the form $Com_{(i,j)}\varphi$ as the CUGO principle states, but not after the act of requesting of the form $Req_{(i,j)}\varphi$. Moreover, the RUGO principle enables us to understand in what sense the option of refusal is allowed for in the act of requesting of the form $Req_{(i,j)}\varphi$. Seeing to it that $K_i\neg O_{(j,i,j)}\varphi$ is one of the three ways of meeting the obligation of the form $O_{(j,i,i)}(\varphi \vee K_iO_{(j,i,j)}\varphi \vee K_i\neg O_{(j,i,j)}\varphi)$. In that sense refusal is a legitimate response to an act of requesting but not to an act of commanding.

We then move on to acts of asking yes-no questions and examine how the RUGO principle works in modeling them. The notion of question as a kind of imperative or request can be found in various authors including Åqvist (1975), Searle (1979), Hintikka (1981), and Searle and Vanderveken (1985).[12] Our analysis can be applied to the formalization of the notion of questions as requests for information in a straightforward manner. Thus we can define the term that represents the type of the acts in which i asks j whether φ is the case or not, $Ask\text{-}if_{(i,j)}\varphi$, as an abbreviation for $Req_{(i,j)}(K_i\varphi \vee K_i\neg\varphi)$.[13]

Then by the RUGO principle, we have:

$$[Ask\text{-}if_{(i,j)}\varphi]O_{(j,i,i)}((K_i\varphi \vee K_i\neg\varphi) \vee K_iO_{(j,i,j)}(K_i\varphi \vee K_i\neg\varphi) \vee K_i\neg O_{(j,i,j)}(K_i\varphi \vee K_i\neg\varphi)).$$

Here i is the agent who asks the question and j the agent whom the question is asked. We will refer to them as "the requester" and "requestee", and examine how well we can treat the situation after the act of this type as the situation in which information is requested.

[12]There are various approaches to questions. Groenendijk and Stokhof (1997) offers a detailed survey of the field, and argues against Searle and Vanderveken's approach, arguing for the semantic approach to imperative sentences. Since we are not dealing with the semantics of natural language imperative sentences, "a priori there is no clash between" their semantic approach and our analysis as they notes (ibid., p. 1074). For more recent works, see Minică (2011).

[13]The author owes this idea to the discussion with Berislav Žarnić.

Now, after the requester i's act of asking, if the requestee j knows the answer and is willing to answer, she (j) can meet the generated obligation by saying "yes" or "no" immediately, since doing so is to see to it that $(K_i\varphi \vee K_i\neg\varphi)$. Then the requester i will know that φ or know that $\neg\varphi$ accordingly. If the requestee j is willing to answer but needs to consult books, maps, databases, or whatever in order to do so, she (j) can promise the requester i that she (j) will answer it later. Then, as the PUGU principle indicates, the requester i will know that the requestee j has committed herself (j) to letting her (i) know that φ or know that $\neg\varphi$. Thus the requestee j has seen to it that $K_iO_{(j,i,j)}(K_i\varphi \vee K_i\neg\varphi)$. If the requestee j cannot answer or decides not to answer for some reason or other, she (j) has to let the requester i know that she (j) will not commit herself (j) to letting her (i) know the answer. Doing so is to see to it that $K_i\neg O_{(j,i,j)}(K_i\varphi \vee K_i\neg\varphi)$. Thus the RUGO principle captures what j has to do after an yes-no question is asked in a natural way.[14]

6 Requesting and commanding again

So far, we have seen that the CUGO principle and the RUGO principle captures how differently acts of commanding and acts of requesting change situations fairly well. But now observe that the following principle is an instantiation of the CUGO principle:

Proposition 11. *If φ is a formula of* DMEDL *and is free of modal operators of the form $O_{(j,i,i)}$, the following formula is valid:*

$$[Com_{(i,j)}(\varphi \vee K_iO_{(j,i,j)}\varphi \vee K_i\neg O_{(j,i,j)}\varphi)]O_{(j,i,i)}(\varphi \vee K_iO_{(j,i,j)}\varphi \vee K_i\neg O_{(j,i,j)}\varphi).$$

Moreover, we can prove the following result:

Theorem 5. *For each act of requesting, there is an act of commanding with much more complex content which updates models of* DMEDL *in exactly the same way as it does.*

[14] As Grice's discussion of the examinee's answer (Grice 1969, p. 106) suggests, however, this model does not work nicely for questions asked by the examiner in an oral exam. If we combine DMEDL with the dynamic logic of propositional commitments developed in Yamada (2012), we will be able to model such a question as a command to the effect that the commandee should commit herself to the truth or falsity of φ.

Proof. By the definitions of updated models, we have:

$$M_{Req(i,j)\varphi} = M_{Com_{(i,j)}(\varphi \vee K_i O_{(j,i,j)}\varphi \vee K_i \neg O_{(j,i,j)}\varphi)} \ .$$

□

Does this mean that acts of requesting are acts of commanding?

We do not think so. As we have seen, an act of requesting of the form $Req(i, j)\varphi$ and an act of commanding of the form $Com(i, j)\varphi$ change the situation in clearly different ways from each other. The identity of the model updated by the act of requesting of the form $Req(i, j)\varphi$ and the model updated by the act of commanding of the form $Com_{(i,j)}(\varphi \vee K_i O_{(j,i,j)}\varphi \vee K_i \neg O_{(j,i,j)}\varphi)$ just means that it is possible to mimic each act of requesting by an act of commanding which has a related but carefully crafted much more complex content. But even an act of commanding of the form $Com_{(i,j)}(\varphi \vee K_i O_{(j,i,j)}\varphi \vee K_i \neg O_{(j,i,j)}\varphi)$ is different from an act of requesting of the form $Req(i, j)\varphi$ in that seeing to it that $K_i \neg O_{(j,i,j)}\varphi$ is a way of obeying $Com_{(i,j)}(\varphi \vee K_i O_{(j,i,j)}\varphi \vee K_i \neg O_{(j,i,j)}\varphi)$ while it is a way of refusing $Req(i, j)\varphi$.[15]

This consideration, however, reminds us of the following fact:

Observation 1. There are other differences between acts of requesting and acts of commanding, and DMEDL does not deal with them.

This is not surprising. As Sbisà (2001, p. 1792) points out, the use of language in communication is "multi-dimensional $\cdots$, ranging from cognitive to emotional facets, from actional to affective ones, from social to the subjective", and DMEDL is not meant to give a comprehensive account of such a complex phenomenon.

As Searle and Vanderveken (1985, p. 201) point out, for example, an agent who issues a command invokes a position of institutional authority, whereas an agent who makes a request does not. This difference enables us to understand why it is sometimes wise for a person not to issue a command but to make a request even in a situation in which she is in a suitable position of authority over the addressee. Invoking her position of authority overtly can be impolite and offensive.[16] In order to deal with the differences of this kind we need to extend

[15] "Commands" of the form $Com_{(i,j)}(\varphi \vee K_i O_{(j,i,j)}\varphi \vee K_i \neg O_{(j,i,j)}\varphi)$ could be used as a way of pretending that a commander has control of his men, but requests of the form $Req_{(i,j)}\varphi$ could not.

[16] Geis (1995) emphasizes the importance of the matters of "face" in criticizing the standard theory of speech acts.

our language and models. To do so, however, will not amount to abandoning what we have developed but to extending it, and we believe that DMEDL has successfully isolated one important dimension in which the workings of acts of requesting, commanding, and promising are compared.

7 Concluding remarks

Observation 1 seems to require us to further reflect on what are captured by the CUGO principle, the PUGO principle, and the RUGO principle. The existence of other differences DMEDL ignores suggests a possibility that there is a class of illocutionary acts whose members are differentiated from each other only by those differences. According to Searle and Vanderveken (1985, p. 201), the difference between acts of commanding and acts of ordering consists in the fact that the position of power an act of ordering invokes need not be institutionalized while the position of power an act of commanding invokes must be institutionally authorized. This in turn suggests that the characteristic the CUGO principle captures, though stated in reference to acts of commanding, is not specific to acts of commanding but is shared by acts of commanding and acts of ordering. And indeed there seems to be a sub-class of directive acts that share this characteristic, namely the class of directive illocutionary acts that do not allow for the option of refusal. This class seems to include at least telling in the directive sense, requiring, and demanding as well as commanding and ordering. Similarly, the characteristic the RUGO principle captures seems to be shared by acts of asking in the directive sense. Whether there are any commissive acts other than promising that share the characteristic the PUGO principle captures, however, does not seem clear and requires further investigation.

The CUGU principle, the PUGU principle, and the RUGU principle, on the other hand, seem to capture the common characteristic shared by all illocutionary acts in their respective specific forms, namely the necessity of the securing of uptake. The understanding to be secured is often considered as the understanding of the intention of the speaker, but the above principles requires something more objective or public, namely the understanding of the changes brought about in the deontic aspects of the situations.

The way these principles are shown to hold was, however, slightly too easy. As we have seen, we need to model the differences in the (un)certainty of agents as

regards what has happened. Since standard technique of doing this is available, our next step will be to extend DMEDL by introducing the product update.

The above reflection also suggests another interesting possibility of further research. The CUGO principle and other principles of "command logic" in fact enable us to reason at the level of higher generality than that of acts of commanding. We can reason generally about the class of directive acts that do not allow for the option of refusal. Other classes of illocutionary acts, of course, may be studied in this way as well.

Acknowledgements The final version of this paper is published in *European Journal of Analytic Philosophy*, Vol. 7, No. 2, pp. 59–82, 2011. The autohr is grateful to the editors of *European Journal of Analytic Philosophy* for allowing the inclusion of this almost final version in ILLC Yearbook. The earlier versions of this paper were presented in the annual meeting of Japan Association for Philosophy of Science, held in Ehime University, Japan, on June 5, 2011, LIRA (Logic and Interactive Rationality) Seminar, held in Institute for Logic, Language and Computation, University of Amsterdam, on November 3, 2011, and Grolog Lecture, held in Faculty of Philosophy, the University of Groningen, November 17, 2011. I am grateful to the participants of these meetings for their helpful comments and critical discussions. I would also like to thank Johan van Benthem, Frank Veltman, Alexandru Baltag, Barteld Kooi, Sonja Smets, Rosja Mastop, and Berislav Žarnić for their insightful questions and comments. And finally, I thank NWO (Netherlands Organisation for Scientific Research), JSPS (Japan Society for the Promotion of Science), ILLC, University of Amsterdam, and Faculty of Letters, Hokkaido University for their supports that enabled me to spend the first part of my sabbatical in the stimulating and productive atmosphere of ILLC.

References

L. Åqvist. *A New Approach to the Logical Theory of Interrogatives*. Verlag Gunter Narr, Tübingen, 1975.

J. L. Austin. *How to do things with words, William James Lectures*, 1955, harvard university. In J. O. Urmson and M. Sbisà, editors, *How to do things with words*. Harvard University Press, Cambridge, 2nd edition, 1975.

A. Baltag, L. S. Moss, and S. Solecki. The logic of public announcements, common knowledge, and private suspicions. In *Proceedings of the 7th conference on Theoretical aspects of rationality and knowledge*, TARK '98, pages 43–56, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1-55860-563-0. URL `http://portal.acm.org/citation.cfm?id=645876.671885`.

M. L. Geis. *Speech acts and conversational interaction*. Cambridge University Press, Cambridge, 1995.

P. Grice. Utterer's meaning and intentions. *The Philosophical Review*, 78, 1969. Reprinted in: Grice (1989), 86-116. Page references are to this edition.

P. Grice. *Studies in the Ways of Words*. Harvard University Press, Cambridge, Mass., 1989.

J. Groenendijk and M. Stokhof. Questions. In J. van Benthem and A. ter Meulen, editors, *Handbook of Logic and Language*, pages 1055–1124. Elsevier Science B. V., Amsterdam, 1997.

J. Hintikka. On the logic of an interrogative model of scientific inquiry. *Synthese*, 47:69–83, 1981.

Ş. Minică. *Dynamic Logic of Questions*. PhD thesis, University of Amsterdam, 2011. ILLC Dissertation Series DS-2011-08.

M. Sbisà. Illocutionary force and degrees of strength in language use. *Journal of Pragmatics*, 33:1791–1814, 2001.

J. Searle. *Expression and Meaning*. Cambridge University Press, Cambridge, 1979.

J. Searle and D. Vanderveken. *Foundation of Illocutionary Logic*. Cambridge University Press, Cambridge, UK., 1985.

D. Traum. Speech acts for dialogue agents. In M. Wooldridge and A. Rao, editors, *Foundations of Rational Agency*, pages 169–201. Kluwer, 1999.

H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*, volume 337 of *Synthese Library*. Springer, Dordrecht, 2007.

T. Yamada. Acts of commanding and changing obligations. In K. Inoue, K. Sato, and F. Toni, editors, *Computational Logic in Multi-Agent Systems, 7th International Workshop, CLIMA VII, Hakodate, Japan, May 2006, Revised Selected and Invited Papers*, volume 4371 of *Lecture Notes in Artificial Intelligence*, pages 1–19, Berlin / Heidelburg / New York, 2007a. Springer-Verlag.

T. Yamada. Logical dynamics of commands and obligations. In T. Washio, K. Satoh, H. Takeda, and A. Inokuchi, editors, *New Frontiers in Artificial Intelligence, JSAI 2006 Conference and Workshops, Tokyo, Japan, June 2006, Revised*

Selected Papers, volume 4384 of *Lecture Notes in Artificial Intelligence*, pages 133–46, Berlin / Heidelburg / New York, 2007b. Springer-Verlag.

T. Yamada. Acts of promising in dynamified deontic logic. In K. Sato, A. Inokuchi, K. Nagao, and T. Kawamura, editors, *New Frontiers in Artificial Intelligence, JSAI 2007 Conference and Workshops, Miyazaki, Japan, June 18-22, 2007, Revised Selected Papers*, volume 4914 of *Lecture Notes in Artificial Intelligence*, pages 95–108, Berlin / Heidelburg / New York, 2008a. Springer-Verlag.

T. Yamada. Logical dynamics of some speech acts that affect obligations and preferences. *Synthese*, 165:295–315, 2008b.

T. Yamada. Dynamic logic of propositional commitments. In M. Trobok, N. Miščvić, and B. Žarnić, editors, *Between Logic and Reality: Modeling Inference, Action, and Understanding*, pages 183–200. Springer-Verlag, Berlin / Heidelburg / New York, 2012.