Workshop on
Logic and Intelligent Interaction

ESSLLI 2008

- **DATE**: August 11 - 15, 2008 (Week 2)
- **TIME**: 11:00 - 12:30
- **WEBSITE**: `ai.stanford.edu/~epacuit/LaII`
- **ORGANIZERS**: Johan van Benthem & Eric Pacuit

Overview

There is a fast-growing interest in logics that deal with intelligent interaction in communities of agents. There is also a fast-growing jungle of formal systems. This workshop is dedicated to promising recent convergences, trying to foster a common sense of what is going on. The sessions will address five core themes in rational agency:

1. **Dynamic and temporal logics of rational agency**: Connections between temporal frameworks (interpreted systems, epistemic temporal logic) and dynamic epistemic logics.
2. **Merging belief revision and information update**: Connections between AGM-style belief revision theory, inference, and learning theory.
3. **Dynamic logics of preference change and aggregation**: Connections between preference logic, dynamic deontic logic, and multi-agent preference merge.
4. **Logics of games, strategies and actions**: Connections between modal approaches to games and strategies, and proof/category- theoretic approaches.
5. **Logics of collective attitudes and social action**: Connections between dynamic logics, judgment aggregation, and social choice, around themes such as collective agency and deliberation.

History

The workshop is based on several events that have addressed issues in rational agency and intelligent interaction. Examples include established conferences such as *TARK: Theoretical Aspects of Rationality and Knowledge* and *LOFT: Logic and the Foundations of Game and Decision Theory*. In addition, a number of recent events have focused on the specific themes of the present ESSLLI workshop:

- NIAS Research Group on Games, Action and Social Software (2006 - 2007)
- Dynamic Logic, Montreal (2007)
- Logic and Rational Interaction (LORI, Beijing 2007)
- Decisions, Games and Logic Workshops (London, 2007; Amsterdam, 2008)

There is also a new European Science Foundation Eurocores Program (LogiCCC) devoted to logic in the broad interdisciplinary sense that we are pursuing here. For more information about this active and growing area of research see

`www.illc.uva.nl/wordpress.`

Workshop Format

The workshop will take place August 11 - 15 during the second week of ESSLLI Hamburg 2008. Each contribution will be given 15 minutes for a short presentation with 5 minutes for discussion. In addition, there is some scheduled time for group discussion about our general themes. A tentative schedule is given below (the final schedule will be posted a few weeks before the workshop on the conference website).

Monday, August 11	
11:00 - 11:10	Opening Remarks
11:10 - 11:50	Semantics, Information and Learning
11:50 - 12:30	Inference and Information
Tuesday, August 12	
11:00 - 11:40	Dynamic Probabilistic Modal Reasoning
11:40 - 12:20	Logics of Belief Change over Time
12:20 - 12:30	Short Discussion
Wednesday, August 13	
11:00 - 11:50	Invited Speaker: R. Ramanujam (Chennai)
11:50 - 12:30	Reasoning about and with Games
Thursday, August 14	
11:00 - 12:20	Logics of Action and Intention
12:20 - 12:30	Short Discussion
Friday, August 15	
11:00 - 11:50	Knowledge in Changing Environments
11:50 - 12:30	Discussion

Invited Lecture: Professor R. Ramanujam will give an invited lecture *Some Automata Theory for Epistemic Logics*.

Abstract. We consider epistemic temporal logics and dynamic epistemic logics from the viewpoint of automata theory. Specifically, what can we say about the knowledge of finite state agents as regular collections of behaviours? We show that there are interesting notions awaiting study relating to monadic second order theories and Kleene type theorems for what we term epistemic languages. We provide answers for some special epistemic theories.

Proceedings

A special issue of the journal Knowledge, Rationality, and Action, will be dedicated to this workshop. Presenters at the workshop will be invited to submit an extended version of their submission to the editors T. Agotnes, J. van benthem & E. Pacuit. There will be an independent reviewing process. More information (including deadlines) will be provided during the workshop.

Contents

Semantics, Information and Learning

Inference and Information

Dynamic Probabilistic Modal Reasoning

Logics of Belief Change over Time

Reasoning about and with Games

Contents

Logics of Action and Intention

Knowledge in Changing Environments

Knowing whether A or B

Maria Aloni
ILLC, Amsterdam
M.D.Aloni@uva.nl

Paul Égré
Institut Jean-Nicod, Paris[*]
paulegre@gmail.com

Abstract

Can we say that s knows whether A or B when s is only able to rule out A, but remains uncertain about B? We discuss a set of examples put forward by J. Schaffer's in favour of a contextualist answer to this problem. We present a context-sensitive and dynamic semantics for knowledge attributions, in which those can depend on the alternatives raised by the embedded question, but also on alternatives raised earlier in the context.

1. Alternative questions in epistemic contexts

The aim of this paper is to discuss the semantics of knowledge attributions of the form "s knows whether A or B", which we may symbolize by $K_s?(A \vee_a B)$, where $?(A \vee_a B)$ denotes an alternative disjunctive question, like "is John in London, or is Mary in London?". More specifically, our aim is to provide a dynamic account of the context-sensitivity of such attributions.

It is standard in linguistic theory to distinguish polar readings and alternative readings of disjunctive questions (see e.g. Haspelmath 2000, Han and Romero 2003). Under the polar reading, a question of the form "is John or Mary in London?" calls for a yes or no answer. The polar reading can be forced in English by asking "is either John or Mary in London?". For the alternative reading, by contrast, the question cannot be answered by yes or no and has to be answered by a sentence like "John is London", or "Mary is not in London", namely by providing information about the truth and falsity of the respective disjuncts.

There is still some debate in the literature about the answerhood conditions of alternative questions, and by way of consequence, about the conditions under which a subject can be said to know whether A or B. In a recent paper (Schaffer 2007), J. Schaffer argues that in a context in which s sees someone on TV, who is actually George Bush, but such that s is not able to discriminate between George Bush and Will Ferrell (because Ferrell is such a good impersonator of Bush), and yet is able to see that it is not Janet Jackson, (1-a) below should be judged false, but (1-b) should count as true:

(1) a. s knows whether George Bush or Will Ferrell is on TV
b. s knows whether George Bush or Janet Jackson is on TV.

The intuition reason for the truth of (1-b), according to Schaffer, is that the question "is Bush or Janet Jackson on TV?" is easier for s to answer than the question "is Bush or Will Ferrell on TV?". In our view, however, ordinary intuitions are less stable: although (1-a) should be incontrovertibly false in the scenario, the status of (1-b) is much less clear. In our opinion, all that s really knows is that *Janet Jackson is not on TV*, which need not be sufficient to fully answer the question "is Bush or Janet Jackson on TV?".

More formally, assuming the partition theory of questions of Groenendijk and Stokhof (1984), an answer of the form "Janet Jackson is not on TV" counts only as a *partial answer* to the question "is Bush or Janet Jackson on TV?". For s to know the complete answer to the question "is Bush or Janet Jackson on TV", s should know more, namely that Bush is on TV and that Janet Jackson is not on TV. The partial answer "Janet Jackson is not on TV" would count as complete if one presupposed that exactly one of the two disjuncts had to be true. In principle, however, there is no more reason to think that "s knows whether Bush or Janet Jackson is on TV" is true than there is to think that "s knows whether Ferrell or Janet Jackson is on TV" is true. In other words, s's ignorance about who exactly *is* on TV seems to override s's partial knowledge about who is *not* on TV.

Despite this, we agree with Schaffer that there is a sense in which, if s is allowed to *ignore* the possibility that Ferrell might be on TV, then s can be said to know whether Bush or Janet Jackson is on TV, simply based on s's knowledge of that partial answer.

2. Dynamics of knowledge attributions

To implement this idea, we propose a question semantics for knowledge in which attributions involving questions can be made sensitive both to the alternatives raised by the question, as well as to alternatives raised earlier in the context. The semantics is dynamic, in so far as the context can be incremented with the considerations of new alternatives, in a way that not simply restricts, but can also increase, the subject's uncertainty.

2.1. Question semantics

Questions in the system are represented by formulas of the form $?p_1, ..., p_n\ \phi$ where ? is a query-operator, $p_1, ..., p_n$ is a possibly empty sequence of propositional variables, and ϕ is a formula of predicate logic with propositional variables. In the case of alternative questions, a question of the form "is ϕ or ψ?" (abbreviated $?(\phi \vee_a \psi)$) is represented by a formula of the form $?p(p \wedge (p = \phi \vee p = \psi))$, which asks which of the propositions ϕ and ψ is true.

Questions denotations are then defined as follows, where $\vec{p}$ stands for the sequence $p_1, ..., p_n$, and $\vec{\alpha}$ for the sequence $\alpha_1, ..., \alpha_n$: $\llbracket ?\vec{p}\ \phi \rrbracket_{M,g} = \{\langle \vec{\alpha}, w \rangle \mid w \in \llbracket \phi \rrbracket_{M,g[\vec{p}/\vec{\alpha}]}\}$. The denotation of an alternative question $?p(p \wedge (p = \phi \vee p = \psi))$ is thus the set of pairs $\langle p, w \rangle$ such that w satisfies p and p is either the proposition expressed by ϕ or the proposition expressed by ψ. From the denotation of a question, we can define the *partition* Part($?\vec{p}\ \phi$) induced by the question $?\vec{p}\,\phi$ as the set of ordered pairs $\langle w, v \rangle$ such that for all proposition $\vec{\alpha}$, $\langle \vec{\alpha}, w \rangle \in \llbracket ?\vec{p}\ \phi \rrbracket_{M,g}$ iff $\langle \vec{\alpha}, v \rangle \in \llbracket ?\vec{p}\ \phi \rrbracket_{M,g}$. Finally, we define the *topics* raised by a question as the set $\text{Top}_{M,g}(?\vec{p}\,\phi) = \{\vec{\alpha} \mid \exists w : \langle \vec{\alpha}, w \rangle \in \llbracket ?\vec{p}\ \phi \rrbracket_{M,g}\}$. For alternative questions, one can check that Part($?(\phi \vee_a \psi)$)=$\{\phi \wedge \neg\psi, \neg\phi \wedge \psi, \neg\phi \wedge \neg\psi, \phi \wedge \psi\}$, and Top($?(\phi \vee_a \psi)$)=$\{\phi, \psi\}$.

2.2. Knowledge and context updates

A *context* C is defined as an ordered pair whose first index s_C is an information state (set of worlds), and whose second index i_C is a sequence of question denotations representing the issues under discussion in C. A context C can be updated either by an assertion P, or by the introduction of a new question Q:

(2) a. $C + P = (s_C \cap \llbracket P \rrbracket, i_C)$
b. $C + Q = (s_C, i_C + \llbracket Q \rrbracket)$

We let $\text{ANS}_w(Q)$ be the true exhaustive answer to Q in w (the cell containing w in Part(Q), and Top(C) denote the union of the topics introduced by all the issues in C, i.e. for $C = (s_c, \llbracket Q_1 \rrbracket, ..., \llbracket Q_n \rrbracket)$: Top($C$) = $\bigcup_{i \in n}$Top(Q_i) $\setminus \{\langle\rangle\}$.

Define $\mathcal{K}_s(w)$ to be the knowledge state of s in w, namely the set of epistemically accessible worlds to s. We then define knowledge as follows:

(3) "s knows Q" is true in world w with respect to context C iff
(i) $\mathcal{K}_s(w) \cap \text{Top}(C) \subseteq \text{ANS}_w(Q)$, if $\text{Top}(C) \neq \emptyset$;
(ii) $\mathcal{K}_s(w) \subseteq \text{ANS}_w(Q)$, otherwise.

3. Schaffer's puzzle

Going back to Schaffer's example, suppose $\mathcal{K}_s(w)$ is a state compatible with Bush being on TV (B) and with Ferrell being on TV (F), but excluding Janet Jackson being on TV (J). The following holds:

(4) a. S knows whether it is Bush or Janet Jackson on TV.
b. true in C+$?(B \vee_a J)$, but false in C+$?(B \vee_a J)$+$?(B \vee_a F)$

(5) a. S knows whether it is Bush or Ferrell on TV.
b. false in C+$?(B \vee_a F)$, and likewise false in C+$?(B \vee_a F)$+$?(B \vee_a J)$.

The semantics predicts that when s's knowledge state is restricted to the topics raised by "is Janet Jackson or Bush on TV?", s will know the answer. But if a further issue comes up after this question was asked, namely "is Bush or Ferrell on TV?", then s may no longer be said to know whether Bush or Janet Jackson is on TV, because the context is incremented with a third alternative (namely the possibility that it might be Ferrell).

4. Perspectives

The semantics here presented can be used to deal with other scenarios involving, in particular, the consideration of skeptical alternatives, whereby the introduction of a new alternative can impair one's initial confidence in the particular answer to a question. We shall explain the extension of the semantics to other types of questions, and discuss possible connections with the topic of unawareness. A further issue, which we elaborate in the paper, concerns the partialization of the semantics, to deal with presupposition failure. Thus, in a situation in which s holds a partial answer to the question, as in Schaffer's scenario, the negation of (1-b) may be judged inappropriate, hence neither true nor false, rather than true at all. The partiality can be derived from the assumption that s's uncertainty should always be symmetric with respect to the alternatives raised by the question.

References

[1] Aloni M. & Égré P. (2008), "Alternative Questions and Knowledge Attributions", manuscript, under review.

[2] Groenendijk, J. and Stokhof M. (1984), *Studies in the Semantics of Questions and the Pragmatics of Answers*. PhD dissertation, University of Amsterdam.

[3] Haspelmath, M. (2000), “Coordination”. To appear in: T. Shopen (ed.) *Language typology and syntactic description*. 2nd ed. Cambridge: Cambridge University Press.

[4] Romero, M. and Han C-H. (2003), “Focus, Ellipsis and the Semantics of Alternative Questions”, in *Empirical Issues in Formal Syntax and Semantics 4*. C. Beyssade, O. Bonami, P. Cabredo Hofherr, F. Corblin (eds), Presses Universitaires de Paris-Sorbonne, Paris, 291-307.

[5] Schaffer, J. (2007), “Knowing the Answer”, *Philosophy and Phenomenological Research*, Vol. LXXV No. 2, 1-21.

Identification through Inductive Verification

Application to Monotone Quantifiers

Nina Gierasimczuk[*]
Institute for Logic, Language, and Computation, University of Amsterdam
Institute of Philosophy, University of Warsaw
nina.gierasimczuk@gmail.com

Abstract

In this paper we are concerned with some general properties of scientific hypotheses. We investigate the relationship between the situation when the task is to verify a given hypothesis, and when a scientist has to pick a correct hypothesis from an arbitrary class of alternatives. Both these procedures are based on induction. We understand hypotheses as generalized quantifiers of types $\langle 1\rangle$ or $\langle 1, 1\rangle$. Some of their formal features, like monotonicity, appear to be of great relevance. We first focus on monotonicity, extendability and persistence of quantifiers. They are investigated in context of epistemological verifiability of scientific hypotheses. In the second part we show that some of these properties imply learnability. As a result two strong paradigms are joined: the paradigm of computational epistemology (see e.g. [6, 5]), which goes back to the notion of identification in the limit as formulated in [4], and the paradigm of investigating natural language determiners in terms of generalized quantifiers in finite models (see e.g.[10]).

Keywords: *identification in the limit, induction, monadic quantifiers, monotonicity, semantics learning, verification.*

1 Introduction

The ‘identification in the limit’ model [4] has found numerous applications in language learning analysis — for the most part in the acquisition of syntax. In contrast the model has been unappreciated in the investigations concerning learning of semantics.

On the other hand, in philosophy of science Gold’s paradigm has been used to account for inductive reasoning and the process of approaching the correct theory about the world. In this domain various semantic properties of hypotheses are of great importance [6, 1].

In the present paper we abstract from the distinction between learning and scientific inquiry. We hope that with this generality our results are relevant for both subjects. Our aim is to analyze semantic properties of inductive verifiability [6] and consider its connection with identification. The first section is devoted to two kinds of verifiability. The introduction of those notions is illustrated with the example of verifiability of monadic quantifiers in section 2. Next we present the basics about identification in the limit. In the culminating chapter 3 we compare the two notions. We conclude with theorems showing that with some restrictions certain types of verification imply identification.

2 Verification

The idea of verification, except for its obvious connections with semantics, is also very important in philosophy of science, where verifying and falsifying seem to be fundamental procedures for establishing an adequate theory and making predictions about the actual world. The semantic procedure of verification consists essentially in what follows:

Verification task Given model M and a sentence φ, answer the question whether $M \models \varphi$.

Let us start with analyzing restrictions we should make on the verification task to be able to proceed with our considerations.

First of all, for the sake of generality we consider M to be infinite. This allows us to talk about infinite procedures being successful in the limit. It is also very important to restrict our attention to computably enumerable structures. The reason is that we are interested in elements of the model being presented one by one — such an inductive procedure essentially requires that it is possible to enumerate them. In connection with this we also require that a presentation of a given model does not include repetitions. This restriction is made to simplify the procedure of counting elements without introducing any additional markers. We also have to say

[*] The author is the receiver of a Foundation for Polish Science Award for Young Researchers (START Programme 2008).

something about φ — the sentence involved in the verification task. We assume that φ has the form of a quantifier sentence, with a quantifier closed under isomorphism. In other words, we assume that hypotheses of our framework are purely about cardinalities or relations between cardinalities, and not about the 'nature' of individual objects.

With the above-explained restrictions in mind, let us now move to define a formal framework of inductive verifiability.

Definition 2.1 Let us consider a model $M = (U, B)$, where U is an infinite, computably enumerable set, and $B \subseteq U$ is some computable unary predicate. Let us assume that λ is an enumeration of the elements of U, without repetitions.

By 'environment of M', ε, we mean an infinite binary sequence such that: if $\lambda_n = x$, then $\varepsilon_n = \chi_B(x)$, where χ_B is the characteristic function of B. ◁

We will use the following notation:

$\varepsilon|n$ is the finite initial segment of ε through position $n-1$ (i.e.: a sequence $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{n-1}$);

SEQ denotes a set of all finite initial segments of all environments;

$set(\varepsilon)$ is a set of elements that occur in ε;

h will refer to a hypothesis;

C is a correctness relation between hypotheses and streams of data. $C(\varepsilon, h)$ is satisfied iff h is correct with respect to ε, i.e., h is true in the model represented by ε;

α is an assessment method — total map from hypotheses and finite data sequences to conjectures, $\alpha : H \times SEQ \to \{0, 1, !\}$.

Conjectures are outputs of α; their meaning is the following:

1 — corresponds to the judgement that the hypothesis is true on the initial "up to now" segment of data;

0 — means that the hypothesis is judged to be false on the initial "up to now" segment of data;

! — appears as an announcement that there will be now mind change about the statement following in the next step (we also refer to it as the *eureka* sign).

2.1 Verification with Certainty

The first type of verification we want to discuss is verification with certainty. It holds when the process of verification is finished after a finite number of steps. We mean 'finished' in the sense that there is a point in the procedure at which the assessment method, α, *decides* that the hypothesis, h, is true and that it can stop computing right there, because h being false is no longer an option. In such a case we can informally say that α is 'sure' or 'certain' about the answer. This is where the name 'verification with certainty' comes from.

Formally, we will require that the step when certainty comes into the picture is marked with the *eureka* symbol '!' and the actual content of this certainty — the hypothesis being true or false — is '1' or '0', respectively, answered in the next step.

Let us first introduce the general notion of producing an answer with certainty.

Definition 2.2 We say that α produces b with certainty on (h, ε) iff there is an n such that:

1. $\alpha(h, \varepsilon|n) = !$, and
2. $\alpha(h, \varepsilon|n+1) = b$,
3. for each $m < n$, $\alpha(h, \varepsilon|m) \neq !$, and
4. all values after $n+1$ are irrelevant.

◁

Verification and falsification with certainty are defined as an adequate production of 0 or 1 with certainty, respectively.

Definition 2.3 We say that α verifies h with certainty on ε (with respect to C) iff α produces 1 with certainty on $(h, \varepsilon) \Leftrightarrow C(\varepsilon, h)$. Definition of refutation with certainty is analogous. ◁

Definition 2.4 We say that h is verifiable with certainty iff there is an α, which for each ε verifies h on ε with certainty iff h is true on ε. ◁

Verification with certainty satisfies the condition of positive introspection of knowledge, i.e., as soon as α answers '!' on h, it 'knows' the logical value of h. Such a situation does not occur in verification in the limit, which is defined below.

2.2 Verification in the Limit

Verification in the limit is much weaker than verification with certainty. In order to define it we exclude the *eureka* sign '!' from the set of possible answers. We restrict the power of the verification procedure α in such a way that it can give only two answers:

1 — corresponds to the fact that the hypothesis is judged to be true on the initial "up to now" segment of data;

0 — the hypothesis is judged to be false on the initial "up to now" segment of data.

As in the previous case, this type of verification consists in giving partial answers to finite initial segments of the environment. This time however the procedure is endless. We are dealing here with an infinite sequence of answers. We say that a procedure verifies a hypothesis in the limit if and only if there is a step in the procedure where the answer is 1 and it stays that way for the rest of the computation.

Definition 2.5 We say that α verifies a hypothesis, h, in the limit iff:

$$\exists n \forall m > n \; \alpha(h, \varepsilon | m) = 1.$$

$\lhd$

Definition 2.6 We say that h is verifiable in the limit iff there is an α, which for each ε verifies h in the limit on ε iff h is true on ε. $\lhd$

In the general case of verification in the limit the fact of verification is not 'visible' to α. Whether a hypothesis has been verified can be judged only from a global perspective. Limiting verification corresponds to the scientific strategy of claiming adequacy of some 'up to now' correct hypothesis as long as possible. There is no guarantee however that in the light of future data it will not be rejected. When dealing with verifiability in the limit a scientist has to remain alert all the time.

3 Application: Verification of Monotone Quantifiers

The restriction made in the previous section, that hypotheses of our framework are purely about cardinalities or relations between cardinalities, and not about the 'nature' of individual objects leads us to treat hypotheses as generalized quantifiers. Informally speaking a given hypothesis can be identified with the class of models in which it is true. The same works for quantifiers. Even if intuitively quantifiers are formal counterparts of (natural language) determiners, we have a theory of generalized quantifiers which instructs us to reduce a quantifier simply to the class of models in which this quantifier is true. So, running the risk of being charged with philosophical insensitivity, we will use the notions of quantifiers and hypotheses interchangeably.

In order to talk about the properties we are interested in we have to provide the relational definition of generalized quantifier.

Definition 3.1 A generalized quantifier Q of type $t = (n_1, \ldots, n_k)$ is a functor assigning to every set M a k-ary relation $\mathsf{Q_M}$ between relations on M such that if $(R_1, \ldots, R_k) \in \mathsf{Q_M}$ then R_i is an n_i-ary relation on M, for $i = 1, \ldots, k$.

$\lhd$

It is quite prevalent in the philosophical literature to link notions of verifiability (with certainty) and falsifiability (with certainty) to the existential and universal quantifier, respectively. In fact, as we are going to see, this intuitive correspondence includes a broader class with quantifiers of some special monotonicity properties. We will discuss this connection below.

3.1 Quantifiers of Type $\langle 1 \rangle$

Let us now focus on properties of generalized quantifiers of type $\langle 1 \rangle$. First we define what it means for a quantifier to be monotone increasing and extendable.

Definition 3.2

(MON$\uparrow$) We say that a quantifier $\mathsf{Q_M}$ of type $\langle 1 \rangle$ is monotone increasing (MON$\uparrow$) iff the following holds: if $A \subseteq A' \subseteq M$, then $\mathsf{Q_M}(A)$ implies $\mathsf{Q_M}(A')$.

(EXT) A quantifier Q of type $\langle 1 \rangle$ satisfies EXT iff for all models M and M': $A \subseteq M \subseteq M'$ implies $\mathsf{Q_M}(A) \implies \mathsf{Q_{M'}}(A)$.

$\lhd$

In other words, monotonicity guarantees that extending the predicate does not change the logical value of the quantifier from true to false. On the other hand extension ensures that adding new elements to the complement of A does not make a true quantifier false.

Comparison of the notions of verifiability with certainty and monotonicity allows us to state the following proposition:

Proposition 3.3 *Let* Q *be a* MON$\uparrow$ *and* EXT *quantifier of type* $\langle 1 \rangle$. *There exists a model* $M = (U, A)$ *with finite* $A \subseteq U$ *such that* $\mathsf{Q}_M(A)$ *iff* Q *is verifiable with certainty for arbitrary computably enumerable models.*

Proof. ($\Rightarrow$) Let us first assume that Q of type $\langle 1\rangle$ is MON$\uparrow$ and EXT, and that there exists a model $M = (U, A)$ with finite $A \subseteq U$ such that $\mathsf{Q}_M(A)$. We use the characteristic function of A, χ_A, to get an infinite sequence, ε_A, of 0's and 1's representing M. ε_A is an environment of M. We run the α procedure on ε_A and $\mathsf{Q}(A)$. Step by step, while being fed, α constructs a model $M' = (U', A')$. This happens in the following way.

First we take $n := 0, U' := \emptyset, A' := \emptyset$.

α reads ε_n: if $\varepsilon_n = 1$, then $|A'| := |A'| + 1$; else $|\bar{A}'| := |\bar{A}'| + 1$. α checks if $\mathsf{Q}(A')$: if it holds, α answers '!' and 1 to the rest of ε_A; otherwise it answers 0 and moves to $n := n + 1$.

The procedure α verifies $\mathsf{Q}(A)$ with certainty. This is because $\mathsf{Q}(A)$ is true in M, and from assumptions about MON $\uparrow$ and EXT, we know there is a finite cardinality of A' which satisfies $\mathsf{Q}(A')$. As soon as α reaches this cardinality there is no possibility that $\mathsf{Q}(A)$ changes its logical value at an extension A', $\bar{A}'$ in M'.

($\Leftarrow$) Let us assume that $M \models \mathsf{Q}(A)$, and that there is a procedure α which verifies with certainty on ε_A. Therefore, there is a point, n, at which α answers ! and then 1. Then we know that $\mathsf{Q}(A')$, where $|A'|$ is equal to the number of 1s in $\varepsilon_A|n$ and $|\bar{A}'|$ is equal to the number of 0s in $\varepsilon_A|n$. What remains of ε is not relevant for the logical value of $\mathsf{Q}(A')$. This means that if $A' \subseteq A''$ then $\mathsf{Q}(A'')$ and if $M' \subseteq M''$ then $Q_{M''}(A')$. This is the same as saying that Q is MON$\uparrow$ and EXT. QED

Having this in mind we can also consider which type $\langle 1\rangle$ quantifiers correspond to the notion of falsifiability with certainty. The answer is as follows:

Proposition 3.4 *Let* Q *be a quantifier of type* $\langle 1\rangle$. Q *is verifiable with certainty iff* $\neg\mathsf{Q}$ *is falsifiable with certainty.*

Proof. ($\Rightarrow$) First assume that Q is verifiable with certainty. That is: there is a procedure α such that for every model M if $M \models \mathsf{Q}(A)$, then α verifies $\mathsf{Q}(A)$ with certainty. We now construct a procedure α' such that it falsifies $\neg\mathsf{Q}$ with certainty.

$$\alpha'(\varepsilon_A|n) = \begin{cases} 1 & \text{if } \alpha(\varepsilon_A|n) = 0, \\ 0 & \text{if } \alpha(\varepsilon_A|n) = 1, \\ ! & \text{if } \alpha(\varepsilon_A|n) = \ !. \end{cases}$$

Since $\neg\mathsf{Q}$ is a complement of Q, this procedure falsifies $\neg\mathsf{Q}$ on A iff $\neg\mathsf{Q}$ is false in M. ($\Leftarrow$) The other direction works the same way. QED

3.2 Quantifiers of Type $\langle 1, 1\rangle$

In the linguistic context it is common to investigate quantifiers of type $\langle 1, 1\rangle$. It is often assumed (see e.g. [9]) that all natural language determiners correspond to so-called CE-quantifiers. CE-quantifiers satisfy three requirements: isomorphism closure (ISOM), extension and conservativity (CONS). (EXT) for quantifiers of type $\langle 1, 1\rangle$ is a natural extension of the definition for type $\langle 1\rangle$. Below we define (CONS).

Definition 3.5 We call a quantifier Q of type $\langle 1, 1\rangle$ conservative iff:

(CONS) $\forall A, B \subseteq M$: $\mathsf{Q}_M(A, B) \iff Q_M(A, A \cap B)$.

$\triangleleft$

CE-quantifiers then have the property that their logical value depends only on the cardinality of the two constituents, $A - B$ and $A \cap B$, in the model. The part of B falling outside of the scope of A does not influence the logical value of a CE-quantifier. For the rest of the present section we will restrict ourselves to CE-quantifiers.

We will also need a notion of left-side monotonicity, which is usually called 'persistence'.

Definition 3.6 We call a quantifier Q of type $\langle 1, 1\rangle$ persistent iff:

(PER) If $A \subseteq A' \subseteq M$ and $B \subseteq M$, then $\mathsf{Q}_{\mathsf{M}}(A, B) \Rightarrow \mathsf{Q}_{\mathsf{M}}(A', B)$.

$\triangleleft$

Persistence guarantees that adding new elements to both important constituents A and $A \cap B$ does not change the logical value of the quantifier from true to false.

We claim the following:

Proposition 3.7 *Let* Q *be a* PER CE*-quantifier of type* $\langle 1, 1\rangle$. *There exists a model* $M = (U, A, B)$ *such that* $A \cap B$ *is finite and* $\mathsf{Q}_M(A, B)$ *iff it is verifiable with certainty.*

Proof. The proof is analogous to the proof of Proposition 1. We simply focus on two constituents of the model: $A - B$ and $A \cap B$, and treat them as $\bar{A}$ and A (respectively) in the proof of Proposition 1. QED

Proposition 3.8 *Let* Q *be a* CE*-quantifier of type* $\langle 1, 1\rangle$. $\neg\mathsf{Q}$ *is falsifiable with certainty iff* Q *is verifiable with certainty.*

Proof. Analogous to the the proof of Proposition 2. QED

Monotonicity has so far given some explanation for differences in the comprehension of quantifiers. The distinction between verifiability and refutability of quantifiers provides new thoughts regarding this problem. It gives some additional psychologically plausible explanation for differences in the difficulty of natural language quantifiers. It can be argued that verification of hypotheses is much easier for people than refutation. In consequence verifiable quantifiers can be considered much easier in natural language processing than refutable ones.

4 Identifiability through Verification

Historically speaking, philosophical analysis of the scientific discovery process led to skepticism. It has been claimed that its creative content cannot be accounted for by any scientific means, in particular by no mathematical or algorithmic model [2]. The natural situation of discovery is indeed so complex and non-uniform that it seems impossible to catch it in an adequate formalism. However, some approximations, which to a certain extent idealize the process, are not only makable, but also already existing and are ready to use. The framework of identification in the limit proposed in [4] started a long line of mathematical investigation of the process of language learning. At first sight scientific discovery and learning might seem distant from each other. In the present paper we assume the adequacy of the identification model for scientific inquiry analysis (for similar approaches see: [6, 7]).

Intuitively, the verification procedure (discussed in the previous section) is a part of scientific discovery. The latter can be seen as a compilation of assuming hypotheses, checking their logical value on data, and changing them to another hypothesis, if needed. In the present section we will introduce the identification formalism and present some ideas and facts about its correspondence to verification.

4.1 Identification

The identification in the limit approach [4] gives a mathematical reconstruction of the process of inductive inference. The task consists in guessing a correct hypothesis on the basis of an inductively given, infinite sequence of data about the world.

The framework includes: a class of hypotheses H, an infinite sequence of data about the world ε, a learning function f (a scientist).

We will explain the general idea of identification in the limit in terms of a simple game between a Scientist and Nature. First, some class of hypotheses, H, is chosen. It is known by both players. Then Nature chooses a single hypothesis, h, from H, to correctly describe the actual world. Then Nature starts giving out atomic information about the world. She does this in an inductive way. Each time the Scientist gets a piece of information, he guesses a hypothesis from the previously defined class on the basis of the sequence of data given so far. Identification in the limit is successful, if the guesses of the Scientist after some finite time stabilize on the correct answer.

Let us now specify the elements of the framework. By hypotheses we again mean quantified formulae, with a logical (closed under isomorphism) quantifier of type $\langle 1 \rangle$ or CE-quantifier of type $\langle 1, 1 \rangle$ (see e.g. [9]). The reason for this is the same as in the case of verification — that we want order- and intension-independent hypotheses, and a clear and relevant binary representation of models. The above-mentioned encoding of models serves as a basis for environments. The learning function, also referred to as the 'scientist', is defined as $f : SEQ \to H$.

Definition 4.1 [Identification in the limit]
We say that a learning function, f:

1. identifies $h \in H$ on ε for $M \models h$ in the limit iff for cofinitely many n, $f(\varepsilon|n) = h$.

2. identifies $h \in H$ in the limit iff it identifies h in the limit on every ε for every M, such that $M \models h$.

3. identifies H in the limit iff it identifies in the limit every $h \in H$.

$\triangleleft$

We can analogously define the much stronger notion of identifiability with certainty. The difference is that in this case the learning function 'knows' when it has identified the correct hypothesis.

Definition 4.2 [Identification with certainty]
We say that a learning function, f:

1. identifies $h \in H$ with certainty on ε for $M \models h$ iff for some n, $f(\varepsilon|n) =!$ and $f(\varepsilon|n + 1) = h$.

2. identifies $h \in H$ with certainty iff it identifies h with certainty on every ε for every $M \models h$.

3. identifies H with certainty iff it identifies with certainty every $h \in H$.

$\triangleleft$

4.2 Comparing verification and identification

In the present section we will state two theorems. They show a connection between identifiability and verifiability.

4.2.1 Certainty setting

Let us take a class of hypotheses, H, and the sequence, ε, of data about the actual world. Assume that H contains only mutually disjoint hypotheses verifiable with certainty, i.e., for every $h \in H$ there is a procedure α, which verifies h with certainty iff it is true in the actual world.

Theorem 4.3 *Every such computably enumerable class H is identifiable with certainty.*

Proof. Assume that H is a computably enumerable class of mutually disjoint hypotheses verifiable with certainty. We define a procedure **Id-Cert** which identifies with certainty every hypothesis from the class H. An example of a run of the procedure is presented in Figure 1.

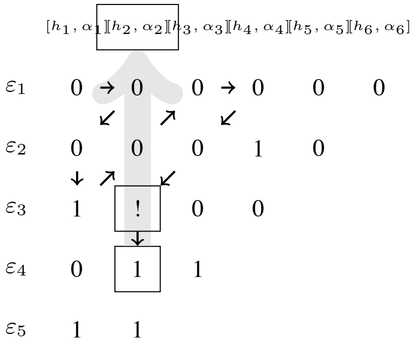


Figure 1. Identifiability with certainty

Since H is computably enumerable we can assume existence of a sequence $(h)_n$ which enumerates H. Each h_n is associated with its verification with certainty procedure α_n. **Id-Cert** works in the following way: it first checks $\alpha_1(h_1, \varepsilon_1)$ (the value of the first hypothesis on the first piece of data), then it proceeds according to the diagonal enumeration of $\alpha_n(h_n, \varepsilon_m)$ until it meets '!'. Then it performs a check for $\alpha_n(h_n, \varepsilon_{m+1})$. If $\alpha_n(h_n, \varepsilon_{m+1}) = 1$, then **Id-Cert** stops and answers h_n. Otherwise it moves back to $\alpha_n(h_n, \varepsilon_m)$ and continues to perform the diagonal procedure.

By assumption every $h \in H$ is verifiable with certainty. Therefore if h_n, for some n, is true on ε, then α_n will eventually produce '!'. And since **Id-Cert** performs a diagonal search it does not miss any answer. Hence, **Id-Cert** identifies every $h \in H$ with certainty, so H is identifiable with certainty. QED

Let us again take a class of hypotheses, H, and the sequence, ε, of data about the actual world. Assume that H contains only hypotheses verifiable with certainty, but this time let us drop the assumption of H being a class of mutually disjoint hypotheses. Then we can prove what follows.

Theorem 4.4 *Every such computably enumerable class H is identifiable in the limit.*

Proof. The proof is very similar to the proof of the previous theorem. We use the same diagonal method. This time however identification does not stop on the first '!' it encounters. Let us assume that '!' happens for ε_n. Instead, it answers the relevant h: the hypothesis which was first recognized to be verified with certainty; then it goes on with the diagonal search looking for a hypothesis, h', which reveals '!' for some ε_m, where $m < n$. If it meets such an h' it keeps answering it as long as no other 'better fitting' hypothesis is found. An example of a run of the procedure is presented in Figure 2.

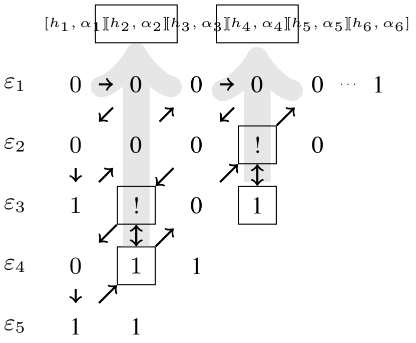


Figure 2. Identifiability with certainty

By assumption every $h \in H$ is verifiable with certainty. Therefore if h_n, for some n, is true on ε, then α_n will eventually produce '!'. And since this identification performs a diagonal search it does not miss any answer. Hence every $h \in H$ is identified in the limit, so H is identifiable in the limit. QED

4.2.2 Limiting setting

Let us again take a computably enumerable class of mutually disjoint hypotheses, H, and a sequence, ε, of data about the actual world. But this time let us agree that H consists of hypotheses that are verifiable in the limit, i.e., for every $h \in H$ there is a procedure α which verifies h in the limit iff h it is true.

Theorem 4.5 *Every such computably enumerable class H is identifiable in the limit.*

Proof. Assume that H is a computably enumerable class of mutually disjoint hypotheses that are verifiable in the limit. This means that for every $h_n \in H$ there is a procedure α_n which verifies h in the limit if and only if h is true. We are now going to define a procedure **Id-Lim** which identifies every hypothesis from the class H. An example of a run of the **Id-Lim** is presented in Figure 3.

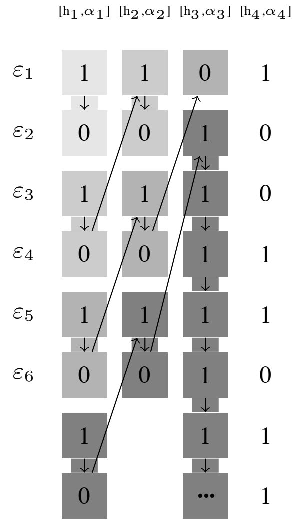


Figure 3. Id-Lim identifiability

Since H is computably enumerable we can assume the existence of the sequence $(h)_n$ enumerating the hypotheses from H. Each of them is associated with its verification in the limit procedure α_n.

The algorithm **Id-Lim** first performs a single check for $\{h_1\}$:

If $\alpha_1(h_1, \varepsilon|1) = 1$, then **Id-Lim** outputs h_1 and moves to $\alpha_1(h_1, \varepsilon|2)$. The answer is repeated until there is an n such that $\alpha_1(h_1, \varepsilon|n) = 0$. In this case it starts the test for $\{h_1, h_2\}$, i.e., starting from $\varepsilon|n+1$ it looks for another 0 in the column (h_1, α_1) answering h_1 as long as α_1 answers 1. When 0 is visited **Id-Lim** moves to $\alpha_2(h_2, \varepsilon_1)$ and performs a single check for h_2. In such manner we try to check $\{h_1\}$, $\{h_1, h_2\}, \{h_1, h_2, h_3\}, \ldots$

Procedure **Id-Lim** never stops. It is successful if after some point its guesses are still the same and correct with respect to ε.

Why does **Id-Lim** work? One can easily observe that **Id-Lim** runs through every finite sequence of 1s. Visiting a point in which $\alpha_n(h_n, \varepsilon_m) = 1$, it answers h_n. If there is a true hypothesis in H, **Id-Lim** will eventually enter an infinite sequence of 1s (in column (h_m, α_m), say), since H consists of hypotheses verifiable in the limit. Once it enters this sequence there is no way out — **Id-Lim** will indefinitely answer h_m. Therefore **Id-Lim** identifies every $h \in H$ in the limit, and hence H is identifiable in the limit.

QED

In case **Id-Lim** identifies some h_n the procedure needs to remember a finite but not predetermined number of points in ε. We would like to have an algorithm which does not run back and forth on the environment. The answer to this is procedure which is introduced below. Let us call it **Id-Lim***. For this procedure it is enough to remember only one point, namely the position in which the procedure finds itself at each moment.

Id-Lim* uses essentially the same idea of column-ruled searching for strings of 1s. It also consecutively performs it for $\{h_1\}$, $\{h_1, h_2\}$, $\{h_1, h_2, h_3\}$, ...The difference is that when it eventually leaves one column, starting a test for a new hypothesis, it does not go back to ε_1. Instead, it simply moves to the value in the next column but in the same row.

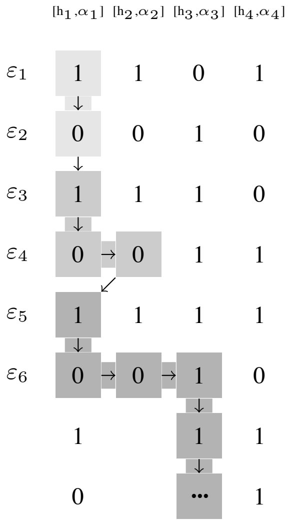


Figure 4. Id-Lim* identifiability

The difference between **Id-Lim** and **Id-Lim*** is mainly in the use of ε. With **Id-Lim*** it is enough to run through ε once without going back. In case of **Id-Lim** every time we fail on some hypothesis and enter a new one, previously not visited, it has to start reading ε from the beginning. **Id-Lim*** also identifies H. It simply leaves out the truth values of hypotheses on some already visited initial segment of ε.

5 Conclusion

The approach presented in this paper can be seen as an attempt to find some general semantic correlates of identification. Inductive verification can be treated as a condition for and a part of the identification process. This fact contributes to the general problem of semantics learning and to modeling the process of scientific inquiry.

Some attempts to approach the problem of learning of semantic constructions are already present in the literature [8, 3]. What is the connection with this framework? The present approach has much to do with the more general idea of model-theoretic learning [1, 7], but it is also related to the work of H.-J. Tiede [8]. In his, slightly different, framework he shows that the class of first-order definable persistent quantifiers of type $\langle 1, 1 \rangle$ is identifiable in the limit. This result is consistent with our considerations. In fact, for the same class of quantifiers we show that it is verifiable with certainty, and that each class containing solely verifiable with certainty structures is identifiable in the limit.

Intuitively there are at least two main parts of human semantic competence. One of them is responsible for producing grammatically correct (syntax domain) or true (semantics domain) hypotheses. The second is a natural correlate of model-checking, i.e., the competence of deciding whether a sentence is true or false in the actual world. The results presented in this paper show how the latter can be embedded in the identification (learning or discovering) process. In this light verification can be seen as a pillar of learning abilities.

References

[1] J. van Benthem and A. Ter Meulen, editors. *Handbook of Logic and Language*. MIT Press, Cambridge, MA, USA, 1997.

[2] P. Feyerabend. *Against Method*. Verso Press, London, 1975.

[3] N. Gierasimczuk. The problem of learning the semantics of quantifiers. In B. ten Cate and H. Zeevat, editors, *Logic, Language, and Computation, TbiLLC 2005*, volume 4363 of *Lecture Notes in Artificial Intelligence*, pages 117–126. Springer, 2007.

[4] E. Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.

[5] S. Jain, D. Osherson, J. S. Royer, and A. Sharma. *Systems that Learn*. MIT Press, Chicago, 1999.

[6] K. Kelly. *The Logic of Reliable Inquiry*. Oxford University Press, Oxford, 1996.

[7] E. Martin and D. Osherson. *Elements of Scientific Inquiry*. MIT Press, Cambridge, 1998.

[8] H.-J. Tiede. Identifiability in the limit of context-free generalized quantifiers. *Journal of Language and Computation*, 1:93–102, 1999.

[9] J. Väänänen. On the expressive power of monotone natural language quantifiers over finite models. *Journal of Philosophical Logic*, 31:327–358, 2002.

[10] J. van Benthem. *Essays in Logical Semantics*. D. Reidel, Dordrecht, 1986.

Inference and Update

Fernando R. Velázquez-Quesada[*]
Institute for Logic, Language and Computation. Universiteit van Amsterdam.
fvelazqu@illc.uva.nl

Abstract

We look at two fundamental logical processes, often intertwined in planning and problem solving: inference and update. Inference is an internal process with which we draw new conclusions, uncovering what is implicit in the information we already have. Update, on the other hand, is produced by external communication, usually in the form of announcements and in general in the form of observations, giving us information that might have been not available (even implicitly) to us before. Both processes have received attention from the logic community, usually separately. In this work, we develop a logical language that allows us to describe them together. We present syntax and semantics, as well as a complete logic for the language; we also discuss similarities and differences with other approaches, and we mention some possible ways the work can be extended.

1. Introduction

Consider the following situation, from [19]:

> You are in a restaurant with your parents, and you have ordered three dishes: fish, meat, and vegetarian. Now a new waiter comes back from the kitchen with the three dishes. What the new waiter can do to get to know which dish corresponds to which person ?

The waiter can ask *"Who has the fish?"*; then, he can ask once again *"Who has the meat?"*. Now he does not have to ask anymore: "two questions plus one inference are all that is needed" ([19]). His reasoning involves two fundamental logical processes: inference and update. The main goal of the present work is to develop a framework in which we can express how they work together.

Inference is an *internal* process: the agent revises her own information in search of what can be derived from it. Update, on the other hand, is produced by *external* communication: the agent gets new information via observations. Both are logical processes, both describe dynamics of information, both are used in every day situations, and still, they have been studied separately.

Inference has been traditionally taken as the main subject of study of logic, "... drawing new conclusions as a means of elucidating or 'unpacking' information that is implicit in the given premises"([20]). Among the most important branches, we can mention Hilbert-style proof systems, natural deduction and tableaux. Recent works, like [7, 8] and [13, 12] have incorporated modal logics to the field, representing inference as a non-deterministic step-by-step process.

Update, on the other hand, has been a main subject of what have been called *Dynamic Epistemic Logic*. Works like [16] and [10] turned attention to the effect public announcements have on the knowledge of an agent. Many works have followed them, including the study of more complex actions ([3, 2]) and the effect of announcements over a more wide propositional attitudes (the soft/hard facts of [17], the knowledge/belief of [4, 5]).

In [20], the author shows how these two phenomena fall directly within the scope of modern logic. As he emphasize, "asking a question and giving an answer is just as 'logical' as drawing a conclusion!". Here, we propose a merging of the two traditions. We consider that both processes are equally important in their own right, but so it is their interaction. In this work, we develop a logical language that join inference and update in a natural way. We first present a modal language to describe inference (section 2). After combining it with epistemic logic (section 3), we give a complete axiomatization. Then we incorporate updates, and we give a set of reduction axioms for the operation (section 4). Finally, we compare our work with other approaches (section 5) and mention some further work we consider interesting (section 6).

[*] Acknowledges a scholarship by **Consejo Nacional de Ciencia y Tecnología** (**CONACyT**) from México. Scholarship holder # 167693.

2. Internal process: an inference language

This section presents a logical language to express inference. The language is based on the work of Jago ([13, 12]), but contain some changes that make it more suitable for our purposes. The agent's information is represented as a set of formulas of a given *internal language*, which in our case is the classical propositional language. Inference steps are then represented as binary relations over such sets, allowing us to use a modal language to talk about them.

Definition 2.1 (Facts and rules) Let $\mathcal{P}$ be a set of atomic propositions, and let $\mathcal{F}_{\mathcal{P}}$ denote the classical propositional language based on $\mathcal{P}$.

- Formulas of $\mathcal{F}_{\mathcal{P}}$ are called *facts* over $\mathcal{P}$.
- A tuple of the form $(\{\lambda_1, \ldots, \lambda_n\}, \lambda)$ (for $n \geq 0$), where each λ_i and λ are facts in $\mathcal{F}_{\mathcal{P}}$, is called a *rule* over $\mathcal{F}_{\mathcal{P}}$. A rule will be also represented as $\lambda_1, \ldots, \lambda_n \Rightarrow \lambda$, and the set of rules over $\mathcal{F}_{\mathcal{P}}$ will be denoted by $R_{\mathcal{F}_{\mathcal{P}}}$.

◁

While facts describe situations about the world, rules describe relations between such situations. Intuitively, a rule $\rho = (\{\lambda_1, \ldots, \lambda_n\}, \lambda)$ indicates that if every λ_i is true, so it is λ. The set of facts $\text{prem}(\rho) := \{\lambda_1, \ldots, \lambda_n\}$ is called the *set of premises of* ρ, and the fact $\text{conc}(\rho) := \lambda$ is called the *conclusion of* ρ.

Definition 2.2 (Internal language) Given a set of atomic propositions $\mathcal{P}$, the *internal language over* $\mathcal{P}$, denoted as $\mathcal{I}_{\mathcal{P}}$, is given by the union of facts in $\mathcal{F}_{\mathcal{P}}$ and rules over $\mathcal{F}_{\mathcal{P}}$, that is, $\mathcal{I}_{\mathcal{P}} = \mathcal{F}_{\mathcal{P}} \cup R_{\mathcal{F}_{\mathcal{P}}}$. ◁

Elements of $\mathcal{I}_{\mathcal{P}}$ will be called in general *formulas* of $\mathcal{I}_{\mathcal{P}}$. The subindexes indicating the set of atomic propositions will be omitted if no confusion arises.

For expressing how the agent's information evolves through inference steps, a (modal) inference language is defined.

Definition 2.3 (Language $\mathcal{IL}$) Let $\mathcal{A}$ be a set of agents and $\mathcal{P}$ a set of atomic propositions. Formulas φ of the *inference language* $\mathcal{IL}$ are given by

$$\varphi ::= \top \mid I_i\,\gamma \mid \neg\varphi \mid \varphi \vee \psi \mid \langle\rho\rangle_i\,\varphi$$

with $i \in \mathcal{A}$ and γ, ρ formulas of the internal language $\mathcal{I}_{\mathcal{P}}$ with ρ a rule. Formulas of the form $I_i\,\gamma$ express *"the agent i is informed about γ"*, while formulas of the form $\langle\rho\rangle_i\,\varphi$ express *"there is an inference step in which agent i applies the rule ρ and, after doing it, φ is the case"*. ◁

The semantic model of $\mathcal{IL}$ is based on a Kripke model: we have a set of worlds and labeled binary relations between them. The main idea is that every world represents the information of the agents at a given stage, while a relation with label $D_{(\rho,i)}$ from a world w to a world w' indicates that the information of agent i at w allows her to perform an inference step with rule ρ, and that the information that results from applying ρ at w is represented by w'. To make formal this intuitive idea, we first need to define what we will understand by the phrases *"the information of i at w allows her to perform an inference step with ρ"* and *"the information that results from applying ρ at w is represented by w'"*. The concepts of *set-matching rule* and *rule-extension of a world* will do the job.

We will use the following abbreviation. Given a universe U, a set $A \subseteq U$ and an element $a \in U$, we denote $A \cup \{a\}$ as $A + a$.

Definition 2.4 (Set-matching rule) Let ρ be a rule in $\mathcal{I}$ and let Γ be a set of formulas of $\mathcal{I}$. We say that ρ *is* Γ-*matching* (ρ can be applied at Γ) if and only if ρ and all its premises are in Γ, that is, $(\text{prem}(\rho) + \rho) \subseteq \Gamma$. ◁

Definition 2.5 (Extension of set of formulas) Let ρ be a rule in $\mathcal{I}$, and let Γ, Γ' be sets of formulas of $\mathcal{I}$. We say that Γ' *is a ρ-extension of* Γ if and only if Γ' is Γ plus the conclusion of ρ, that is, $\Gamma' = \Gamma + \text{conc}(\rho)$. ◁

With the notions of Γ-matching rule and ρ-extension of Γ, we can give a formal definition of the models where formulas of $\mathcal{IL}$ are interpreted.

Definition 2.6 (Inference model) Let $\mathcal{A}$ be a set of agents and let $\mathcal{P}$ be a set of atomic propositions. An *inference model* is a tuple $M = (W, D_{(\rho,i)}, Y_i)$ where

- W is a non-empty set of worlds.
- $Y_i : W \to \wp(\mathcal{I}_{\mathcal{P}})$ is the *information set function* for each agent $i \in \mathcal{A}$. It assigns to i a set of formulas of the internal language in each world w.
- $D_{(\rho,i)} \subseteq (W \times W)$ is the *inference relation* for each pair (ρ, i), with ρ a rule in $\mathcal{I}_{\mathcal{P}}$ and i an agent in $\mathcal{A}$. The relation represents the application of a rule, so if $D_{(\rho,i)}\,ww'$, then ρ is $Y_i(w)$-matching and $Y_i(w')$ is a ρ-extension of $Y_i(w)$.

◁

Note that the definition of $D_{(\rho,i)}$ just states the property any tuple should satisfy in order to be in the relation. The relation is not induced by the property, so it is possible to have two worlds w and w' such that there is a rule ρ that is $Y_i(w)$-matching and $Y_i(w')$ is a ρ-extension of $Y_i(w)$, and still do not have the pair (w, w') in $D_{(\rho,i)}$. One of the goals of the work is to make the basic definitions as general as possible, and then analyze the different concepts of inference and information we can get by asking for extra properties of the inference relation[1] and of the information sets (as we do later for the case of truthful information, that is, knowledge). This allows us to represent agents that are not as powerful reasoners as those represented with classic epistemic logic, and it may play an important role when studying agents with diverse reasoning abilities (cf. the discussion in section 6).

The concepts of *set-matching rule* and *rule-extension of a world* have their *possible world* version. We say that ρ is w-matching for i if it is $Y_i(w)$-matching, and we say that w' is a ρ-extension of w for i if $Y_i(w')$ is a ρ-extension of $Y_i(w)$.

Definition 2.7 Given an inference model $M = (W, D_{(\rho,i)}, Y_i)$ and a world $w \in W$, the relation $\models$ between the pair M, w and $\top$ (the always true formula), negations and disjunctions is given as usual. For the remaining formulas, we have

$$\begin{array}{lll} M, w \models I_i\,\gamma & \textit{iff} & \gamma \in Y_i(w) \\ M, w \models \langle\rho\rangle_i\,\varphi & \textit{iff} & \text{there is } w' \in W \text{ such that} \\ & & D_{(\rho,i)}\,ww' \text{ and } M, w' \models \varphi \end{array}$$

◁

3. The real world: an epistemic inference language

We have a language that express the agent's information and how it evolves through inferences. Still, we cannot talk about the real world or about the agent's uncertainty. In this section, we extend the current language to express those notions.

Syntactically, we extend the inference language with classical epistemic logic. We add basic formulas of the form p (for p an atomic proposition) and we close it under the modal operator P_i (for i an agent).

Definition 3.1 (Epistemic inference language) Let $\mathcal{A}$ be a set of agents and let $\mathcal{P}$ be a set of atomic propositions. The formulas of the *epistemic inference language* $\mathcal{EI}$ are given by

$$\varphi ::= \top \mid p \mid I_i\,\gamma \mid \neg\varphi \mid \varphi \vee \psi \mid P_i\,\varphi \mid \langle\rho\rangle_i\varphi$$

with $i \in \mathcal{A}$, $p \in \mathcal{P}$ and γ, ρ formulas of the internal language $\mathcal{I}_\mathcal{P}$ with ρ a rule. ◁

The propositional connectives $\wedge$, $\rightarrow$ and $\leftrightarrow$ are defined as usual; the modal operators K_i and $[\rho]_i$ are defined as the dual of P_i and $\langle\rho\rangle_i$, respectively.

As argued by van Benthem in [18], the operator K_i should be read as a more implicit notion, describing not the information the agent actually has, but the maximum amount of information she can get under her current uncertainty (i.e., without external interaction). In our framework, *explicit* information is represented with formulas of the form $I_i\,\gamma$, indicating that γ is part of the agent's information set; *implicit* information is represented with formulas of the form $K_i\,\varphi$, indicating what the agent can eventually get if she has enough explicit information (i.e., enough formulas and rules) and enough time to perform the adequate inference steps.

Semantically, we combine inference models with classic Kripke models. Each world has two components: information sets containing the facts and rules each agent is informed about, and a valuation indicating the truth value of atomic propositions. We also have two binary relations: the inference one indicating how inference steps modify information sets, and the epistemic one indicating the worlds each agent considers possible.

Definition 3.2 (Epistemic inference model) Let $\mathcal{A}$ be a set of agents and let $\mathcal{P}$ be a set of atomic propositions. An *epistemic inference model* is a tuple $M = (W, \sim_i, D_{(\rho,i)}, V, Y_i)$ where:

- W is a non-empty set of worlds.
- $V : W \to \wp(\mathcal{P})$ is a *valuation function*.
- $Y_i : W \to \wp(\mathcal{I}_\mathcal{P})$ is the *information set function* for agent i.
- $D_{(\rho,i)}$ is the *inference relation* for each pair (ρ, i), just as in definition 2.6. It satisfies an extra requirement: if $D_{(\rho,i)}\,ww'$, then $V(w) = V(w')$.
- $\sim_i$ is the *epistemic relation* for agent i. The relation satisfy the following property: for all worlds w, w', u, u': if $w \sim_i u$ and $D_{(\rho,i)}\,ww'$, $D_{(\rho,i)}\,uu'$ for some rule ρ, then $w' \sim_i u'$.

◁

We have two new restrictions: one for the inference relation and one relating it with the epistemic relation. It is worthwhile to justify them.

1. The relation $D_{(\rho,i)}$ describes inference, an agent's internal process that changes her information but does not change the real situation. If an agent can go from w to w' by an inference step, w and w' should satisfy the same propositional letters.

[1] In fact, the definition of $D_{(\rho,i)}$ restricts inferences to *deductive* ones. Within the proposed framework, it is possible to represent other inference processes, as mentioned in section 6.

2. This property, called *no miracles* in [21] and related with the *no learning* property of [11], reflects the following idea: if two worlds are epistemically indistinguishable and the same rule is applied at both of them, then the resulting worlds should be epistemically indistinguishable too.

Definition 3.3 Given an epistemic inference model $M = (W, \sim_i, D_{(\rho,i)}, V, Y_i)$ and a world $w \in W$, the relation $\models$ between the pair M, w and $\top$, negations and disjunctions is given as usual. For the remaining formulas, we have:

$$\begin{array}{lll} M, w \models p & \textit{iff} & p \in V(w) \\ M, w \models I_i\,\gamma & \textit{iff} & \gamma \in Y_i(w) \\ M, w \models P_i\,\varphi & \textit{iff} & \text{there is } u \in W \text{ such that} \\ & & w \sim_i u \text{ and } M, u \models \varphi \\ M, w \models \langle\rho\rangle_i\,\varphi & \textit{iff} & \text{there is } w' \in W \text{ such that} \\ & & D_{(\rho,i)}\,ww' \text{ and } M, w' \models \varphi \end{array}$$

A formula φ is *valid in a epistemic inference model* M (notation $M \models \varphi$) if $M, w \models \varphi$ for all worlds w in M. A formula φ is *valid in the class of models* **M** (notation $\mathbf{M} \models \varphi$) if φ is valid in M ($M \models \varphi$) for all M in **M**. ◁

As it is currently defined, epistemic inference models do not impose any restriction to the information sets: *any* propositional formula of $\mathcal{I}$ can be in *any* information set $Y_i(w)$. We can have non-veridical information sets (if we have $\gamma \in Y_i(w)$ and $M, w \not\models \gamma$ for some $w \in W$) describing situations where the information of the agent is not true, or even inconsistent ones (if we have γ and $\neg\gamma$ in $Y_i(w)$ for some $w \in W$), describing situations where her information is contradictory.

In the present work we focus on a special class of models: those in which the information sets of the agents describe *knowledge*. We ask for the epistemic relation to be an equivalence one, and we ask for all formulas of an information set to be true at the correspondent world.

Definition 3.4 (Class $\mathbf{EI}_K$) The class of epistemic inference models $\mathbf{EI}_K$ contains exactly those models in which each $\sim_i$ is an equivalence relation and for every world $w \in W$, if $\gamma \in Y_i(w)$ then $M, w \models \gamma$. The following table summarize the properties of models in this class.

P1	$D_{(\rho,i)}\,ww'$ implies ρ is w-matching and w' is a ρ-extension of w (for i).
P2	If $D_{(\rho,i)}\,ww'$, then w and w' satisfy the same propositional letters.
P3	If $D_{(\rho,i)}\,ww'$, $D_{(\rho,i)}\,uu'$ and $w \sim_i u$ for some rule ρ, then $w' \sim_i u'$.
P4	$\sim_i$ is an equivalence relation.
P5	$\gamma \in Y_i(w)$ implies $M, w \models \gamma$.

◁

Our first result is a syntactic characterization of formulas of $\mathcal{EI}$ that are valid on models of $\mathbf{EI}_K$. Non-defined concepts, like a (modal) logic, Λ-consistent / inconsistent set and maximal Λ-consistent set (for a normal modal logic Λ) are completely standard, and can be found in chapter 4 of [6].

Definition 3.5 (Logic EI_K) The logic EI_K is the smallest set of formulas of $\mathcal{EI}$ that is created from the set of axioms [2] and a set of rules of table 1. ◁

Axioms		
P	All propositional tautologies	
E-K	$K_i\,(\varphi \rightarrow \psi) \rightarrow (K_i\,\varphi \rightarrow K_i\,\phi)$	
E-Dual	$P_i\,\varphi \leftrightarrow \neg K_i\,\neg\varphi$	
I-K	$[\rho]_i\,(\varphi \rightarrow \psi) \rightarrow ([\rho]_i\,\varphi \rightarrow [\rho]_i\,\psi)$	
I-Dual	$\langle\rho\rangle_i\,\varphi \leftrightarrow \neg[\rho]_i\,\neg\varphi$	
T	$\varphi \rightarrow P_i\,\varphi$	
4	$P_i\,P_i\,\varphi \rightarrow P_i\,\varphi$	
B	$\varphi \rightarrow K_i\,P_i\,\varphi$	
A1	$[\rho]_i\,I_i\;\mathrm{conc}(\rho)$	
A2	$\langle\rho\rangle_i\,\top \rightarrow I_i\,(\mathrm{prem}(\rho) + \rho)$	
A3	$I_i\,\gamma \rightarrow [\rho]_i\,I_i\,\gamma$	
A4	$\langle\rho\rangle_i\,I_i\,\gamma \rightarrow I_i\,\gamma$	with $\gamma \neq \mathrm{conc}(\rho)$.
A5	$(p \rightarrow [\rho]_i\,p) \wedge (\neg p \rightarrow [\rho]_i\,\neg p)$	with $p \in \mathcal{P}$.
A6	$(\langle\rho\rangle_i\,\varphi \wedge P_i\,\langle\rho\rangle_i\,\psi) \rightarrow \langle\rho\rangle_i\,(\varphi \wedge P_i\,\psi)$	
A7	$I_i\,\gamma \rightarrow \gamma$	

Rules	
MP	Given φ and $\varphi \rightarrow \psi$, prove ψ
E-Gen	Given φ, prove $K_i\,\varphi$
I-Gen	Given φ, prove $[\rho]_i\,\varphi$

Table 1. Axioms and rules for EI_K.

Theorem 3.6 (Soundness) *The logic EI_K is sound with respect to the class $\mathbf{EI}_K$.*

Proof. *For soundness, we just need to prove that axioms of EI_K are valid in $\mathbf{EI}_K$, and that its rules preserve validity. We omit the details here.* QED

Strong completeness is equivalent to satisfiability of consistent set of formulas, as mentioned in Proposition 4.12 of [6].

Theorem 3.7 (Completeness) *The logic EI_K is strongly complete with respect to the class $\mathbf{EI}_K$.*

Proof. *We define the canonical model M^{EI_K} for the logic EI_K. With the the Lindenbaum's Lemma, the Existence Lemma and the Truth Lemma, we show that every EI_K-consistent set of formulas is satisfiable in M^{EI_K}. Finally, we show that M^{EI_K} is indeed a model in $\mathbf{EI}_K$. See section A.1 for details.* QED

[2] Formulas of the form $I_i\,\Gamma$ are abbreviations of $\bigwedge_{\gamma\in\Gamma} I_i\,\gamma$, for a finite $\Gamma \subseteq \mathcal{I}$.

4. External interaction: explicit observations

So far, our language can express the agent's *internal* dynamics, but it cannot express *external* ones. We can express how inference steps modify the explicit information, but we cannot express how both explicit and implicit one are affected by external observations. Here we add the other fundamental source of information; in this section, we extend the language to express updates. For easiness of reading and writing, we remove subindexes referring to agents.

Updates are usually represented as operations that modify the semantic model. In *Public Announcement Logic* (PAL), for example, an announcement is defined by an operation that removes the worlds where the announced formula does not hold, restricting the epistemic relation to those that are not deleted.

In our semantic model, we have a finer representation of the agent's information. We have explicit information (her information sets) but we also have implicit one (what she can add to her information set via inference). Then, we can extend PAL by defining different kinds of model operations, affecting explicit and implicit information in different forms, and therefore expressing different ways the agent processes the new information. Here, we present one of the possible definitions, what we have called *explicit observations*.

Definition 4.1 (Explicit observation) Let $M = (W, \sim, D_\rho, V, Y)$ be an epistemic inference model, and let γ be a formula of the internal language. The epistemic inference model $M_{+\gamma!} = (W', \sim', D'_\rho, V', Y')$ is given by

- $W' := \{ w \in W \mid M, w \models \gamma \}$
- $\sim' := \{ (w, u) \in W' \times W' \mid w \sim u \}$
- $D'_\rho := \{ (w, u) \in W' \times W' \mid D_\rho wu \}$
- $V'(w) := V(w)$ for $w \in W'$
- $Y'(w) := Y(w) + \gamma$ for $w \in W'$

$\lhd$

Our explicit observation operation behave as the standard public announcement with respect to worlds, valuation and relations. With respect to the information set functions, we have chosen a simple definition: once a formula is announced, it will become part of the agent's explicit information. The choice is also a good one, since the operation is closed for models in $\mathbf{EI}_K$.

Proposition 4.2 *If M is a model in $\mathbf{EI}_K$, so it is $M_{+\gamma!}$.*

Proof. *See section A.2.* QED

The new language $\mathcal{EEI}$ extends $\mathcal{EI}$ by closing it under explicit observations. Take a formula γ in the internal language; if φ is a formula in $\mathcal{EEI}$, so it is $[+\gamma!]\,\varphi$. The semantics for formulas already in $\mathcal{EI}$ is defined as before (definition 3.3). For explicit observation formulas, we have the following.

Definition 4.3 Let M be a model in $\mathbf{EI}_K$, and let $w \in W$ be a world in it. Then:

$$M, w \models [+\gamma!]\,\varphi \quad \textit{iff} \quad \begin{array}{l} M, w \models \gamma \text{ implies} \\ M_{+\gamma!}, w \models \varphi \end{array}$$

$\lhd$

Our second result is a syntactic characterization of the formulas in $\mathcal{EEI}$ that are valid in models in $\mathbf{EI}_K$. By proposition 4.2, the explicit observation operation is closed for models in $\mathbf{EI}_K$, so we can rely on the logic EI_K: all we have to do is give a set of reduction axioms for formulas of the form $[+\gamma!]\,\varphi$. The standard reduction axioms for atomic propositions, negations, disjunctions and epistemic formulas work for $\mathcal{EEI}$ too; we just have to add axioms indicating how information set formulas and inference formulas are affected.

Theorem 4.4 *The logic EEI_K, built from axioms and rules of EI_K (see table 1) plus axioms and rules in table 4.4, is sound and strongly complete for the class $\mathbf{EI}_K$.*

Axioms		
EO-1	$[+\gamma!]\,p \;\leftrightarrow\; (\gamma \rightarrow p)$	
EO-2	$[+\gamma!]\,\neg\varphi \;\leftrightarrow\; (\gamma \rightarrow \neg[+\gamma!]\,\varphi)$	
EO-3	$[+\gamma!]\,(\varphi \vee \psi) \;\leftrightarrow\; ([+\gamma!]\,\varphi \vee [+\gamma!]\,\psi)$	
EO-4	$[+\gamma!]\,K\,\varphi \;\leftrightarrow\; (\gamma \rightarrow K\,[+\gamma!]\,\varphi)$	
EO-5	$[+\gamma!]\,I\,\gamma \;\leftrightarrow\; \top$	
EO-6	$[+\gamma!]\,I\,\delta \;\leftrightarrow\; (\gamma \rightarrow I\,\delta)$	for $\delta \neq \gamma$
EO-7	$[+\gamma!]\,[\rho]\,\varphi \;\leftrightarrow\; (\gamma \rightarrow [\rho]\,[+\gamma!]\,\varphi)$	
Rules		
EO-Gen	Given φ, prove $[+\gamma!]\,\varphi$	

Table 2. Axioms and rules for explicit observations.

Proof. *Soundness comes from the validity of the new axioms and the validity-preserving property of the new rule. Strong completeness comes from the fact that, by a repetitive application of such axioms, any explicit observation formula can be reduced to a formula in $\mathcal{EI}$, for which EI_K is strongly complete with respect to $\mathbf{EI}_K$.* QED

The language $\mathcal{EEI}$ can express uncertainty (as classic epistemic logic does), inference (as the modal approaches of [7, 8, 13, 12]) and update (as PAL). Moreover, it can express its combinations. With it, we are able to talk about the merging of *internal* dynamics, expressing the way the agent *"unpacks"* her implicit information, with external ones, expressing how her interaction with her environment modifies what she is informed about.

We have provided semantics for the language; semantics that reflect the nature of each process. Inferences are represented as relations between information sets. This reflects the idea that, with enough initial explicit information, the agent may get all the implicit information by the adequate rule applications. Update, on the other hand, is defined as a model operation. It is a process that not only provides explicit information, but also modifies implicit one. This reflects the idea that updates yields information that might have not been available to the agent before.

Among the semantic models, we distinguish the class $\mathbf{EI}_K$, which contains those where the agent's information is in fact knowledge. We give a syntactic characterization of the valid formulas in $\mathbf{EI}_K$ by means of the sound and complete logic EEI_K.

5. Comparison with other works

The present work is a combination of three main ideas: the representation of explicit information as set of formulas, relations between such sets to represent inferences and model operations to represent updates. The first two have been used in some other works; we present a brief comparison between some of them and our approach.

5.1. Fagin-Halpern's logics of awareness

Fagin and Halpern presented in [9] what they called *logic of general awareness* ($\mathcal{L}_A$). Given a set of agents, formulas of the language are given by a set of atomic propositions $\mathcal{P}$ closed under negation, conjunction and the modal operators A_i and L_i (for an agent i). Formulas of the form $A_i\varphi$ are read as *"the agent i is aware of φ"*, and formulas of the form $L_i\varphi$ are read as *"the agent i implicitly believes that φ"*. The operator B_i, which expresses explicit beliefs, is defined as $B_i\varphi := A_i\varphi \wedge L_i\varphi$.

A *Kripke structure for general awareness* is defined as a tuple $M = (W, \mathfrak{A}_i, \mathfrak{L}_i, V)$, where $W \neq \emptyset$ is the set of possible worlds, $\mathfrak{A}_i : W \to \wp(\mathcal{L}_A)$ is a function that assigns a set of formulas of $\mathcal{L}_A$ to the agent i in each world (her awareness set), the relation $\mathfrak{L}_i \subseteq (W \times W)$ is a serial, transitive and Euclidean relation over W for each agent i ($\mathcal{L}_A$ deals with beliefs rather than knowledge) and $V : \mathcal{P} \to \wp(W)$ is a valuation function.

Given a Kripke structure for general awareness $M = (W, \mathfrak{A}_i, \mathfrak{L}_i, V)$, semantics for atomic propositions, negations and conjunctions are given in the standard way. For formulas of the form $A_i\,\varphi$ and $L_i\,\varphi$, we have

$$\begin{array}{lll} M, w \models A_i\varphi & \textit{iff} & \varphi \in \mathfrak{A}_i(w) \\ M, w \models L_i\varphi & \textit{iff} & \text{for all } u \in W, \\ & & \mathfrak{L}_i wu \text{ implies } M, u \models \varphi \end{array}$$

It follows that $M, w \models B_i\varphi$ *iff* $\varphi \in A_i(w)$ and, for all $u \in W$, $\mathfrak{L}_i wu$ implies $M, u \models \varphi$.

Given the similarities between the functions $\mathfrak{A}_i$ and Y_i and between the relations $\mathfrak{L}_i$ and $\sim_i$, formulas $A_i\varphi$ and $L_i\varphi$ in $\mathcal{L}_A$ behaves exactly like $I_i\,\varphi$ and $K_i\,\varphi$ in $\mathcal{EEI}$. The difference in the approaches is in the dynamic part.

For the internal dynamics (inference), the language $\mathcal{L}_A$ does not express changes in the agent's awareness sets. Later in the same paper, Fagin and Halpern explore the incorporation of time to the language by adding a deterministic serial binary relation $\mathfrak{T}$ over W to represent steps in time. Still, they do not indicate what the process(es) that change the awareness sets is (are).

In our approach, pairs in the inference relation $D_{(\rho,i)}$ have a specific interpretation: they indicate *steps in the agent's reasoning process*. Because of this, we have a particular definition of how they should behave (properties **P1**, **P2**, and **P3**). Moreover, external dynamics (observations), which are not considered $\mathcal{L}_A$, are represented in a different way, as model operations.

There is another conceptual difference. In $\mathcal{L}_A$, elements of the awareness sets are just formulas; in $\mathcal{EI}$, elements of the information sets are not only formulas (what we have called *facts*) but also *rules*. The information of the agent consists not only on facts, but also on rules that allow her to infer new facts. It is not that the agent knows that after a rule application her information set will change; it is that she knows the *process* that leads the change. We interpret a rule as an object that can be part of the agent's information, and whose presence is needed for the agent to be able to apply it.

5.2. Duc's dynamic epistemic logic

In [7] and [8], Ho Ngoc Duc proposes a dynamic epistemic logic to reason about agents that are neither logically omniscient nor logically ignorant.

The syntax of the language is very similar to the inference part of our language. There is an internal language, the classic propositional one (PL), to express agent's knowledge. There is also another language to talk about how this knowledge evolves. Formally, *At* denotes the set of formulas of the form $K\gamma$, for γ in PL. The language $\mathcal{L}_{BDE}$ contains *At* and is closed under negation, conjunction and the modal operator $\langle F\rangle$. Formulas of the form $K\gamma$ are read as *"γ is known"*; formulas of the form $\langle F\rangle\varphi$ are read as *"φ is true after some course of thought"*.

A model M is a tuple (W, R, Y), where $W \neq \emptyset$ is the set of *possible worlds*, $R \subseteq (W \times W)$ is a transitive binary relation and $Y : W \to \wp(At)$ associates a set of formulas of *At* to each possible world. A *BDE*-model is a model M such that: (1) for all $w \in W$, if $K\gamma \in Y(w)$ and Rwu, then $K\gamma \in Y(u)$; (2) for all $w \in W$, if $K\gamma$ and $K(\gamma \to \delta)$ are in

$Y(w)$, then $K\delta$ is in $Y(u)$ for some u such that Rwu; (3) if γ is a propositional tautology, then for all $w \in W$ there is a world u such that Rwu and $K\gamma \in Y(u)$. Such restrictions guarantees that the set of formulas will grow as the agent reasons, and that her knowledge will be closed under modus ponens and will contain all tautologies at some point in the future.

Given a *BDE*-model, the semantics for negation and conjunctions are standard. The semantics of atomic and reasoning-steps formulas are given by:

$$\begin{array}{lll} M, w \models K\gamma & \textit{iff} & K\gamma \in Y(w) \\ M, w \models \langle F \rangle \varphi & \textit{iff} & \text{there is } u \in W \text{ such that} \\ & & Rwu \text{ and } M, u \models \varphi \end{array}$$

Note that the language does not indicate what a *"course of though"* is; again, our framework is more precise. Also, it does not consider sentences about the world. Finally, the language is restricted to express what the agent can infer through some *"course of though"*, but it does not express external dynamics, as explicit observations in $\mathcal{EEI}$ do.

6. Further work

In order to give a finer representation of the inference process, we have chosen to represent information as set of formulas. This is also a solution for the famous *logical omniscience* problem, since sets of formulas do not need to satisfy *a priori* any particular property, like being closed under some consequence relation. Among other approaches for the problem, there is the non-classical worlds approach for epistemic logic. The idea is to add worlds in which the usual rules of logic do not hold. The knowledge of the agents is affected since non-classical worlds may be considered possible. It would be interesting to look at this approach as an alternative for representing the agent's explicit information, and see what the differences are.

Our framework do not represent in a completely faithful way the intuitive idea of the application of a rule. It is possible to have a world in which a rule can be applied, and not to have a world that results from its application. We can focus on models on which, if a rule is applicable, then *there is a world* that results from its application. This forces us to change the defined explicit observation operation since, in general, the resulting model will not have the required property: the added formula can make applicable a rule that was not applicable before. The immediate solution is to create all needed worlds, but this iterative process complicates the operation, and the existence of reduction axioms is not so clear anymore.

As mentioned in the text, properties **P4** and **P5** characterize models in which the information the agent has is in fact knowledge, that is, the epistemic relation is an equivalence one and formulas in all information sets are true at the correspondent world. It would be interesting to be able to talk about not only knowledge but also *beliefs*. Some recent works ([17, 4, 5] among others) combine these two notions, giving us a nice way of studying these two propositional attitudes together.

Property **P1** defines not only the situation when a rule can be applied (whenever a rule a rule and all its premises are in the agent's information set), but also what results from the application (the given information set extended by the conclusion of the rule). The property indeed restricts our models to those that use rules in a *deductive* way, that is, to those that represent just *deductive* inference. There are other interesting inference processes, like *abduction* or *belief revision*; they are not deductive, but they are important and widely used, with particular relevance on incomplete information situations. Within the proposed framework, we can represent different inference processes, and we can study how all of them work together.

For the external dynamics, we mentioned that this finer representation of knowledge allows us to define different kinds of observations. Since we represent both explicit and implicit information, we can define different model operations, allowing us to explore the different ways an agent process new information.

In the context of agent diversity ([14, 15]), a finer representation of the inference process allows us to make a distinction between agents with different reasoning abilities. The rules an agent has in her information set may be very different from those in the information set of another, and they will not be able to perform the same inference steps. Moreover, some of them may be able to perform several inference steps at once instead of a single one. The idea works also for external dynamics: agents may have different observational power. It will be interesting to explore how agents that differs in their reasoning and observational abilities interact with each other.

Acknowledgments

The author would like to thank Johan van Benthem for his invaluable ideas and suggestions that led to the present work. He also would like to thank the participants of the ILLC's *Seminar on Logics for Dynamics of Information and Preferences*; their comments helped to improve earlier versions.

A. Technical appendix

A.1. Proof of completeness

As mentioned, the key observation is that a logic Λ is strongly complete with respect to a class of structures if and only if every Λ-consistent set of formulas is satisfiable on some structure of the given class (Proposition 4.12 of [6]). Using the the canonical model technique, we show that every EI_K-consistent set of formulas is satisfiable in a model in $\mathbf{EI}_K$. Proofs of Lindenbaum's Lemma, Existence Lemmas and Truth Lemma are standard.

Lemma A.1 (Lindenbaum's Lemma) *For any EI_K-consistent set of formulas Σ, there is a maximal EI_K-consistent set Σ^+ such that $\Sigma \subseteq \Sigma^+$.*

Definition A.2 (Canonical model) The canonical model of the logic EI_K is the epistemic inference model $M^{\mathsf{EI}_K} = (W^{\mathsf{EI}_K}, \sim_i^{\mathsf{EI}_K}, D_{(\rho,i)}^{\mathsf{EI}_K}, V^{\mathsf{EI}_K}, Y_i^{\mathsf{EI}_K})$, where:

- W^{EI_K} is the set of all maximal EI_K-consistent set of formulas.
- $w \sim_i^{\mathsf{EI}_K} u$ *iff* for all φ in $\mathcal{EI}$, $\varphi \in u$ implies $P_i\, \varphi \in w$ (equivalently, $w \sim_i^{\mathsf{EI}_K} u$ *iff* for all φ in $\mathcal{EI}$, $K_i\, \varphi \in w$ implies $\varphi \in u$).
- $w D_{(\rho,i)}^{\mathsf{EI}_K} w'$ *iff* for all φ in $\mathcal{EI}$, $\varphi \in w'$ implies $\langle \rho \rangle_i\, \varphi \in w$ (equivalently, $w D_{(\rho,i)}^{\mathsf{EI}_K} w'$ *iff* for all φ in $\mathcal{EI}$, $[\rho]_i\, \varphi \in w$ implies $\varphi \in w'$).
- $V^{\mathsf{EI}_K}(w) := \{\, p \in \mathcal{P} \mid p \in w \,\}$.
- $Y_i^{\mathsf{EI}_K}(w) := \{\, \gamma \in \mathcal{I} \mid I_i\, \gamma \in w \,\}$.

◁

Lemma A.3 (Existence Lemmas) *For any world $w \in W^{\mathsf{EI}_K}$, if $P_i\, \varphi \in w$, then there is a world $u \in W^{\mathsf{EI}_K}$ such that $w \sim_i^{\mathsf{EI}_K} u$ and $\varphi \in u$. For any world $w \in W^{\mathsf{EI}_K}$, if $\langle \rho \rangle_i\, \varphi \in w$, then there is a world $w' \in W^{\mathsf{EI}_K}$ such that $D_{(\rho,i)}^{\mathsf{EI}_K}\, ww'$ and $\varphi \in w'$.*

Lemma A.4 (Truth Lemma) *For all $w \in W^{\mathsf{EI}_K}$, we have $M^{\mathsf{EI}_K}, w \models \varphi$ iff $\varphi \in w$.*

By the mentioned Proposition of [6], all we have to show is that every EI_K-consistent set of formulas is satisfiable, so take any such set Σ. By Lindenbaum's Lemma, we can extend it to a maximal EI_K-consistent set of formulas Σ^+; by the Truth Lemma, we have $M^{\mathsf{EI}_K}, \Sigma^+ \models \Sigma$, so Σ is satisfiable in the canonical model of EI_K at Σ^+. Now we have to show that the canonical model M^{EI_K} is indeed a model in $\mathbf{EI}_K$.

Axioms **T**, **4** and **B** are canonical for reflexivity, transitivity and symmetry, respectively, so $\sim_i^{\mathsf{EI}_K}$ is an equivalence relation and property **P4** is fulfilled. It remains to show that M^{EI_K} satisfy **P1**, **P2**, **P3** and **P5**. We have removed the agent's subindexes for easiness of writing and reading.

Remember that any maximal EI_K-consistent set Φ is closed under modus ponens, that is, if φ and $\varphi \rightarrow \psi$ are in Φ, so it is ψ.

P1 Suppose $D_\rho^{\mathsf{EI}_K}\, ww'$; we want to show that $(\mathrm{prem}(\rho) + \rho) \subseteq Y^{\mathsf{EI}_K}(w)$ and that $Y^{\mathsf{EI}_K}(w') = Y^{\mathsf{EI}_K}(w) + \mathrm{conc}(\rho)$.

For the first part, $D_\rho^{\mathsf{EI}_K}\, ww'$ implies $M^{\mathsf{EI}_K}, w \models \langle \rho \rangle \top$, so $\langle \rho \rangle \top \in w$. By axiom **A2** and modus ponens closure, we have $I\, (\mathrm{prem}(\rho) + \rho) \in w$. Then, $\mathrm{prem}(\rho)$ and ρ are in $Y^{\mathsf{EI}_K}(w)$.

For the second part, we will show both inclusions, i.e., we will show that $Y^{\mathsf{EI}_K}(w) + \mathrm{conc}(\rho) \subseteq Y^{\mathsf{EI}_K}(w')$ and $Y^{\mathsf{EI}_K}(w') \subseteq Y^{\mathsf{EI}_K}(w) + \mathrm{conc}(\rho)$.

- Take any $\gamma \in Y^{\mathsf{EI}_K}(w)$; then, $I\, \gamma \in w$. By axiom **A3** and the modus ponens closure, $[\rho]\, I\, \gamma \in w$. Since $D_\rho^{\mathsf{EI}_K}\, ww'$, we have $I\, \gamma \in w'$ and then $\gamma \in Y^{\mathsf{EI}_K}(w')$.

  It remains to show that $\mathrm{conc}(\rho) \in Y^{\mathsf{EI}_K}(w')$. Since axiom **A1** is in w and $D_\rho^{\mathsf{EI}_K}\, ww'$, we have $I\ \mathrm{conc}(\rho) \in w'$ and therefore $\mathrm{conc}(\rho) \in Y^{\mathsf{EI}_K}(w')$.
- Take any $\gamma \in (Y^{\mathsf{EI}_K}(w') - \mathrm{conc}(\rho))$; then, $I\, \gamma \in w'$. Since $D_\rho^{\mathsf{EI}_K}\, ww'$, we have $\langle \rho \rangle\, I\, \gamma \in w$ and, by axiom **A4**, we have $I\, \gamma \in w$; then, $\gamma \in Y^{\mathsf{EI}_K}(w)$. Hence, $Y^{\mathsf{EI}_K}(w') - \mathrm{conc}(\rho) \subseteq Y^{\mathsf{EI}_K}(w)$, and therefore $Y^{\mathsf{EI}_K}(w') \subseteq Y^{\mathsf{EI}_K}(w) + \mathrm{conc}(\rho)$.

P2 Suppose $D_\rho^{\mathsf{EI}_K}\, ww'$; we want to show that w and w' satisfy the same propositional letters. Note that we have **A5** in w, and then both $p \rightarrow [\rho]\, p$ and $\neg p \rightarrow [\rho]\, \neg p$ are in w since it is a maximal consistent set.

If $M^{\mathsf{EI}_K}, w \models p$ then, by definition of V^{EI_K}, we have $p \in w$. But $(p \rightarrow [\rho]\, p) \in w$ and, by the modus ponens closure, $[\rho]\, p \in w$. Then, since $D_\rho^{\mathsf{EI}_K}\, ww'$, we have $p \in w'$, so $M^{\mathsf{EI}_K}, w' \models p$.

If $M^{\mathsf{EI}_K}, w \not\models p$, then $M^{\mathsf{EI}_K}, w \models \neg p$; by definition of V^{EI_K}, we have $\neg p \in w$. But $(\neg p \rightarrow [\rho]\, \neg p) \in w$, so the modus ponens closure implies $[\rho]\, \neg p \in w$. Then, since $D_\rho^{\mathsf{EI}_K}\, ww'$, we have $\neg p \in w'$, so $M^{\mathsf{EI}_K}, w' \models \neg p$, i.e., $M^{\mathsf{EI}_K}, w' \not\models p$.

P3 Note that axiom **A6** is a Sahlqvist formula (a very simple Sahlqvist formula indeed; see section 3.6 of [6] for details). Its first-order local correspondent is the formula

$$(\forall w')(\forall u)(\forall u')\big((D_\rho\, ww' \wedge w \sim u \wedge D_\rho\, uu') \\ \rightarrow (D_\rho\, ww' \wedge u \sim u')\big)$$

which is equivalent to our desired property

$$\chi(w) := (\forall w')(\forall u)(\forall u') \\ ((D_\rho\, ww' \wedge w \sim u \wedge D_\rho\, uu') \rightarrow u \sim u')$$

By theorem 4.42 of [6], we know that **A6** is canonical for $\chi(w)$, i.e., the canonical frame for any normal modal logic containing **A6** has the property $\chi(w)$. In particular, M^{EI_K} has the property.

P5 We want to show that $\gamma \in Y^{\mathsf{EI}_K}(w)$ implies $M^{\mathsf{EI}_K}, w \models \gamma$. Suppose $\gamma \in Y^{\mathsf{EI}_K}(w)$; by definition of $Y^{\mathsf{EI}_K}(w)$, we have $I\,\gamma \in w$; by axiom **A7** and the modus ponens closure, $\gamma \in w$; by the Truth Lemma, $M^{\mathsf{EI}_K}, w \models \gamma$.

A.2. Proof of Proposition 4.2

We will show that $M_{+\gamma!} = (W', \sim', D'_\rho, V', Y')$ satisfy **P1**-**P5**.

P1 Suppose $D'_\rho wu$; we want to show that $(\text{prem}(\rho)+\rho) \subseteq Y'(w)$ and that $Y'(u) = Y'(w) + \text{conc}(\rho)$. If $D'_\rho wu$, then $w, u \in W'$ and $D_\rho wu$. Since M satisfy **P1**, we have $(\text{prem}(\rho) + \rho) \subseteq Y(w)$ and $Y(u) = Y(w) + \text{conc}(\rho)$. By definition of Y' and the fact that $w, u \in W'$, we have $(\text{prem}(\rho) + \rho) \subseteq Y'(w)$ and $Y'(u) = Y'(w) + \text{conc}(\rho)$.

P2 Suppose $D'_\rho wu$; we want to show that w, u satisfy the same propositional letters in M. Since $D'_\rho wu$, w and u are in W' and $D_\rho wu$. By property **P2** of M, we know that w and u satisfy the same propositional letters in M; by definition of V', w and u satisfy the same propositional letters in $M_{+\gamma!}$.

P3 Suppose $w_1 \sim' u_1$ and $D'_\rho w_1 w_2, D'_\rho u_1 u_2$ for some rule ρ; we want to show that $w_2 \sim' u_2$. By $w_1 \sim' u_1$, $D'_\rho w_1 w_2$ and $D'_\rho u_1 u_2$, we have $w_1 \sim u_1$, $D_\rho w_1 w_2$ and $D_\rho u_1 u_2$, with $w_1, w_2, u_1, u_2 \in W'$. By **P3** of M, $w_2 \sim u_2$; by definition of $\sim'$, we get $w_2 \sim' u_2$.

P4 It follows from the definition that if $\sim$ is an equivalence relation, so it is $\sim'$.

P5 Suppose $\gamma \in Y'(w)$; we want to show that $M', w \models \gamma$. If $\gamma \in Y'(w)$, then $w \in W'$ and $\gamma \in Y(w)$. By **P5** of M, we get $M, w \models \gamma$; then, by definition of V', $M', w \models \gamma$.

References

[1] T. Agotnes and N. Alechina, editors. *Proceedings of the Workshop on Logics for Resource-Bounded Agents, organised as part of the 18th European Summer School on Logic, Language and Information (ESSLLI)*, Malaga, Spain, August 2006.

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

[3] 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.

[4] A. Baltag and S. Smets. Conditional doxastic models: A qualitative approach to dynamic belief revision. In *Proceedings of the 13th Workshop on Logic, Language, Information and Computation (WoLLIC 2006)*, volume 165, pages 5–21, 2006.

[5] A. Baltag and S. Smets. Dynamic belief revision over multi-agent plausibility models. Available at http://www.vub.ac.be/CLWF/SS/loft.pdf, 2006.

[6] P. Blackburn, M. de Rijke, and Y. Venema. *Modal logic*. Cambridge University Press, New York, NY, USA, 2001.

[7] H. N. Duc. Logical omniscience vs. logical ignorance on a dilemma of epistemic logic. In *EPIA '95: Proceedings of the 7th Portuguese Conference on Artificial Intelligence*, pages 237–248, London, UK, 1995. Springer-Verlag.

[8] H. N. Duc. Reasoning about rational, but not logically omniscient, agents. *Journal of Logic and Computation*, 7(5):633–648, 1997.

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

[10] J. Gerbrandy. *Bisimulations on Planet Kripke*. PhD thesis, Institute for Logic, Language and Computation (University of Amsterdam), 1999.

[11] J. Y. Halpern and M. Y. Vardi. The complexity of reasoning about knowledge and time: Synchronous systems. Technical Report RJ 6097, IBM Almaden Research Center, 1988.

[12] M. Jago. *Logics for Resource-Bounded Agents*. PhD thesis, University of Nottingham, July 2006.

[13] M. Jago. Rule-based and resource-bounded: A new look at epistemic logic. In Agotnes and Alechina [1], pages 63–77.

[14] F. Liu. Diversity of agents. In Agotnes and Alechina [1], pages 88–98.

[15] F. Liu. *Changing for the Better. Preference Dynamics and Agent Diversity*. PhD thesis, Institute for logic, Language and Computation (Universiteit van Amsterdam), Amsterdam, The Netherlands, February 2008. ILLC Dissertation series DS-2008-02.

[16] J. A. Plaza. Logics of public communications. In M. L. Emrich, M. S. Pfeifer, M. Hadzikadic, and Z. W. Ras, editors, *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems*, pages 201–216, 1989.

[17] J. van Benthem. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics*, 14(2), 2004.

[18] J. van Benthem. Epistemic logic and epistemology: The state of their affairs. *Philosophical Studies*, 128:49–76, March 2006.

[19] J. van Benthem. Logic and reasoning: Do the facts matter? *Studia Logica special issue* "Psychologism in Logic?", 2008.

[20] J. van Benthem. Tell it like it is: Information flow in logic. *Journal of Peking University (Humanities and Social Science Edition)*, 1:80–90, 2008.

[21] J. van Benthem and E. Pacuit. The tree of knowledge in action. In G. Governatori, I. Hodkinson, and Y. Venema, editors, *Proceedings of Advances in Modal Logic, 2006 (AiML 2006)*. King's College Press, 2006.

Quasi-merging and Pure-arbitration on Information for the Family of Adaptive Logics ADM[*]

Giuseppe Primiero
Giuseppe.Primiero@UGent.be
Centre for Logic and Philosophy of Science, Ghent University, Belgium

Joke Meheus
Joke.Meheus@UGent.be
Centre for Logic and Philosophy of Science, Ghent University, Belgium

Abstract

The present paper introduces two new information merging protocols for the family of adaptive logics $\mathbf{ADM}$*, for which majority merging has been defined in previous work. The new adaptive operators reflect the negotiation processes of quasi-merging and pure-arbitration known from the Integrity Constraints framework. The Adaptive Variant Counting selection provides results equivalent to the GMax family of merging operators: it selects a collective model for a multi-set of belief bases established on the number of disagreements verified by the various models. The Adaptive Minimax Counting selection is a quasi-merging operator which applies a minimax function and it obtains a larger spectrum of possibilities than the previous selection: it simulates the behaviour of the Max family of operators from the Integrity Constraints framework, but it avoids some of its counterintuitive results.*

Keywords: Information Fusion, Negotiation Protocols, Arbitration, Quasi-Merging, Adaptive Logics.

1. Introduction

The analysis of processes of intelligent interaction in multi-agent systems has grown constantly in the logical literature of the last decade, with diversificated approaches and aims. The applications vary from the formalization of interactive processes of collective deliberation, especially relevant with respect to the formulation of judgement aggregation strategies, to information fusion architectures.

The analysis of contents involved in a decision process focuses naturally on the agreements among agents, in order to perform the most satisfactory selection of common goals and judgements in the group. Obviously, such a process might not be entirely satisfactory, and the presence of disagreements expressing a certain degree of internal dissatisfaction cannot be completely ruled out by the negotiation and the consequent aggregation protocols. The formalization of selection procedures in view of such inconsistent data is the aim of the frameworks defining knowledge merging operators, also known as information fusion operators.

The merging of contents from contradictory sources, whose study goes back to [5], has applications in distributed databases and information systems. General properties for the logical approaches to merging procedures for knowledge bases containing inconsistent information have been studied in [7], [6], [4], and more recently surveyed in [11] and [10].

The first definition of an operator for merging information has been given in [21] and later considerably reworked in [14] and [15]. In the latter work, the idea of *arbitration* comes from an intuitive modification of the more standard revision operator from the AGM-paradigm in [1]: it refers to merging as the revision of an older base with the information of a newer base, without any order of priority of the latter over the former. The process requires instead preservation of information from one base in some cases and from the other in other cases. This general principle has been modified by the use of weights on the bases, to indicate the relative importance of the information rather than strict priority. Weights have been expressed as priority values (as in [9]), they have been assigned either to propositional terms (see e.g. [8]) or to the set of models of formulas (as in [21]), and finally they have been formulated as possibility values (see [23]).

A major distinction has been introduced by the defini-

[*] Research for this paper was supported by subventions from Ghent University and from the Research Foundation - Flanders (FWO - Vlaanderen).

tion of the *majority* protocol. In [16] an operator is defined which avoids the typical restrictions of the majority principle, formulated taking into account formulae in disjunctive form within the bases to be merged. This allows to represent bases that only partially support their contents. The ground distinction between the arbitration and majority operators - see e.g. [13] - can be reflected in the following terms: whereas majority merging operators aim at minimizing collective dissatisfaction, arbitration operators aim at maximizing individual satisfaction. This distinction is of the greatest importance with respect to the results of collective deliberation procedures and the restrictions due to results as the one of the judgement aggregation paradox.

These two sub-classes of merging operators are further defined in the more general and standard framework of information merging under intergrity constraints in [12]. This framework allows for defining families of three distinct operators:

1. the $\triangle^{\Sigma}$ operator satisfies the postulates for majority merging and it corresponds to the merging operator defined in [16];
2. the $\triangle^{GMax}$ operator satisfies a pure arbitration procedure, and it represents a new merging method;
3. finally, the $\triangle^{Max}$ operator is called a quasi-merging operator and it represents a pseudo-arbitration operator corresponding to the one defined in [15].

Moreover, in [13] it is shown that another family of operators called $\triangle^{n}$ can be defined which belongs simultaneously to the two main subclasses.

The standard approach to these protocols uses a definition of distance between the involved belief bases and the possible interpretations. The standard one is the Dalal distance from [8]: the intuitive idea behind this definition is to measure the number of atoms that have different truh values among each base and every interpretation, so to find the collective model that retains the most of each base; a variant is represented by the Satoh distance, defined in [22]. The various merging protocols apply an ordering on the values resulting from the definition of distance according to different functions, in order to obtain the desired negotiation process.

A different approach to the resolution of merging processes of conflicting belief bases has been introduced in [17] in view of the dynamic semantics of adaptive logics (see [2, 3] for a general introduction to the standard format of Adaptive Logics). The crucial change of perspective given by this new approach is represented by the focus on disagreements occurring among the agents involved in the negotiation process: the explicit derivation of conflicts in the collective decision process allows for the formulation of a consequence set that reflects the various aggregation methods in terms of unavoidable disagreements. The resulting framework is the family of logics $\mathbf{ADM}$, for *Adaptive Doxastic Merging*.

The first effective result obtained for adaptive merging is the majority protocol for bases with partial support defined in terms of the logic $\mathbf{ADM}^{c}$, for *Adaptive Doxastic Merging by Counting*, formulated in [20]. As it is shown in [16], the protocol which satisfies all the due postulates for majority has to take into account the requirements on partially supported contents, and the family $\mathbf{ADM}$ makes use of so called abnormal formulas that are designed precisely to accomplish this aim. The Counting strategy selects from the set of models of a given premise set, providing a protocol of majority merging corresponding to the generalization under Integrity Constraints represented by the $\triangle^{\Sigma}$ operator. Moreover, the use of fully versus partially supported contents allows for the mentioned notion of weights to be reformulated in a new light: weights express the support each agent gives to contents, in order for his or her beliefs to be accepted by the group in the fusion procedure. This allows for commutativity to be entirely preserved also among weighted bases.

The application of the majority protocol to the judgment aggregation paradox is considered in [18]: it provides a non-paradoxical though inefficient solution. Effectivity is obtained by modifying the agenda of interaction, which in turn amounts to give up the Universal Domain condition. To this aim the formulation of the logic $\mathbf{ADM}^{c}$ is slightly more simple in view of the fact that all bases in the case of the paradox express full support to their contents.

The next step in this resarch is represented by the formulation of an arbitration protocol for the family of logics $\mathbf{ADM}$, mimicking the results of the $\triangle^{GMax}$ operator. This result is first presented in [19], where the logic $\mathbf{ADM}^{c+}$ for *Adaptive Doxastic Merging by Variant Counting* is introduced. The semantic selection defined for this adaptive logic in standard format shows a basic correspondance between a pre-order on satisfied disagreements among the agents and a lexicographic order of Dalal's distances. The formulation of the arbitration protocol is considered in the light of the problem of fusion of information from heterogeneous databases: the fusion architecture based on this protocol shows its potential in applications where the treatment of incomplete or only partially verified data might be crucial to an effective fusion procedure.

In the present paper we shall recover the basics of the logic $\mathbf{ADM}^{c+}$ and of its selection procedure. Starting from its basis, it will be possible to define a third selection procedure on the models of a premise set, derived from the minimax rule for decision theory, and thus performing the same results as the $\triangle^{Max}$ quasi-merging operator. This selection procedure shall be introduced as the logic $\mathbf{ADM}^{c-mm}$, for *Adaptive Doxastic Merging by Minimax Counting*. With this last result the family of logics $\mathbf{ADM}$ is shown to be a general framework to define all the various negotiation processes modelled by the standard merging operators, in particular those of the general Integrity Constraints framework.

The structure of this paper is as follows. In section 2 we will consider briefly the quasi-merging and arbitration Integrity Constraints operators. In section 3 we will introduce the preliminaries needed for the adaptive logics of the family $\mathbf{ADM}$, and in section 4 and 5 we will respectively define the semantic selection procedures that give rise to the logics $\mathbf{ADM}^{c+}$ and $\mathbf{ADM}^{c-mm}$. Section 6 presents a standard example where both strategies are applied. In the final section further steps for the research on the adaptive procedures of merging are surveyed.

2. Integrity Constraints Merging

In this section we introduce the Integrity Constraints (IC) merging protocols from [12] that are going to be mimicked by different strategies in the $\mathbf{ADM}$ family of adaptive logics. In the following of this paper $\mathcal{L}$ will refer to the standard language of classical propositional logic (henceforth $\mathbf{CL}$) that is formed from a finite set of atoms $\mathcal{P}$ in the usual way. The set of literals $\mathcal{P}^{\pm}$ contains atoms and their negations. Letters from the greek alphabet $\varphi, \psi, \ldots$ are used as metavariables for sentences of $\mathcal{L}$. As is common, the abbreviation $\bigvee(\Delta)$ will stand for the disjunction of the members of Δ, where Δ is a set of formulas. A *belief base* T is a finite set of sentences of $\mathcal{L}$. Integrity constraints μ are a finite set of sentences, i.e. a belief base with respect to which the final merged state must be consistent. By Γ one refers to a multi-set consisting of n belief bases, $\Gamma = \{T_1, \ldots, T_n\}$. The formula $\bigwedge \Gamma$ denotes the conjunction of the belief bases of Γ, i.e. $\bigwedge \Gamma = \bigwedge\{T_1, \ldots, T_n\}$. A $\mathbf{CL}$-model is a function $\mathcal{P} \to \{0, 1\}$. We shall use $\mathcal{M}$ to denote the set of all $\mathbf{CL}$-models. A model M is a model of T iff all the members of T are true in it. $Mod(\Gamma)$ will be the set of models of the multi-set Γ and $Cn(\Gamma)$ will denote the consequence set of Γ. The result of a merging procedure on a multi-set Γ under constraints μ shall be denoted as $\triangle_{\mu}(\Gamma)$. The union of multi-sets will be denoted by $\sqcup$.

2.1 IC Pure-Arbitration

The IC framework defines selection methods of the collective models of various belief bases by operators satisfying the following postulates:

IC0 $\triangle_{\mu}(\Gamma) \vdash \mu$;

IC1 If μ is consistent, then $\triangle_{\mu}(\Gamma)$ is consistent;

IC2 If $\bigwedge \Gamma$ is consistent with μ, then $\triangle_{\mu}(\Gamma) = \bigwedge \Gamma \wedge \mu$;

IC3 If $\Gamma_1 \leftrightarrow \Gamma_2$ and $\mu_1 \leftrightarrow \mu_2$, then $\triangle_{\mu_1}(\Gamma_1) \leftrightarrow \triangle_{\mu_2}(\Gamma_2)$;

IC4 If $T \vdash \mu$ and $T' \vdash \mu$, then $\triangle_{\mu}(T \sqcup T') \wedge T \nvdash \mu$ implies that $\triangle_{\mu}(T \sqcup T') \wedge T' \nvdash \mu$;

IC5 $\triangle_{\mu}(\Gamma_1) \wedge \triangle_{\mu}(\Gamma_2) \vdash \triangle_{\mu}(\Gamma_1 \sqcup \Gamma_2)$;

IC6 $\triangle_{\mu}(\Gamma_1) \wedge \triangle_{\mu}(\Gamma_2)$ is consistent, then $\triangle_{\mu}(\Gamma_1 \sqcup \Gamma_2) \vdash \triangle_{\mu}(\Gamma_1) \wedge \triangle_{\mu}(\Gamma_2)$;

IC7 $\triangle_{\mu_1}(\Gamma) \wedge \mu_2 \vdash \triangle_{\mu_1 \wedge \mu_2}(\Gamma)$;

IC8 If $\triangle_{\mu_1}(\Gamma) \wedge \mu_2$ is consistent, then $\triangle_{\mu_1 \wedge \mu_2}(\Gamma) \vdash \triangle_{\mu_1}(\Gamma) \wedge \mu_2$.

On the basis of these postulates, the consistency on merging and the irrelevance of syntax are principles of the greatest importance to define quasi-merging and pure-arbitration. The former is given in the following informal definition:

Definition 2.1 [Principle of Consistency on Merging] If two subgroups agree on at least one alternative, the result of global merging will be exactly those alternatives the two groups agree on. $\triangleleft$

and it is formally obtained by the combination of postulates **IC5** and **IC6**. The principle of syntax irrelevance says informally:

Definition 2.2 [Principle of Syntax Irrelevance on Merging] If two bases are syntactically equivalent and so are their integrity constraints, then the merging of one base under one set of integrity constraints shall be equivalent to the merging of the other base under the other set of constraints. $\triangleleft$

and it is formally given by postulate **IC3**.

For the introduction and explanation of the pure-arbitration and quasi-merging protocols we will refer to a preorder on the set of models of a premise set. A preorder over the set of $\mathbf{CL}$-models is a reflexive and transitive relation on $\mathcal{M}$. Where $\leq$ is a preorder, $<$ is defined as: $M < M'$ iff $M \leq M'$ and $M' \nleq M$. Where M is a subset of $\mathcal{M}$, a model M is said *minimal* in M with respect to $\leq$ iff $M \in \mathsf{M}$ and there is no $M' \in \mathsf{M}$ such that $M' < M$. $Min(\mathsf{M}, \leq)$ shall denote the set of models that are minimal in M with respect to $\leq$.

Given two models M_1, M_2 and a belief base T, a preorder $M_1 \leq M_2$ holds if and only if $dist(M_1, T) \leq dist(M_2, T)$. The value of $dist(M_1, M_2)$ between two models M_1 and M_2 according to the Dalal distance refers to the number of atoms whose valuation differs in the two models. Given the set $Mod(T)$ of possible models of the base T, the distance between a $\mathbf{CL}$-model M and T is given as follows:

$$dist(M,T) = min(dist(M,M')) \text{ for each } M' \in Mod(T). \qquad (1)$$

The selection of collective models $Mod(\triangle_{\mu}^{GMax}(\Gamma))$ performed according to the IC arbitration operator using this notion of distance works in the following way. Consider belief bases T_1, T_2 whose alternatives are preferred respectively under Integrity Contraints μ_1, μ_2; assume that each of the set of alternatives is equally preferred under the union of the bases $T_1 \sqcup T_2$; the subset of preferred alternatives under the disjunction of the integrity constraints coincides with the preferred alternatives of each base. Model-theoretically this means that there is a total preorder on the plausibility of the models with respect to the belief bases. Plausibility is obtained as an ordering by a notion of distance as Dalal's one and an aggregation function $\oplus$. Such an ordering says that if M_1 is more plausible than M_2 for T_1 and more plausible than M_3 for T_2, and M_2 and M_3 are equally plausible for the union of bases $T_1 \sqcup T_2$, than M_1 has to be more plausible than both M_2 and M_3 for $T_1 \sqcup T_2$. The result of the merging procedure is the belief base whose models are the most plausible ones for the given set of individual bases, according to given rationality criteria.

In [12], it is shown that the aggregation function $\oplus$ satisfying the arbitration protocol is the *leximax* function. Consider the multi-set $\Gamma = \{T_1, \ldots, T_n\}$; for each model M consider the list $D = (dist_1^M, \ldots, dist_n^M)$ of distances between M and the n belief bases in Γ, i.e. the list of distances $dist_i^M = dist(M, T_i)$. Let L_Γ^M be the list obtained from D by sorting its members in descending order. Denote now by $\leq_{lex}$ the lexicographic order among sequences of integers of the same length. For any two models M_1 and M_2, a total preorder $M_1 \leq_\Gamma M_2$ holds in view of Γ if and only if $L_\Gamma^{M_1} \leq_{lex} L_\Gamma^{M_2}$. Given a multi-set Γ holding under contraints μ, the $\triangle_\mu^{GMax}$ operator is then defined as follows:

$$Mod(\triangle_\mu^{GMax}(\Gamma)) = Min(Mod(\mu), \leq_\Gamma). \qquad (2)$$

This operator satisfies the typical postulate for arbitration:

$$\left.\begin{array}{r} \triangle_{\mu_1}(T_1) \leftrightarrow \triangle_{\mu_2}(T_2) \\ \triangle_{\mu_1 \leftrightarrow \neg\mu_2}(T_1 \sqcup T_2) \leftrightarrow (\mu_1 \leftrightarrow \neg\mu_2) \\ \mu_1 \nvdash \mu_2 \\ \mu_2 \nvdash \mu_1 \end{array}\right\} \begin{array}{l} \Rightarrow \triangle_{\mu_1 \vee \mu_2} \\ (T_1 \sqcup T_2 \leftrightarrow \\ \triangle_{\mu_1}(T_1)) \end{array} \qquad (3)$$

which says that if a set of alternatives preferred among one set of integrity constraints μ_1 for a belief base T_1 corresponds to the set of alternatives preferred among another set of integrity constraints μ_2 for base T_2, and if the alternatives that belong to a set of integrity constraints but not to the other are equally preferred for the whole group $(T_1 \sqcup T_2)$, then the subset of preferred alternatives among the disjunction of integrity constraints coincides with the preferred alternatives of each base among their respective integrity constraints (see [12], p.778).

2.2 IC Quasi-Merging

The second family of merging operators considered is a less fine-grained one and it is defined by the so-called quasi-merging $\triangle^{Max}$ operator in terms of the $minimax$ function. Let $\Gamma = \{T_1, \ldots, T_n\}$ be the usual belief set, M a model and d the standard Dalal's distance value. The Max operator considers first the maximal distance between an interpretation and a belief base

$$d_{Max}(M, \Gamma) = Max_{T \in \Gamma} dist(M, T); \qquad (4)$$

then a preorder on the set of interpretations M is defined:

$$M_1 \leq_\Gamma^{Max} M_2 \text{ iff } d_{Max}(M_1, \Gamma) \leq d_{Max}(M_2, \Gamma) \qquad (5)$$

which says that a model M_1 comes before in the preorder than a model M_2 if and only if the maximal distance between the former and the multi-set Γ is lower than the same distance between the latter and Γ. The resulting $\triangle_\mu^{Max}(\Gamma)$ operator is obtained as the one with lower position (minimal value) in the obtained pre-order:

$$Mod(\triangle_\mu^{Max}(\Gamma)) = Min(Mod(\mu), \leq_\Gamma^{Max}). \qquad (6)$$

In the following sections we shall introduce the Adaptive Logic **ADM** with two adaptive strategies, namely *Variant Counting* and *Minimax Counting*: their role is to formulate adaptive merging procedures whose results are comparable to those of the $\triangle^{GMax}$ and $\triangle^{Max}$ operators.

3. The Adaptive Logic for Merging

The formulation of the logics belonging to the family **ADM** is based on the langugae $\mathcal{L}^B$, which enables one to represent a *set* of belief bases by a single set of premises. It also enables one to consider (modal) models that validate all the premises, rather than having to consider models for each of the belief bases separately. Where $\mathcal{I} = \{0, 1, \ldots\}$ is a set of indexes, the multi-modal language $\mathcal{L}^{\mathcal{B}}$ is $\mathcal{L}$ extended with a belief operator b_i, for any $i \in \mathcal{I}$. Each different base is given an index b_i with $i \in \mathcal{I} \setminus 0$. The operator b_0 is used exclusively for the beliefs selected for the merging state, or for the constraints holding in such state. Intuitively, $b_i\varphi$ (for $i > 0$) will express that agent i believes or supports φ; the formula $b_0\varphi$ means that all agents agree on φ or that their decision is constrained by the holding of φ. The premise set Γ refers to a multi-set of indexed belief bases $\Gamma = \{T_1, \ldots, T_n\}$. When the two adaptive strategies are introduced, the operator $\triangle^{c+}$ (eventually $\triangle_\mu^{c+}$ when some set of constraints μ is given) is used for the result of the Variant Counting strategy and $Mod(\triangle^{c+}(\Gamma))$ to refer to the subset of $Mod(\Gamma)$ correspondingly selected; the operator $\triangle^{c-mm}$ ($\triangle_{\mu}^{c-mm}$ respectively) is used for the result obtained by the Minimax Counting Strategy, $Mod(\triangle^{c-mm}(\Gamma))$ referring to the subset of $Mod(\Gamma)$ selected by that strategy.

Let us consider as an example a set of belief bases

$$\begin{array}{l} T_1 = \{p \vee q\} \\ T_2 = \{\neg p\} \\ T_3 = \{\neg q\}. \end{array}$$

These belief bases are given a modal translation in any of the logics belonging to the family **ADM** as the premise set $\Gamma = \{b_1(p \vee q), b_2\neg p, b_3\neg q\}$. This means that for any $T_i \models \phi$, in a **DM** premise set there is a doxastic formula $b_i\phi$ holding in $\mathcal{L}^B$. A literal φ is *fully supported* by some belief base T if $T \vDash \varphi$. A literal φ is *partially supported* by a belief base T if there is a set of literals Δ such that $\varphi \in \Delta, T \vDash \bigvee(\Delta), \nvDash \bigvee(\Delta)$, and there is no $\Delta' \subset \Delta$ such that $T \vDash \bigvee(\Delta')$. As usual $\bigvee(\Delta)$ stands for the disjunction of the members of the set of literals Δ. So, for the previous premise set where $T_1 = \{p \vee q\}$, T_1 partially supports p and q; whereas T_2 fully supports $\neg p$ and T_3 fully supports $\neg q$.

All the logics belonging to the family **ADM** are adaptive logics in standard format. This format is extensively discussed in [3]. They all share the same first element needed for their definition, i.e. the lower limit logic (**LLL**); they all share the same second element in the definition, i.e. the set of abnormal formulas; and they all differ for the last element, i.e. the adaptive strategy which selects the abnormal models holding for a given premise set.

The basis of the adaptive logics of the **ADM** family is the so-called lower limit logic **DM**: this is a multi-modal version of the modal logic **D**. In addition to all **CL**-axioms, the logic **DM** validates

- Necessitation Rule: if $\vdash_{\mathbf{CL}} \varphi$ then $\vdash_{\mathbf{DM}} b_i\varphi$;
- Distribution: $b_i(\varphi \supset \psi) \supset (b_i\varphi \supset b_i\psi)$;
- Consistency: $b_i\varphi \supset \neg b_i\neg\varphi$.

Semantically, the models of each logic ($\mathbf{ADM}^c$, $\mathbf{ADM}^{c+}$, $\mathbf{ADM}^{c-mm}$) of a given premise set Γ are obtained by making a selection of the **DM**-models of Γ. This selection will establish the valid models, and the contents of the corresponding consequence sets are the result of the negotiation procedure.

The semantics of the lower limit logic **DM** is a standard possible world semantics, with multiple accessibility relations. A **DM**-model is a quadruple $\langle \mathcal{W}, w_o, \mathcal{R}, v\rangle$ where $\mathcal{W}$ is a set of possible worlds, $w_o \in \mathcal{W}$ is the actual world, $\mathcal{R}$ is a set of serial accessibility relations R_i ($i \in \mathcal{I}$) over $\mathcal{W}$, and $v : \mathcal{P} \times \mathcal{W} \rightarrow \{0, 1\}$ is an assignment function.

The valuation function defined by a model M is characterized as follows:

C1 where $A \in \mathcal{P}$, $v_M(A, w) = v(A, w)$;

C2 $v_M(\neg A, w) = 1$ iff $v_M(A, w) = 0$;

C3 $v_M(A \vee B, w) = 1$ iff $v_M(A, w) = 1$ or $v_M(B, w) = 1$;

C4 $v_M(A \wedge B, w) = 1$ iff $v_M(A, w) = 1$ and $v_M(B, w) = 1$;

C5 $v_M(A \supset B, w) = 1$ iff $v_M(A, w) = 0$ or $v_M(B, w) = 1$;

C6 $v_M(b_i\varphi, w) = 1$ iff $v_M(\varphi, w') = 1$ for all w' such that R_iww'.

The standard semantic notions are defined as usual: a model M verifies A iff $v_M(A, w_0) = 1$, $\Gamma \models_{\mathbf{DM}} A$ iff all **DM**-models of Γ verify A, and $\models_{\mathbf{DM}} A$ iff all **DM**-models verify A.

In order to establish which contents of the premise set are finally merged, the adaptive machinery formulates all the disagreements that occurr in view of each agent's belief base. These are formalised in terms of a special class of formulas, called *abnormalities*, that are eventually verified by some models of the given premise set in the lower limit logic. In the case of the previously introduced premise set $\Gamma = \{b_1(p \vee q), b_2\neg p, b_3\neg q\}$, and in view of the fact that one tries to merge as much as possible of its content, some of the **DM**-models of Γ verify the formula $b_3\neg q \supset b_0\neg q$, whereas others falsify it; or, what comes to the same, verify $b_3\neg q \wedge \neg b_0\neg q$. An abnormality is precisely a formula of the form

$$b_i\varphi \wedge \neg b_0\varphi \tag{7}$$

i.e. a formula expressing a (full) support by some agent i for a literal φ which is not merged in view of someone's disagreement. In all **DM**-models of Γ, at least one instance of such an abnormality is verified. In a simple example, where $\Gamma = \{b_1p, b_2p, b_3\neg p\}$, there will be two types of **DM**-models: those that verify b_0p and those that verify $\neg b_0p$. Models that verify b_0p, necessarily verify the abnormality $b_3\neg p \wedge \neg b_0p$; those that verify $b_0\neg p$ necessarily verify $b_1p \wedge \neg b_0p$ and $b_2p \wedge \neg b_0p$. The selection tells us which type of models should be chosen.

As one is considering arbitration on bases that express also partial support, among the **DM**-models of Γ there are models verifying a different kind of abnormalities. As far as an abnormality with respect to T_1 from the previous example is concerned, one has to account for the rejection of a partially supported content. An abnormality involving a base expressing partial support might be due to conflicts arising with respect to each partially supported literal. This is formulated in the following form:

$$b_i(\varphi_1 \vee \ldots \vee \varphi_n) \wedge (\neg b_i\varphi_1 \wedge \ldots \wedge \neg b_i\varphi_n) \wedge \neg b_0(\varphi_1 \vee \ldots \vee \varphi_n) \tag{8}$$

where all φ_i are literals. The union of sets of disagreements for fully and partially supported literals will form our set of abnormalities:

Definition 3.1 [Set of Abnormalities] $\Omega = \{b_i\varphi \wedge \neg b_0\varphi \mid i \in \mathcal{I} \setminus 0, \varphi \in \mathcal{P}^{\pm}\} \cup \{b_i(\varphi_1 \vee \ldots \vee \varphi_n) \wedge (\neg b_i\varphi_1 \wedge \ldots \wedge \neg b_i\varphi_n) \wedge \neg b_0(\varphi_1 \vee \ldots \vee \varphi_n) \mid i \in \mathcal{I} \setminus 0, \varphi_1, \ldots, \varphi_n \in \mathcal{P}^{\pm}, n > 1\}$. ◃

In each adaptive logic obtained by the lower limit logic $\mathbf{DM}$, a disjunction of abnormalities may be $\mathbf{DM}$-derivable without any of its disjuncts being $\mathbf{DM}$-derivable. Consider again $\Gamma = \{b_1(p \vee q),\ b_2\neg p,\ \ b_3\neg q\}$. From this, neither $b_1p \wedge \neg b_0p$ nor $b_2\neg p \wedge \neg b_0\neg p$ is $\mathbf{DM}$-derivable, but the disjunction $(b_1p \wedge \neg b_0p) \vee (b_2\neg p \wedge \neg b_0\neg p)$ is. Disjunctions of abnormalities will be called *Dab*-formulas, and the abbreviation $Dab(\Delta)$ is used to refer to them:

Definition 3.2 [Disjunctions of Abnormalities] $Dab(\Delta)$ stands for $\bigvee(\Delta)$ where $\Delta \subseteq \Omega$. ◃

If Δ is a singleton, $Dab(\Delta)$ is a single abnormality; if $\Delta = \emptyset$, any disjunction $A \vee Dab(\Delta)$ corresponds to A. A *Dab*-formula that is $\mathbf{DM}$-derivable from Γ will be called a *Dab*-consequence of Γ:

Definition 3.3 [*Dab*-Consequence] $Dab(\Delta)$ is a *Dab*-consequence of a premise set Γ iff $\Gamma \models_{\mathbf{DM}} Dab(\Delta)$. ◃

If $Dab(\Delta)$ is a *Dab*-consequence of a set Γ, then so is any $Dab(\Delta')$ such that $\Delta' \supset \Delta$. This is why a further definition is needed:

Definition 3.4 [Minimal *Dab*-Consequence] A disjunction of abnormalities $Dab(\Delta)$ is a *minimal Dab*-consequence of Γ iff $\Gamma \models_{\mathbf{DM}} Dab(\Delta)$ and there is no $\Delta' \subset \Delta$ such that $\Gamma \models_{\mathbf{DM}} Dab(\Delta')$. ◃

It is in view of the derivability of *Dab*-formulas from a premise set that the *adaptive strategy* is needed. Intuitively, the adaptive strategy specifies what it means, in the case of disjunctions of abnormalities, that the abnormalities are false *unless and until proven otherwise*. Given the same lower limit logic and the same set of abnormalities, there are different ways to interpret a set of premises *as normally as possible*: the precise interpretation of this ambiguous phrase is determined by the adaptive strategy. In the present case, one will distinguish between the interpretation of a premise set as normally as possible in view of the Variant Counting Strategy in $\mathbf{ADM}^{c+}$, and the interpretation in view of the Minimax Counting Strategy in $\mathbf{ADM}^{c-mm}$.

4. Variant Counting for Arbitration

The selection by variant Counting is applied to the consequence set of the lower limit logic $\mathbf{DM}$ and it gives rise to the adaptive logic $\mathbf{ADM}^{c+}$. It considers the various *Dab*-consequences of a premise set Γ in view of the number of disagreements involving each agent. This corresponds to a selection of the formulas verified in any given model on the basis of the number of contents held true by each agent and involving a disagreement with another agent.

Consider first all the formulas $A \in Dab(\Delta)$ such that $\Gamma \models_{\mathbf{DM}} Dab(\Delta)$ and the b-operator indexed 1 occurs in A: typically, this will be the set of all the abnormalities derivable from a premise set Γ that are of the form $b_1\phi \wedge \neg b_0\phi$ or of the form $b_1(\phi_1 \vee \ldots \vee \phi_n) \wedge (\neg b_1\phi_1 \wedge \ldots \wedge \neg b_1\phi_n) \wedge \neg b_0(\phi_1 \vee \ldots \vee \phi_n)$. Call this set Ω^1. Then consider the set of all formulas of the same kind occurring with b-operator indexed 2 and call this set Ω^2, and so on up to index n. The set Ω is in turn the union of all the various Ω^i sets:

Definition 4.1 [The set of indexed abnormalities]

$$\Omega = \bigcup_{i=1}^{n} \Omega^i. \tag{9}$$

◃

It is obvious that one can consider now the set of abnormalities with a given index as a proper subset of Ω.

For each model of a given premise set, consider now the abnormal formulas of a certain Ω^i verified by that model:

Definition 4.2 [The abnormal part of a model with index i] $Ab^i(M) = \{A \mid A \in \Omega^i \text{ and } M \models A\}$. ◃

For any model M_j of a given premise set, let $\mathcal{C}^i_{M_j} = |Ab^i(M_j)|$ denote the cardinality of its abnormal part with respect to Ω^i:

Definition 4.3 [Abnormal cardinality of a model] Given a model M_j of a premise set Γ and its abnormal part $Ab^i(M_j)$, its abnormal cardinality $\mathcal{C}^i_{M_j}$ is the number of abnormal formulas $A \in \Omega^i$ verified in the model M_j. ◃

The abnormal cardinality $\mathcal{C}^i_{M_j}$ expresses the number of disagreements that agent i faces with respect to the literals verified by the model M_j. For each model M, we construct the list $(\mathcal{C}^1_M, \ldots, \mathcal{C}^n_M)$, where n is the number of elements of $\mathcal{I}$. Let L^M_Γ be the list obtained by $(\mathcal{C}^1_M, \ldots, \mathcal{C}^n_M)$ by sorting its elements in descending order. Let now $\leq_{lex}$ be the lexicographic order between sequences of integers of the same length. On the basis of the ordering $\leq_{lex}$, a total preorder $\leq^{\mathcal{C}}_{\Gamma}$ holds among the various models $M_1, \ldots, M_n$ of Γ in the following way:

Definition 4.4 [Preorder by Minimal Abnormal Cardinality] A total preorder $\leq^{\mathcal{C}}_{\Gamma}$ holds between models of a premise set Γ according to the following definition

$$M_i \leq^{\mathcal{C}}_{\Gamma} M_j \text{ iff } L^{M_i}_{\Gamma} \leq_{lex} L^{M_j}_{\Gamma}. \tag{10}$$

◁

According to this definition, the pre-order on the models of a premiset set Γ is obtained by ordering models according to their abnormal cardinalities. Where M_{Γ} stands for the set of $\mathbf{DM}$-models of Γ, the Variant Counting strategy $\triangle^{c+}(\Gamma)$ will select among those models the minimal ones with respect to the ordering obtained by $\leq^{\mathcal{C}}_{\Gamma}$:

Definition 4.5 [Selection of Models by $\mathbf{ADM}^{c+}$]

$$Mod(\triangle^{c+}(\Gamma)) = Min(\mathsf{M}_{\Gamma}, \leq^{\mathcal{C}}_{\Gamma}). \tag{11}$$

◁

The definition of the merging operator $\triangle^{c+}$ reflects a selection of abnormal models of the premise set that corresponds to the models satisfying the median possible choices that are preferred. In terms of the fair syncretic assignment presented in [13], the arbitration protocol satisfies the following conditions: the models of the premise set are the more plausible interpretations for the pre-order associated to that set; two equivalent knowledge sets have the same associated pre-orders. In the case of the adaptive selection this means that the abnormal models of a premise set selected by $\triangle^{c+}$ are those with lower position in the pre-order given by abnormal cardinalities and that two equivalent premise sets have the same pre-orders of abnormal cardinalities for their models.

The main condition of arbitration as fair syncretic assignment is satisfied as follows: if the ordering of abnormal cardinalities of M_i for base T_1 is lower than that of M_j for the same base (i.e. $M_i <^{\mathcal{C}}_{T_1} M_j$) and the same holds for M_i with respect to M_k for T_2 (i.e. $M_i <^{\mathcal{C}}_{T_2} M_k$), and if M_j and M_k are equally abnormal for $T_1 \sqcup T_2$ (i.e. $M_j \simeq^{\mathcal{C}}_{T_1 \sqcup T_2} M_k$), then M_i is less abnormal than M_j and M_k for $T_1 \sqcup T_2$ (i.e. $M_i <^{\mathcal{C}}_{T_1 \sqcup T_2} M_{j,k}$). Correspondingly, the following principle is formulated:

Definition 4.6 [Arbitration by Ordering on Abnormal Cardinalities] If for models M_i, M_j, M_k holds that $|Ab^i(M_i)| < |Ab^i(M_j)|$ and $|Ab^i(M_i)| < |Ab^i(M_k)|$; and if $|Ab^i(M_j)| = |Ab^i(M_k)|$; then $(M_i) <^{\mathcal{C}}_{\Gamma} M_{j,k}$ and $M(\triangle^{c+}(\Gamma)) = M_i$.

$$\left.\begin{array}{r} |Ab^i(M_i)| <_{T_1 \in \Gamma} |Ab^i(M_j)| \\ |Ab^i(M_i)| <_{T_2 \in \Gamma} |Ab^i(M_k)| \\ |Ab^i(M_j)| \simeq_{T_1 \sqcup T_2 \in \Gamma} |Ab^i(M_k)| \end{array}\right\} \Rightarrow M_i <^{\mathcal{C}}_{\Gamma} M_{j,k}.$$

◁

5. Minimax Counting for Quasi-Merging

The pseudo-arbitration operator from [15] has the main property of being constrained to only two bases and to require consistency to be obtained without the principle of average on bases to be preserved. This means that the negotiation procedure is performed among the belief bases rather than among the propositional letters having different truth values. If the operator is applied to two bases that support only respectively inconsistent literals, it will provide their disjunction without taking into account any combination of consistent contents. The $\triangle^{Max}$ operator from [12] is meant to model the very same procedure of arbitration, without the restriction imposed on the number of belief bases involved in the negotiation process. This operator is a less fine-grained one than the $\triangle^{GMax}$, because it provides a larger spectrum of possible results, and therefore it is called a quasi-merging operator.

In the present section a new adaptive semantic selection for $\mathbf{ADM}$ is introduced: it is called *Minimax Counting*, it gives rise to the adaptive logic $\mathbf{ADM}^{c-mm}$ and it aims at providing the same kind of negotiation process that is reflected by the result of the $\triangle^{Max}$ operator. The resulting $\triangle^{c-mm}$ operator for the Minimax Adaptive Counting applies the minimax rule to the selection of $\mathbf{DM}$-models of a premise set in view of their abnormal cardinality. The *Minimax Counting* selection presents an important difference with the standard $\triangle^{Max}$ operator: the latter, as any IC merging operator, does not satisfy the Majority Independence postulate (see [12], p.779). This postulate states that the result of merging is fully independent of the popularity of the views and it simply takes into account each different view:

$$\forall n \, \triangle_{\mu} \, (\Gamma \sqcup \Gamma'_1, \ldots, \Gamma'_n) \leftrightarrow \triangle_{\mu}(\Gamma \sqcup \Gamma'). \tag{12}$$

From this follows that the $\triangle^{Max}$ operator does not satisfy the **IC6** postulate, which togheter with its counterpart the **IC5** postulate allows the merging to satisfy always the alternatives for which there is no disagreement (consistency). The selection performed according to $\triangle^{Max}$ provides therefore a range of alternatives that contains also some of the choices for which none of the agents has expressed explicit preference. On the other hand, $\triangle^{c-mm}$ is based on the formulation of disagreements and their selection: anything which is not involved in any disagreement is obviously merged. This restricts slightly the range of results offered by the previous pseudo-arbitration operator (because it avoids some counter-intuitive results), but it still provides a larger spectrum of possibilities than the $\triangle^{c+}$ operator by using the minimax function.

The selection still makes use of the notion of abnormal cardinality $\mathcal{C}^i_{M_j}$ as given in Definition 4.3; it moreover refers for each model M to the list L^M_Γ obtained by ordering in descending order the list $(\mathcal{C}^1_M, \ldots, \mathcal{C}^n_M)$, where n is the number of elements of $\mathcal{I}$. A new maximal distance Max is defined as the first element in each list L^M_Γ for each model M:

Definition 5.1 [Maximal Abnormal Distance] $Max(M, \Gamma) = \mathcal{C}^i_M$ and there is no index k such that $|Ab^k(M)| > |Ab^i(M)|$. ◁

i.e. the highest of the values $\mathcal{C}^i$ for each model M and the first element of each L^M_Γ list. The Maximal Abnormal Distance expresses the highest number of disagrements possible in each abnormal model for any given agent. On its basis one derives a new total pre-order for the abnormal models in the following way:

Definition 5.2 [Preorder by Maximal Abnormal Distance] A total preorder $\leq^{Max}_\Gamma$ holds between models of a premise set Γ according to the following definition

$$M_i \leq^{Max}_\Gamma M_j \text{ iff } Max(M_i, \Gamma) \leq Max(M_j, \Gamma). \tag{13}$$

◁

Where M_Γ stands for the set of $\mathbf{DM}$-models of Γ, the Minimax Counting strategy of $\mathbf{ADM}^{c-mm}$ will select the minimal models with respect to the ordering obtained by $\leq^{Max}_\Gamma$:

Definition 5.3 [Selection of Models by $\mathbf{ADM}^{c-mm}$]

$$Mod(\triangle^{c-mm}(\Gamma)) = Min(\mathsf{M}_\Gamma, \leq^{Max}_\Gamma). \tag{14}$$

◁

The result of this selection is therefore obtained by restricting the $\mathbf{DM}$-models to their highest abnormal part and then selecting those that verify the minimal number of disagreements. In this way the result expresses a negotiation procedure that accounts for all the possible consistent combinations of contents, in view of full agreements and constraints.

6. An Example

The application of the various IC operators is shown in [12] in terms of an example which will now be considered for the operators $\triangle^{c+}$ and $\triangle^{c-mm}$. The formulation of the example is the following:

> At a meeting of a block of flat co-owners, the chairman proposes for the coming year the construction of a swimming pool, of a tennis court and a private car park. But if two of these three items are built, the rent will increase significantly ([12], p.787).

In the following, the letters p, q, r stand respectively for the construction of the swimming pool, the tennis court and the private car park. The rent increase will be denoted by s, which is implied by each conjunction of two out of the three items: $\mu = ((p \wedge q) \vee (p \wedge r) \vee (q \wedge r)) \rightarrow s$. The set of chioces of the co-owners is represented by $\Gamma = \{T_1 \sqcup T_2 \sqcup T_3 \sqcup T_4\}$. The first two of the co-owners want to build the three items and do not care about the rent (i.e. $(s \vee \neg s)$ holds in T_1 and T_2); the third does not want the rent increase nor anything built; the fourth wants the last two items (i.e. $(p \vee \neg p)$ holds in T_4), though he does not want the rent to increase:

$$\begin{aligned}
&T_1 = \{p \wedge q \wedge r\}\\
&T_2 = \{p \wedge q \wedge r\}\\
&T_3 = \{\neg p \wedge \neg q \wedge \neg r \wedge \neg s\}\\
&T_4 = \{q \wedge r \wedge \neg s\}.
\end{aligned}$$

Our premise set in $\mathbf{DM}$ is of the form $\Gamma = \{b_1(p \wedge q \wedge r), b_2(p \wedge q \wedge r), b_3(\neg p \wedge \neg q \wedge \neg r \wedge \neg s), b_4(q \wedge r \wedge \neg s)\}$. The adaptive procedure requires in the first instance the formulation of the disagreements in terms of Dab-consequences of Γ:

$$\begin{aligned}
Dab(\Delta_1) &= (b_1 p \wedge \neg b_0 p) \vee (b_3 \neg p \wedge \neg b_0 \neg p)\\
Dab(\Delta_2) &= (b_1 q \wedge \neg b_0 q) \vee (b_3 \neg q \wedge \neg b_0 \neg q)\\
Dab(\Delta_3) &= (b_1 r \wedge \neg b_0 r) \vee (b_3 \neg r \wedge \neg b_0 \neg r)\\
Dab(\Delta_4) &= (b_2 p \wedge \neg b_0 p) \vee (b_3 \neg p \wedge \neg b_0 \neg p)\\
Dab(\Delta_5) &= (b_2 q \wedge \neg b_0 q) \vee (b_3 \neg q \wedge \neg b_0 \neg q)\\
Dab(\Delta_6) &= (b_2 r \wedge \neg b_0 r) \vee (b_3 \neg r \wedge \neg b_0 \neg r)\\
Dab(\Delta_7) &= (b_4 q \wedge \neg b_0 q) \vee (b_3 \neg q \wedge \neg b_0 \neg q)\\
Dab(\Delta_8) &= (b_4 r \wedge \neg b_0 r) \vee (b_3 \neg r \wedge \neg b_0 \neg r)
\end{aligned}$$

These provide the following Ω^i sets of indexed abnormalities (where $!b_i\varphi$ will abbreviate $b_i\varphi \wedge \neg b_0\varphi$ provided $\varphi \in \mathcal{P}^\pm$ and $!b_i(\varphi_1 \vee \ldots \vee \varphi_n)$ will abbreviate $b_i(\varphi_1 \vee \ldots \vee \varphi_n) \wedge (\neg b_i\varphi_1 \wedge \ldots \wedge \neg b_i\varphi_n) \wedge \neg b_0(\varphi_1 \vee \ldots \vee \varphi_n)$ provided each $\varphi_i \in \mathcal{P}^\pm$ and $n > 1$):

$$\begin{aligned}
&\Omega^1 = \{!b_1 p, !b_1 q, !b_1 r\}\\
&\Omega^2 = \{!b_2 p, !b_2 q, !b_2 r\}\\
&\Omega^3 = \{!b_3 \neg p, !b_3 \neg q, !b_3 \neg r\}\\
&\Omega^4 = \{!b_4 q, !b_4 r\}.
\end{aligned}$$

Let us now consider our models, with respect to which abnormal cardinalities shall be calculated:

$$
\begin{aligned}
M_1 &= b_0p, b_0q, b_0r, b_0s\\
M_2 &= b_0p, b_0q, b_0r, b_0\neg s\\
M_3 &= b_0p, b_0q, b_0\neg r, b_0s\\
M_4 &= b_0p, b_0q, b_0\neg r, b_0\neg s\\
M_5 &= b_0p, b_0\neg q, b_0r, b_0s\\
M_6 &= b_0p, b_0\neg q, b_0r, b_0\neg s\\
M_7 &= b_0p, b_0\neg q, b_0\neg r, b_0s\\
M_8 &= b_0p, b_0\neg q, b_0\neg r, b_0\neg s\\
M_9 &= b_0\neg p, b_0q, b_0r, b_0s\\
M_{10} &= b_0\neg p, b_0q, b_0r, b_0\neg s\\
M_{11} &= b_0\neg p, b_0q, b_0\neg r, b_0s\\
M_{12} &= b_0\neg p, b_0q, b_0\neg r, b_0\neg s\\
M_{13} &= b_0\neg p, b_0\neg q, b_0r, b_0s\\
M_{14} &= b_0\neg p, b_0\neg q, b_0r, b_0\neg s\\
M_{15} &= b_0\neg p, b_0\neg q, b_0\neg r, b_0s\\
M_{16} &= b_0\neg p, b_0\neg q, b_0\neg r, b_0\neg s
\end{aligned}
$$

In view of μ the models M_2, M_4, M_6, M_{10} are rejected, i.e. any model satisfying $((p \wedge q) \vee (p \wedge r) \vee (q \wedge r)) \wedge \neg s$ is ignored. The initial assumption that $s \vee \neg s$ holds for T_1 and T_2, i.e. that though these agents express preference for the construction of all the three items, they still would approve if the three items might be built without increasing the rent ($\neg s$), means that with respect to $\neg s$ there is no disagreement, and none can be explicitely formulated within Γ. This in turn means that $\Gamma \models_{\mathbf{DM}} b_0\neg s$ holds, and the result of the selection shall be consistent with it. Hence, from the previous list all the models that still verify b_0s shall be removed as well. This leaves the following list:

$$
\begin{aligned}
M_8 &= b_0p, b_0\neg q, b_0\neg r, b_0\neg s\\
M_{12} &= b_0\neg p, b_0q, b_0\neg r, b_0\neg s\\
M_{14} &= b_0\neg p, b_0\neg q, b_0r, b_0\neg s\\
M_{16} &= b_0\neg p, b_0\neg q, b_0\neg r, b_0\neg s.
\end{aligned}
$$

6.1. Arbitration

For each of the remaining models one calculates the abnormal cardinality with respect to the indexed sets of abnormalities. For each model M_j and any indexed set of abnormalities Ω^i there will be a value to $\mathcal{C}^i_{M_j}$. These values are listed in the following table, where at the intersection of each M_j and Ω^i one has the value of $\mathcal{C}^i_{M_j}$, and in the last column each list $L^{M_j}_{\Gamma}$ obtained by sorting the elements of $(\mathcal{C}^1_{M_j}, \ldots, \mathcal{C}^n_{M_j})$ in descending oder:

	Ω^1	Ω^2	Ω^3	Ω^4	$L^{M_j}_{\Gamma}$
M_8	2	2	1	2	$(2, 2, 2, 1)$
M_{12}	2	2	1	1	$(2, 2, 1, 1)$
M_{14}	2	2	1	1	$(2, 2, 1, 1)$
M_{16}	3	3	0	2	$(3, 3, 2, 0)$

The lexicographic order $\leq^{\mathcal{C}}_{\Gamma}$ among the sequences of each $L^{M_j}_{\Gamma}$ gives the total preorder among the various models:

$$M_{12,14} \leq^{\mathcal{C}}_{\Gamma} M_8 \leq^{\mathcal{C}}_{\Gamma} M_{16}. \tag{15}$$

The result of merging according to $Min(\mathsf{M}_{\Gamma}, \leq^{\mathcal{C}}_{\Gamma})$ is:

$$
\begin{aligned}
\triangle^{c+}_{\mu}(\Gamma) = b_0((\neg p \wedge q \wedge \neg r \wedge \neg s) \vee \\
(\neg p \wedge \neg q \wedge r \wedge \neg s)).
\end{aligned}
\tag{16}
$$

The preferred choice by the group of co-owners is therefore to build either the tennis court or the private car park without increasing the rent. This is also the result of the pure arbitration $\triangle^{GMax}$ operator from [12].

6.2. Quasi-merging

By the same example it will be shown now how the $\triangle^{c-mm}$ operator for Minimax Adaptive Counting works. From the very same premise set $\Gamma = \{b_1(p \wedge q \wedge r), b_2(p \wedge q \wedge r), b_3(\neg p \wedge \neg q \wedge \neg r \wedge \neg s), b_4(q \wedge r \wedge \neg s)\}$, the same derivable disjunctions of abnormalities and list of Ω^i sets, one derives the same list of values for abnormal cardinalities in each of the possible models, and the same lexicographic order of these values.

The models that allow the combination $b_0((p \wedge q) \vee (p \wedge r) \vee (q \wedge r) \wedge s)$ are obviously still rejected in view of the constraint μ; and it still holds in the merging state $b_0\neg s$ in view of the absence of disagreements with respect to this literal. The rejection of any other model in which s holds – which leaves only models $M_8, M_{12}, M_{14}, M_{16}$ – is of the greatest importance in order to show the result of our minimax selection.

By the original $\triangle^{Max}$ operator from [12], one cannot avoid that some of the models are selected in which at least two between p, q, r are negated (i.e. only one of the item is allowed to be built by the group of co-owners), and nonetheless s is satisfied (i.e. the rent is increased). This result is counterintuitive in view of the required constraint, but it is also undesirable in view of intelligent interaction by our agents. Our $\triangle^{c-mm}$ operator avoids this undesirable result.

According to Definition 5.1, one selects the Maximal Abnormal Distance for each of the allowed models out of the lexicographic order of abnormal cardinalities:

	Ω^1	Ω^2	Ω^3	Ω^4	$Max(M_j, \Gamma)$
M_8	2	2	1	2	2
M_{12}	2	2	1	1	2
M_{14}	2	2	1	1	2
M_{16}	3	3	0	2	3

from which the following preorder based on $\leq^{Max}_{\Gamma}$ is obtained:

$$M_{8,12,14} \leq^{Max}_{\Gamma} M_{16}. \tag{17}$$

The selection of models M_8, M_{12}, M_{14} with the minimal values provides the following alternatives:

$$\begin{aligned}\triangle_{\mu}^{c-mm}(\Gamma) = b_0((p \wedge \neg q \wedge \neg r \wedge \neg s) \vee \\ (\neg p \wedge q \wedge \neg r \wedge s) \vee \\ (\neg p \wedge \neg q \wedge r \wedge \neg s)).\end{aligned} \tag{18}$$

The preferred choice by the group of co-owners is therefore to build one of the three items without increasing the rent. This results avoids the other alternatives allowed by the $\triangle^{Max}$ operator according to which one among the tennis court or the private park is built and the rent is increased (the latter condition being not necessary in view of the constraints).

7. Conclusion

The formulation of the family of adaptive logics $\mathbf{ADM}$, started with the definition of a Majority merging selection in [20], has been in this paper further developed by the definition of selection procedures corresponding to pure-arbitration and quasi-merging protocols. A next obvious step of this research is represented by the formulation of a selection procedure for $\mathbf{ADM}$ that reflects the $\triangle^n$ operators from [13], a set of operators that belong simultaneously to the two main sub-families, majority and arbitration.

A number of application contexts, such as those presented in [19] for heterogenous databases and in [18] for judgment aggregation procedures, provide the settings for testing the computational limits and effectiveness of the procedures. With respect to these open questions, a number of positive and negative results can be formulated, in line with those valid for other general merging protocols.

References

[1] C. Alchourrón, P. Gärdernfors, and D. Makinson. On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic,*, 50:510–530, 1985.

[2] D. Batens. A general characterization of adaptive logics. *Logique & Analyse*, 173–175:45–68, 2001.

[3] D. Batens. A universal logic approach to adaptive logics. *Logica universalis*, 1:221–242, 2007.

[4] S. Benferhat, D. Dubois, and H. Prade. Some syntactic approaches to the handling of inconsistent knowledge bases: A comparative study. Part 1: The flat case. *Studia Logica*, 58:17–45, 1997.

[5] A. Borgida and T. Imielinski. Decision making in committees: a framework for dealing with inconsistency and nonmonotonicity. In *Proceedings Workshop on Nonmonotonic Reasoning*, pages 21–32, 1984.

[6] L. Cholvy. *Reasoning about merged information*, volume 3, pages 233–263. Kluwer Academic Publisher, 1998.

[7] L. Cholvy and T. Hunter. Fusion in logic: a brief overview. In *4th European Conference on Symbolic and Quantitative Approaches to reasoning with Uncertainty*, volume 1244 of *Lectures Notes in Computer Science*, pages 86–95, 1997.

[8] M. Dalal. Investigations into theory of knowledge base revision. In *Proc. AAAI-88*, pages 449–479. St. Paul, MN, 1988.

[9] R. Fagin, J. Ullman, and M. Vardi. On the semantics of updates in databases. In *Second ACM SIGACT-SIGMOD*, pages 352–365, 1983.

[10] E. Grégoire and S. Konieczny. Logic-based approaches to information fusion. *Information Fusion*, 7:4–18, 2006.

[11] S. Konieczny and R. Pino-Pérez. On the logic of merging. In *Proceedings of the Sixth International Conference on Principles of Knowledge Representation and Reasoning (KR'98)*, pages 488–498, 1998.

[12] S. Konieczny and R. Pino-Pérez. Merging information under constraints: A logical framework. *Journal of Logic and Computation*, 12(5):773–808, 2002.

[13] S. Konieczny and R. Pino-Pérez. On the frontier between arbitration and majority. In *Proceedings of the 8^{th} International Conference on Principles of Knowledge Representation and Reasoning*, pages 109–120, 2002.

[14] P. Liberatore and M. Schaerf. Arbitration: a commutative operator for belief revision. In *Proceedings of the Second World Conference on the Fundamentals of Articial Intelligence*, pages 217–228, 1995.

[15] P. Liberatore and M. Schaerf. Arbitration (or how to merge knowledge bases). *IEEE Transactions on Knowledge and Data Engineering*, pages 76–90, 1998.

[16] J. Lin and A. Mendelzon. Knowledge base merging by majority. In *Dynamic Worlds: from the Frame Problem to Knowledge Management*. Kluwer, 1999.

[17] G. Primiero. Belief merging based on adaptive interaction. In J. van Benthem, S. Ju, and F. Veltman, editors, *A meeting of the minds - Proceedings of the Workshop on Logic, rationality and Interaction*, volume 8 of *Texts in Computer Science*, pages 313–320. College Publications, 2007.

[18] G. Primiero. Aggregating collective judgements by selecting disagreements. Accepted paper at LOFT08, 8^{th} Conference on Logic and the Foundations of Game and Decision Theory, Amsterdam 3-5 July 2008, 2008.

[19] G. Primiero and J. Meheus. Adaptive arbitration by variant counting on commutative bases with weights. In *Proceedings of the $11th$ International Conference on Information Fusion*, 2008. forthcoming.

[20] G. Primiero and J. Meheus. Majority merging by adaptive counting. *Knowledge, Rationality and Action (Synthese)*, forthcoming.

[21] P. Revesz. On the semantics of arbitration. *Journal of Algebra and Computation*, 7(2):133–160, 1997.

[22] K. Satoh. Nonmonotonic reasoning by minimal belief revision. In *Proceedings of the International Conference on Fifth Generation Computer Systems*, pages 455–462, 1988.

[23] L. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1(1):3–28, 1978.

Extending Probabilistic Dynamic Epistemic Logic

Joshua Sack
joshua.sack@gmail.com

Abstract

This paper aims to extend in two directions the probabilistic dynamic epistemic logic provided in Kooi's paper [8]. Kooi's probabilistic dynamic epistemic logic adds to probabilistic epistemic logic sentences that express consequences of public announcements. The first extension offered in this paper is to add a previous time operator to a probabilistic dynamic epistemic logic similar to Kooi's. The other is to involve action models and update products in a probabilistic dynamic epistemic logic setting. This would allow for more kinds of actions, such as private announcements.

1 Introduction

Probabilistic epistemic logic has been developed to express interaction between both qualitative and quantitative beliefs. This logic lets us formally express statements such as "Bob believes the probability of φ to be at least 1/2" or "Ann considers the probability of ψ to be 1/4". As we are often concerned about how beliefs and probabilities change over time, there have been papers written that mix probability, belief, and time. Examples include, [7] and [3], which use probabilistic systems of runs, and [8], which combines probability with public announcement logic. The probabilistic systems of runs provides a natural way to view time, both past and future, but conditions need to be imposed in order to ensure that agents' probability measures change in a realistic way. Public announcement logic, and more generally dynamic epistemic logic (DEL), provides a mechanical procedure for changes in belief upon receipt of public information, and [8] extends this mechanical procedure to show how a probability measure may change given public information. But DEL has limitations in its ability to express features of the past and future. By adding temporal logic to DEL in a non-probabilistic setting, the paper [13] captures both some of the temporal flexibility of the system of runs as well as the mechanical method offered by DEL of going from one stage in time to the next. One goal of this paper is to involve probability in the combination of temporal logic and DEL, focusing on the inclusion of a previous-time operator and exploring the possibility of completeness.

Another goal is to go beyond public announcements. There are other forms of information exchange that are of interest, such as semi-private announcements, where the fact that a message was sent to someone is not a secret, and completely-private announcements, where non-recipients of the message are completely unaware of the fact that there is a message at all. An illuminating example involving semi-private announcements is given in [3]. The sequence of events provides a context motivating why there are stages in time in which an agent's sample space should differ from the set of states the agent considers possible. A non-probabilistic mechanical method for changing beliefs according to semi-private announcements was given in [2] and [1]. There, semi-private announcements are encoded in action models, and a product is defined between a model and an action model to produce an updated model that encodes the updated beliefs. This goal then is to involve action models in probabilistic dynamic epistemic logic.

2 Probabilistic Public Announcement Logic with a Previous Time Operator

As the underlying structure for public announcement logic is the epistemic model, the underlying structure for probabilistic public announcement logic is the probabilistic epistemic model, which adds probability spaces to an epistemic model.

Definition 2.1 [Probabilistic Epistemic Model] Let Φ be a set of proposition letters, and $\mathbf{I}$ be a set of agents. A probabilistic epistemic model is a tuple $\mathbf{M} = (X, \{\xrightarrow{i}\}_{i\in\mathbf{I}}, \|\cdot\|, \{\mathbf{P}_{i,x}\})$, where

- X is a set of "states" or "possible worlds"
- $\xrightarrow{i} \subseteq X^2$ is an epistemic relation for each agent $i \in \mathbf{I}$, that is $x \xrightarrow{i} y$ if i considers y possible from x
- $\|\cdot\|$ is a function assigning to each proposition letter p the set of states where it is true.

- $\mathbf{P}_{i,x}$ is a probability space for each agent i and state x, that is $P_{i,x} = (S_{i,x}, \mathcal{A}_{i,x}, \mu_{i,x})$, where
  - $S_{i,x} \subseteq X$ is set called the sample space
  - $\mathcal{A}_{i,x}$ is a σ-algebra over $S_{i,x}$ (that is, a collection of subsets of $S_{i,x}$ that is closed under complements and countable unions). We the sets in the σ-algebra "measurable sets".
  - $\mu_{i,x} : \mathcal{A}_{i,x} \to [0, 1]$ is a probability measure over $S_{i,x}$ (that is, $\mu_{i,x}(S_{i,x}) = 1$ and for each countable collection $A_1, A_2, \ldots$ of pairwise disjoint sets in $\mathcal{A}$, $\mu(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} \mu(A_k)$).

$\triangleleft$

For the rest of this section, we restrict the set X to be finite and the set $\mathcal{A}_{i,x}$ to be the power set $\mathcal{P}(S_{i,x})$. Thus we need not specify the σ-algebra $\mathcal{A}_{i,x}$ until the next section. It is recommended that $S_{i,x} \subseteq \{z : x \xrightarrow{i} z\}$, as every outcome in the sample space is a state the agent considers possible. For technical convenience in definition 2.2, we will not impose such a restriction. One might assume that the converse of the recommendation should hold too, thus making i's sample space $S_{i,x}$ equal to the set $\{z : x \xrightarrow{i} z\}$ of states i considers possible, but the example in the beginning of the next section motivates why we prefer not to make this restriction either. The example presents a situation in which an agent does not know enough to assign a probability to everything she considers possible, and although there are different ways of handling the uncertainty about the probability, omitting some states from the sample space is an attractive solution. As sample spaces are defined for each state, the agent may still be uncertain about which sample space is correct.

Public announcement logic is concerned with how an agent revises his/her beliefs given new information, knowing that this information is received by all other agents. Of greatest interest is new information that is consistent with the agent's beliefs, and PAL provides an enlightening mechanical procedure called an *update* for producing a new epistemic model from an old one given the new information. But although the updates provide a reasonable method of revising beliefs upon consistent information, they do not upon inconsistent information. The goal of probabilistic public announcement logic is to provide an update procedure that shows how to produce a new probabilistic epistemic model from an old one given information that is not only consistent with the agents' beliefs, but is also given positive probability. The case where the probability of the new information is 0 poses difficulties, and the goal of the definition of such a case is more to provide technical convenience than to offer a realistic result.

Definition 2.2 [updates] Given a probabilistic epistemic model $\mathbf{M} = (X, \{\xrightarrow{i}\}, \| \cdot \|, \{\mathbf{P}_{i,x}\})$ and a subset Y of X, the update of $\mathbf{M}$ given Y is written $\mathbf{M} \otimes Y$ and is the model $(X', \{\xrightarrow{i}\}', \| \cdot \| ', \{\mathbf{P}'_{i,x}\})$, where

- $X' = X \cap Y$
- $x \xrightarrow{i}{}' y$ iff $x, y \in Y$ and $x \xrightarrow{i} y$
- $\|p\|' = \|p\| \cap Y$
- If $\mu_{i,x}(Y) = 0$, then let $\mathbf{P}'_{i,x}$ be the only probability space definable on the singleton x. Otherwise, let $\mathbf{P}'_{i,x}$ be defined by
  - $S'_{i,x} = S_{i,x} \cap Y$
  - For each subset $Z \subseteq Y$, $\mu'_{i,x}(Z) = \mu_{i,x}(Z)/\mu_{i,x}(Y)$

$\triangleleft$

This definition differs from the one in [8] in that here updating removes states while in [8] it removes relational connections but not states. Probability is updated in the same way as long as the set Y has positive probability. The reason for these differences is to aid in proving completeness for the language that adds a previous time operator to probabilistic public announcement logic. Although completeness is still under construction (and the semantics in [8] may later demonstrate itself to be easier), the following discussions about the previous time operator may provide some intuition for why this new semantics may help, particularly for the conditions in definition 2.4.[1]

A natural choice for semantics that includes a previous time operator for this language is to involve structures that consist of a list of all past and present models. This is what was done in [13]

Definition 2.3 [History] A history H is a list of models $(\mathbf{M}_0, \mathbf{M}_1, \ldots, \mathbf{M}_n)$, where for each k, $\mathbf{M}_k = (X_k, \{\xrightarrow{i}_k \}, \| \cdot \|_k, \{\mathbf{P}_{k i,x}\})$, and $M_{k+1} = M_k \otimes X_{k+1}$. $\triangleleft$

Given a history $H = (\mathbf{M}_0, \mathbf{M}_1, \ldots, \mathbf{M}_n)$, let $\widehat{P}(H) = (\mathbf{M}_0, \mathbf{M}_1, \ldots \mathbf{M}_{n-1})$ be the previous history, $\widehat{\mathbf{M}}(H) = \mathbf{M}_n$ be the last (most recent) model in the list, and let $\widehat{X}(H) = X_n$. We may write $x \in H$ for $x \in \widehat{X}(H)$.

[1] Definition 2.4 makes use of sets A_n consisting of all states corresponding to time n. It is helpful that this set is equal to the set B_n consisting of all of the more recent versions of states that satisfied the formula that induced the update. Finding an appropriate characterization of B_n so far appears more difficult using the semantics of [8].

Language

Let Φ be a set of proposition letters and $\mathbf{I}$ be a set of agents. We define by mutual recursion a multi-sorted language $\mathcal{L}$ with sentences and terms for each agent. The sentences (also called formulas) are given by

$$\varphi ::= \mathsf{true} \,|\, p \,|\, \neg\varphi \,|\, \varphi_1 \wedge \varphi_2 \,|\, \Box_i\varphi \,|\, [\varphi_1]\varphi_2 \,|\, t_i \geq q \,|\, \overline{Y}\varphi$$

where t is a term, $p \in \Phi$, and $i \in \mathbf{I}$.

The terms for agent i are given by

$$t_i ::= qP_i(\varphi) \,|\, t_i + u_i$$

where $q \in \mathbb{Q}$ is a rational number, t_i and u_i are terms for agent i, and φ is a sentence.

The semantics is defined by a function $[\![\cdot]\!]$ from formulas to functions f that map each history H to a subset of $\widehat{X}(H)$, the carrier set of the most recent model in H. Then

- $[\![\mathsf{true}]\!]$ is the function that maps each H to the whole set $\widehat{X}(H)$,
- $[\![\neg\varphi]\!](H) = \widehat{X}(H) - [\![\varphi]\!](H)$
- $[\![\varphi \wedge \psi]\!](H) = [\![\varphi]\!](H) \cap [\![\psi]\!](H)$.
- $x \in [\![\Box_i\varphi]\!](H)$ if and only $y \in [\![\Box_i\varphi]\!](H)$ for every y in which $x \xrightarrow{i} y$, where $\xrightarrow{i}$ is i's epistemic relation in $\widehat{\mathbf{M}}(H)$ (the most recent model in H).
- $x \in [\![[\varphi_1]\varphi_2]\!](H)$ if and only if either $x \notin [\![\varphi_1]\!](H)$ or $x \in [\![\varphi_2]\!](H \otimes [\![\varphi_1]\!](H))$.
- $x \in [\![q_1P_i(\varphi_1) + \cdots + q_nP_i(\varphi_n) \geq q]\!](H)$ if and only if $q_1\mu_{i,x}([\![\varphi_1]\!](H)) + \cdots + q_n\mu_{i,x}[\![\varphi_n]\!](H)) \geq q)$.
- $x \in [\![\overline{Y}\varphi]\!](H)$ if and only if $H = (\mathbf{M})$ has just one model or $x \in [\![\varphi]\!](\widehat{P}(H))$.

We have the usual modal abbreviations, such as $\Diamond_A\varphi \equiv \neg\Box_A\neg\varphi$ and $\langle\psi\rangle\varphi \equiv \neg[\psi]\neg\varphi$, and we let $\widehat{Y}\varphi \equiv \neg\overline{Y}\neg\varphi$, which asserts that there is a previous time and φ is true then. Here are some abbreviations in the language that express a variety of inequalities and equality.

- $t \leq q \equiv -t \geq -q$
- $t < q \equiv \neg(t \geq q)$
- $t > q \equiv \neg(t \leq q)$
- $t = q \equiv t \leq q \wedge t \geq q.$
- $t \geq s \equiv t - s \geq 0$
- $t = s \equiv t - s \geq 0 \wedge s - t \geq 0$

Proof system

Include axioms of proposition logic together with the following:

$\Box_i$-normality	$\Box_i(\varphi \to \psi) \to (\Box_i\varphi \to \Box_i\psi)$
$[\varphi]$ -normality	$[\varphi](\psi_1 \to \psi_2) \to ([\varphi]\psi_1 \to [\varphi]\psi_2)$
$\overline{Y}$-normality	$\overline{Y}(\varphi \to \psi) \to (\overline{Y}\varphi \to \overline{Y}\psi)$
Update partial functionality	$[\varphi]\neg\psi \leftrightarrow (\varphi \to \neg[\varphi]\psi)$
$\overline{Y}$-partial functionality	$\widehat{Y}\psi \leftrightarrow (\widehat{Y}\,\mathsf{true} \to \overline{Y}\psi)$
Future atomic permanence	$(\varphi \to p) \leftrightarrow [\varphi]p$
Past atomic permanence	$\overline{Y}p \leftrightarrow (\widehat{Y}\,\mathsf{true} \to p)$
Update yesterday	$[\varphi]\overline{Y}\psi \leftrightarrow (\varphi \to \Box_i[\varphi]\psi)$
Probability yesterday 0	$\widehat{Y}(\sum_{k=1}^n q_kP_i(\varphi_k) = 0) \to (\sum_{k=1}^n q_kP_i(\widehat{Y}\varphi_k) = 0)$
Probability yesterday 1	$\widehat{Y}(\sum_{k=1}^n q_kP_i(\varphi_k) = \sum_{k=1}^n q_kP_i(\,\mathsf{true}))$ $\to (\sum_{k=1}^n q_kP_i(\widehat{Y}\varphi_k) = \sum_{k=1}^n q_kP_i(\,\mathsf{true}))$
Epistemic-yesterday mix	$\overline{Y}\Box_i\varphi \to \Box_i\overline{Y}\varphi$
Epistemic update	$[\varphi]\Box_i\psi \leftrightarrow (\varphi \to \Box_i\psi)$
Probability update	$P_i(\varphi) > 0 \to ([\varphi]\sum_{k=1}^n q_kP_i(\varphi_k) \geq q$ $\leftrightarrow (\varphi \to \sum_{k=1}^n q_kP_i(\varphi \wedge [\varphi]\varphi_k) \geq qP_i(\varphi)))$
Probability 0 update	$P_i(\varphi) = 0 \to ([\varphi](\sum_{k=1}^n q_kP_i(\varphi_k)) \geq q$ $\leftrightarrow (\varphi \to \sum_{k=1}^n q_kP_i(\,\mathsf{true}) \geq q)$
Non-initial time	$\widehat{Y}\,\mathsf{true} \to \Box_i\widehat{Y}\,\mathsf{true} \wedge P_i(\widehat{Y}\,\mathsf{true}) = 1$
Initial time	$\overline{Y}\,\mathsf{false} \to \Box_i\overline{Y}\,\mathsf{false} \wedge P_i(\overline{Y}\,\mathsf{false}) = 1$
0 terms	$\sum_{k=1}^n q_kP_i(\varphi_k) \geq q$ $\leftrightarrow (\sum_{k=1}^n q_kP_i(\varphi_k)) + 0P_i(\varphi_{k+1}) \geq q$
Permutation	$\sum_{k=1}^n q_kP_i(\varphi_k) \geq q \to \sum_{k=1}^n q_{j_k}P_i(\varphi_{j_k}) \geq q$ where$j_1, \ldots, j_n$ is a permutation of $1, \ldots, n$
Addition	$\sum_{k=1}^n q_kP_i(\varphi_k) \geq q \wedge \sum_{k=1}^n q'_kP_i(\varphi_k) \geq q'$ $\to \sum_{k=1}^n (q_k + q'_k)P_i(\varphi_k) \geq (q + q')$
Multiplication	$(\sum_{k=1}^n q_kP_i(\varphi_k) \geq q)$ $\leftrightarrow (\sum_{k=1}^n dq_kP_i(\varphi_k) \geq dq)$ where $d > 0$
Dichotomy	$(t \geq q) \vee (t \leq q)$
Monotonicity	$(t \geq q) \to (t > q')$ where $q > q'$
Nonnegativity	$P_i(\varphi) \geq 0$
Probability of truth	$P_i(\,\mathsf{true}) = 1$
Additivity	$P_i(\varphi \wedge \psi) + P_i(\varphi \wedge \neg\psi) = P_i(\varphi)$

Include axioms of proposition logic together with the following:

$\Box_i$-necessitation	From $\vdash \varphi$ infer $\vdash \Box_i\varphi$
$[\varphi]$ -necessitation	From $\vdash \varphi$ infer $\vdash [\varphi]\varphi$
$\overline{Y}$-necessitation	From $\vdash \varphi$ infer $\vdash \overline{Y}\varphi$
Equivalence	
From $\vdash \varphi \leftrightarrow \psi$, infer $\vdash P_i(\varphi) = P_i(\psi)$	

Approach to completeness

A general strategy for proving weak completeness is to start with a consistent formula and then prove that it is satisfiable. Modal logic and probabilistic epistemic logic provide techniques for finding filtrations that may satisfy either the formula or a formula provably equivalent to the first. But such filtrations are single models, not lists of models. There is a similar difficulty for DEL to use filtrations, because the semantics of DEL involves the construction of a new model. But it turns out that in DEL, a semantics for a subset of formulas, each called normal form formulas, can be defined in which no new model needs to be constructed in order to determine whether a formula is true. In addition, each formula is provably equivalent to a normal form formula, and the two semantics relate to each other in a natural and convenient way: a formula is true given one semantics if and only if it is true given the other.

To employ a strategy similar to this, we need an alternative set of models and an alternative semantics. Let us define a non-standard model as a probabilistic epistemic model together with a binary relation Y. Ideally xYz can be read as "x is one stage later than z", but this interpretation may be difficult to achieve unless there are some restrictions placed on this new non-standard model. We thus provide the following definition:

Definition 2.4 [non-standard history] Let

$$\mathcal{M} = (X, \{\xrightarrow{i}\}_{i\in\mathbf{I}}, \|\cdot\|, \{\mathbf{P}_{i,x}\}, Y)$$

be a non-standard model. Define

$$A_0 = \{x : \text{there is no } z \text{ such that } xYz\},$$

and for each $n > 0$,

$$A_n = \{x : \text{there is a } z \text{ such that } xY^nz \\ \text{and there is no } z \text{ such that } xY^{n+1}z\}$$

Define for each set A and binary relation R,

$$R(A) = \{z : \text{there is a } x \in A \text{ such that } xRz\}$$

Then $\mathcal{M}$ is a non-standard history if the following conditions hold:

1. Partial functionality of Y: if xYz and xYz', then $z = z'$.
2. Bounded age: There exists N such that for all x there is no z for which xY^Nz.
3. Epistemic synchronicity: if $x \xrightarrow{i} z$, then for each n, xY^nx' for some x' iff zY^nz' for some z'.
4. Probabilistic synchronicity: if $x, z, w \in X$ and $x, z \in S_{i,w}$, then for each n, xY^nx' for some x' iff zY^nz' for some z'.
5. Update product relation condition a: if $x \xrightarrow{i} z$ and zYz' then there exists x' such that xYx' and $x' \xrightarrow{i} z'$.
6. Update product relation condition b: if xYx', $x' \xrightarrow{i} z'$, and zYz', then $x \xrightarrow{i} z$.
7. Update product sample space condition a: for each $n \geq 1$, $i \in \mathbf{I}$, $x \in A_n$, and z such that xYz, if $\mu_{i,z}(Y(A_n)) > 0$, then $Y(S_{i,x}) = Y(A_n) \cap S_{i,z}$.
8. Update product sample space condition b: for each $n \geq 1$, $i \in \mathbf{I}$, $x \in A_n$, and z such that xYz, if $\mu_{i,z}(Y(A_n)) = 0$, then $S_{i,x} = \{x\}$.
9. Update product probability condition a: for each $n \geq 1$, $i \in \mathbf{I}$, $x \in A_n$, and z such that xYz, if $\mu_{i,z}(Y(A_n)) > 0$, then for each $A \subseteq S_{i,x}$,
$$\mu_{i,x}(A) = \frac{\mu_{i,z}(Y(A))}{\mu_{i,z}(Y(A_n))}$$
10. Update product probability condition b: for each $n \geq 1$, $i \in \mathbf{I}$, $x \in A_n$, and z such that xYz, if $\mu_{i,z}(Y(A_n)) = 0$, then $\mu_{i,x}(\{x\}) = 1$.
11. Update product valuation condition: if xYz then $x \in \|p\|$ iff $z \in \|p\|$

$\triangleleft$

Semantics can be defined for formulas that do not include public announcement operators $[\varphi]$. These formulas constitute a language which we call *normal form*. The operator $\overline{Y}$ is treated as the box modality for the relation Y. The semantics for the other operators remain the same.

To show that every formula is provably equivalent to one in normal form, we employ a term rewriting system similar to one used in [1]. Our term rewriting system will make use of the following algebraic semantics.

Definition 2.5 [Signature] We define Δ to be the following signature. It is multi-sorted, with one sort for sentence terms s and another for weight terms t_i for each agent $i \in \mathbf{I}$. Here are the symbols in the signature:

1. Each $p \in \Phi$ and true is a constant symbol of sort s.

2. $\neg, \Box_A, \Box^*_{\mathcal{B}}, \overline{Y}$ are function symbols of type $s \to s$

3. $\wedge$, $\to$, and [ ] are binary function symbols of type $s \times s \to s$

4. P_i is a function of type $\mathbb{Q} \times s \to t_i$

5. $+_i$ is a function of type $t_i \times t_i \to t_i$

6. $\geq_i$ is a function of type $t_i \times \mathbb{Q} \to s$

7. triv_i is a function of type $t_i \to t_i$

8. bay_i is a function of type $s \times t_i \to t_i$

◁

We will in general write $[s]s$ for $[\,](s, s)$, and we choose to write $+_i$ and $\geq_i$ in infix notation too. In addition, we often drop the subscripts when it is understood from context. The functions bay and triv are just tools for reducing formulas of the form $[x](t \geq_i q)$ in the next definition. The choice of symbol bay is supposed to indicate a relationship to Bayesian updating, and the choice of the symbol triv is to indicate that the probabilities are trivialized (that is, we will take the probability of true).

Let $\mathcal{L}^+$ be the algebraic language defined by this signature, and let $\mathcal{L}^+(X)$ be the language L augmented with a set of variables X. Syntactically, occurrences of a variable must agree on the sort, that is, $x +_i x$ implies that x is a weight term for i, and $x \geq_i x$ is not allowed, since the first occurrence of x would have to be a weight term and the second a sentence term.

A term rewriting system is a collection of rewrite rules, written $\varphi \rightsquigarrow \psi$, where $\varphi, \psi \in \mathcal{L}^+(X)$. Executing a rewrite rule on a formula χ would identify a substitution instance of φ in χ and replace it with a substitution instance of ψ (using the same assignment of variables to terms). For our purposes, we use the following rewrite system.

Definition 2.6 [Rewriting system $\mathcal{R}$] Here is a rewriting system of use to us

(r1)	$x \to y$	$\rightsquigarrow$	$\neg(x \wedge \neg y)$
(r2)	$[x]\,\mathsf{true}$	$\rightsquigarrow$	true
(r3)	$[x]p$	$\rightsquigarrow$	$x \to p$
(r4)	$[x]\neg y$	$\rightsquigarrow$	$x \to \neg[x]y$
(r5)	$[x](y \wedge z)$	$\rightsquigarrow$	$[x]y \wedge [x]z$
(r6)	$[x]\Box_A y$	$\rightsquigarrow$	$x \to \Box_A[x]y$
(r7)	$[x]\overline{Y}z$	$\rightsquigarrow$	$x \to z$
(r8)	$\mathsf{triv}(P_i(q, x))$	$\rightsquigarrow$	$P_i(q, \mathsf{true})$
(r9)	$\mathsf{triv}(t_1 + t_2)$	$\rightsquigarrow$	$\mathsf{triv}(t_1) + \mathsf{triv}(t_2)$
(r10)	$\mathsf{bay}(x, P_i(q, z))$	$\rightsquigarrow$	$P_i(q, x \wedge [x]z)$
(r11)	$\mathsf{bay}(x, t_1 + t_2)$	$\rightsquigarrow$	$\mathsf{bay}(x, t_1) + \mathsf{bay}(x, t_2)$
(r12)	$[x](t \geq_i q)$	$\rightsquigarrow$	$(P_i(-1, x) \geq 0 \wedge (x \to (\mathsf{triv}(t) \geq q)))\vee$ $(\neg(P_i(-1, x) \geq 0)\wedge$ $(x \to (\mathsf{bay}(x, t) +_i P_i(-q, x) \geq 0)))$

◁

These rules correspond to either biconditional axioms schema or provable biconditiionals, which is the core reason why a rewritten formula is provably equivalent to the first. Note that there is a natural translation between our original language $\mathcal{L}$ and the algebraic language $\mathcal{L}^+$. The rewriting is done in $\mathcal{L}^+$ and the provable equivalence is determined between corresponding formulas in $\mathcal{L}$.

But it is also important that we can apply rules finitely many times in order to obtain a term corresponding to a formula in normal form. We first observe that no rule can be applied to terms if and only if the terms correspond to formulas in in normal form. But we must also show that only finitely many applications of the rules can be applied, something that we may doubt, given that rule (r12) appears to produce a much more complicated term. But the following interpretation of symbols in the signature can help us show that the rewriting system terminates.

Definition 2.7 [Interpretation of Signature] Let us overload the symbol $[\![\cdot]\!]$ to indicate interpretation. Our signature has a carrier $\mathbb{N}_{\geq 3}$ for sentences, and a carrier $\mathbb{N}_{\geq 3}$ for actions. The function symbols are then interpreted as the following arithmetic functions on these numbers:

$[\![\,\mathsf{true}]\!]$	$= 3$	$[\![\mathsf{bay}]\!](a, b)$	$= a^b$
$[\![p]\!]$	$= 3$	$[\![\mathsf{triv}]\!](b)$	$= 3^b$
$[\![\neg]\!](a)$	$= a + 1$	$[\![Y_\Box]\!](a)$	$= a + 1$
$[\![\wedge]\!](a, b)$	$= a + b$	$[\![[\,]]\!](a, b)$	$= a^{b+4}$
$[\![\to]\!](a, b)$	$= a + b + 3$	$[\![P_i]\!](q, a)$	$= a + 5$
$[\![\Box_A]\!](a)$	$= a + 2$	$[\![\geq_i]\!](a, q)$	$= a$
$[\![+_i]\!](a, b)$	$= a + b$		

We recursively extend this interpretation to all terms and sentences. ◁

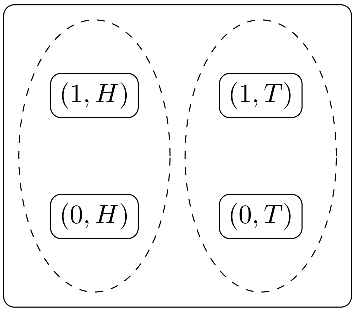
It turns out that every application of a term rewriting rule results in a term with a strictly smaller interpretation. Thus the interpretation of the term is an upper bound to the number of times rules can be applied before terminating.

To show completeness, we start with a consistent formula in $\mathcal{L}$, translate it into $\mathcal{L}^+$, and then apply rewrite rules until we obtain a term in which no more rules can be applied. The result of translating this term back into $\mathcal{L}$ is a normal form formula provably equivalent to the original formula. We then show for every non-standard history $\mathcal{H}$ and state x in the non-standard history, there is an actual history H and a state in that history that agrees with x on the truth of every normal form formula. We then form a filtration for the non-standard semantics. A number of model transformations will likely be needed to turn the filtration into a history. One transformation that might be useful is an unravelling (or partial unravelling) about a point in the canonical model that satisfies the consistent formula. This was done in [12]. But producing a discrete probability after unravelling is not always straightforward or possible. Thus it would be helpful to get a better grasp of issues regarding updating probabilistic epistemic models that have unmeasurable sets (sets not in the σ-algebra). This is the greatest challenge discussed in the next section concerning a dynamic probabilistic epistemic logic.

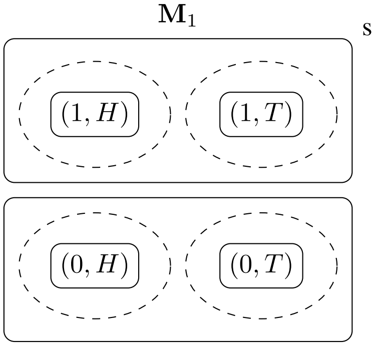
3 Involving Action Models

The example in [3] is as follows. There are two agents: i and k. Agent k receives a bit: 0 or 1. Agent i is aware that k learns what the bit is, but i does not know what the bit is. Then agent k flips a fair coin, and observes the result. Again i is aware that k learns the result of the flip, but does not learn the result. Viewing heads as 1 and tails as 0, agent k performs action s if the coin agrees with the bit, and d if it does not.

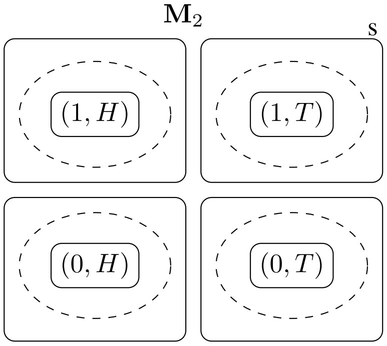
Fagin and Halpern viewed this experiment through a system of runs. There are 4 possible runs of this example based on the outcome of the bit together with the outcome of the coin. The action d or s is determined from the first two outcomes. Let us consider 4 states, one for each run: $(1,H),(1,T),(0,H),(0,T)$. Before agent k performs action s or d, agent i considers all four possible. But what should agent i's probability space be at each state? Three possibilities are discussed in [3] and are depicted below. The solid box depicts what i would consider to be the sample spaces from each state within its borders, and the dotted lines depict the smallest non-empty sets in the σ-algebra for i from each element in the sample space. Note that i's probability spaces from one state to another need not always differ.



$\mathbf{M}_1$



$\mathbf{M}_2$



$\mathbf{M}_3$

The diagram on the left depicts the situation where the sample space should consist of all 4 states, thus making the sample space the same as the set of states considered possible. Notice that the set $\{(1,H),(1,T)\}$ is not measurable (that is, is not in the σ-algebra). Unlike a fair coin, we do not know enough about how the bit is to be assigned to give it a probability. But even without knowing the probability of the bits 1 and 0, we can determine that s (represented by $\{(1,H),(0,T)\}$) has probability 1/2, and similarly for d. So the second diagram provides model where i can give a probability to s, but still does not give a probability to bits 1 or 0. The set of states i considers possible is still all four states, but now the sample spaces are different from that. Thus i can also be uncertain about whether the bit turned out to be 0 or 1, but can still determine that the probability of s is 1/2 in each case, thus concluding that the probability is 1/2. In the third diagram, agent i believes the probability of any outcome is either 1 or 0, but does not know which.

It is suggested that these three diagrams may be viewed as three different stages of the example. The first diagram would correspond to the time before the bit is given. The second diagram would correspond to the time after the bit is given but before the coin is flipped. The third diagram would correspond to the time after the coin is flipped. The transition from one stage to the next might make more sense if we view the probabilities as objective. Indeed, these diagrams would just as appropriately represent k's probabilities at the three stages. But one consequence of using these objective probabilities is that i's degree of certainty about the probabilities changes as a result of learning that k was informed of something. From the first diagram to the second, i becomes more certain about the probability of events s and d, but in both steps, i becomes less certain about the probability space. Being more sensitive to probabilities as subjective, we may prefer that there is more information revealed to i between stages, to help the change in the degree of certainty. Suppose that at first i is not aware of any plan for k to perform either action s or d, and hence does not wonder about the probabilities of these two events. The sequence of actions may be as follows:

1. k receives the bit.
2. k informs i of his plan to perform action s if the result of the coin matches the bit and d if it does not match.
3. k observes the outcome of the coin.
4. k offers i a bet that the result was heads.

The first diagram would correspond to i's probabilities before any of these actions have taken place. The second diagram would correspond to i's probabilities after the second action, but before the third. The third diagram would correspond to i's probabilities after all four actions have taken place. Perhaps action two, where k informs i of the plan to perform either s or d prompts i to rework his probabilities so that s and d are assigned probabilities. The contribution of the fourth action, where k offers i a bet, may be more convincing. Now i considers it highly unlikely that k would have offered the bet knowing that he would lose, yet i could not have a quantitive grasp of this likelihood, and thus could not assign a probability.

Now in terms of probabilistic epistemic logic, the first model poses a difficulty; if there is any formula for the bit 1, the bit 0, the action s, or the action d, then the set of states making that formula true is not measurable. A temporary fix is proposed in [3], which is to use the inner measure function. If $(S, \mathcal{A}, \mu)$ is a probability space, then the inner measure of μ is μ_* defined by

$$\mu_*(T) = \sup\{\mu(A) : A \in \mathcal{A}, A \subseteq T\}$$

for each $T \subseteq S$. We could alternatively use the outer measure:

$$\mu^*(T) = \inf\{\mu(A) : A \in \mathcal{A}, T \subseteq A\}$$

for each $T \subseteq S$. Inner and outer measures are related according to $\mu^*(T) = 1 - \mu_*(\overline{T})$, where $\overline{T}$ is the complement of T in S. Thus either the inner or outer measure can be taken as primitive in the language, and the other can be defined according to the other. Either the inner or outer measure lets us define the semantics of formulas such as $P_i(\varphi) \geq 1/2$ even when the set of states making φ true is unmeasurable. But as the inner measure and the outer measure need not be a measures themselves, the probability axioms would fail. One suggestion given in [3] is that we explicitly require that the σ-algebra be large enough to contain all sets corresponding to each formula. But non-measurable cases have been considered with a more relaxed set axioms

In defining an update product between a probabilistic epistemic model and an action model, we shall first consider the general case, where sets of states that make the formulas true need not be measurable. In particular, the formulas inducing the update might not correspond to measurable sets (the elements of the σ-algebra). We thus define an "outer (or inner) probability dynamic epistemic logic". To make this task manageable, we restrict the models to be finite.

Definition 3.1 [action model] An *action model* $(\Sigma, \{\xrightarrow{i}\}, \{\mathcal{P}_{i,\sigma}\}, \mathsf{pre})$ is a probabilistic epistemic model with the valuation function $\|\cdot\|$ replaced by a function pre which assigns to each $\sigma \in \Sigma$ a function that assigns to each probabilistic epistemic model a subset of the carrier set of that model. Each element $\sigma \in \Sigma$ is called an *action type.* ◁

We define the update product between a probabilistic epistemic model and an action model in two stages. We first define the product between the original probabilistic epistemic model and an action signature (which is just a probabilistic epistemic frame (no valuation, and without the pre function)), and then relativize the result according to the pre function. The first product is called the unrestricted product. The second is called the relativization.

Definition 3.2 [unrestricted product] The unrestricted product between a probabilistic epistemic model $\mathbf{M}$ and an action model $\mathbf{\Sigma}$ is $\mathbf{M} \otimes_U \mathbf{\Sigma}$ with the following components:

1. $X_\otimes = X \times \Sigma$
2. $(x, \sigma) \xrightarrow{i} (z, \tau)$ iff $x \xrightarrow{i} z$ and $\sigma \xrightarrow{i} t$
3. $\|p\|_\otimes = \|p\| \times \Sigma$
4. We define $\mathcal{P}_{i,(x,\sigma)}$ as follows:

(a) The sample space is the Cartesian product $S_{i,(x,\sigma)} = S_{i,x} \times S_{i,\sigma}$

(b) The σ-algebra $\mathcal{A}_{i,(x,\sigma)}$ is the smallest σ-algebra containing

$$\{A \times B : A \in \mathcal{A}_{i,x}, B \in \mathcal{A}_{i,\sigma}\}$$

(c) The probability measure is defined as

$$\mu_{i,(x,\sigma)}(A) = \sum_{k=1}^{n} \mu_{i,x}(B_k)\mu_{i,\sigma}(C_k)$$

where $B_k \in \mathcal{A}_{i,x}$, $C_k \in \mathcal{A}_{i,\sigma}$, and $A = \biguplus_{i=1}^{n} B_k \times C_k$

◁

This product is a probabilistic epistemic model. The usual definition for product measures (for finite spaces) is given to our new probability measure. Product measures need not be restricted to finite spaces, and hence this unrestricted product can be defined between any probabilistic epistemic model and action model of infinite size.

But the relativization of our probabilistic epistemic model requires some restriction be placed on the probabilistic epistemic model. Requiring the carrier set of the probabilistic epistemic model to be finite is sufficient and still allows us to explore a wealth of examples.

Definition 3.3 [relativization] The relativization of a probabilistic epistemic model $\mathbf{M}$ to $Y \subseteq X$ is given by $\mathbf{M} \otimes_R Y$ with the following components:

1. $X_Y = Y$
2. $x \xrightarrow{i}_Y z$ iff $x \xrightarrow{i} z$ and $x, z \in Y$
3. $\|p\|_Y = \|p\| \cap Y$
4. For $x \in Y$, if $\mu^*_{i,x}(Y) = 0$, then define $\mathcal{P}_{i,x}$ to be the trivial probability space on the singleton x. Otherwise

   (a) $S_{Y\,i,x} = S_{i,x} \cap Y$

   (b) $\mathcal{A}_{Y\,i,x}$ is the σ-algebra generated by $\{A \cap Y : A \in \mathcal{A}_{i,x}\}$

   (c) The probability measure is defined by

$$\mu_{Y\,i,x}(A) = \frac{\mu^*_{i,x}(B)}{\mu^*_{i,x}(Y)}$$

◁

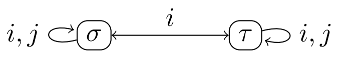
The choice to update using outer measures rather than inner measures is mostly arbitrary. The outer measure, however is less likely to be zero. When the σ-algebra $\mathcal{A}$ of a space $(S, \mathcal{A}, \mu)$ is finite, the outer measure of μ applied to a set $T \subseteq S$ becomes

$$\mu^*(T) = \bigcap\{\mu(A) : A \in \mathcal{A}, T \subseteq A\}$$
$$= \mu(\bigcap\{A : A \in \mathcal{A}, T \subseteq A\})$$

Thus the outer measure of a (not necessarily measurable) set is equal to the measure of an appropriate measurable set. This is not guaranteed in the infinite case. But this property helps us guarantee that the updated function is indeed a measure. The most difficult case is the additivity condition. If $A_1, \ldots, A_n$ is a set of pairwise disjoint sets measurable in the relativized model, let $\hat{A}_i = \bigcap\{B : A_i \subseteq B, B \in \mathcal{A}_{i,x}\}$, where $\mathcal{A}_{i,x}$ is the σ-algebra for the original model. Unlike the A_i, the $\hat{A}_i$ are necessarily measurable in the first space. Also $\hat{A}_1, \ldots, \hat{A}_n$ is pairwise disjoint, for if Y is the set with which we relativized, then $B = \hat{A}_j \cap \hat{A}_k \subseteq \overline{Y}$ (otherwise A_j and A_k would not be disjoint). But $A_j \subseteq \hat{A}_j - B$ and $\hat{A}_j - B \in \mathcal{A}_{i,x}$, thus $\hat{A}_j = \hat{A}_j - B$, and so we conclude that $B = \emptyset$. Also observe that $\widehat{\bigcup A_i} = \bigcup \hat{A}_i$. We can then make use of this and the fact that $\mu^*(C) = \mu(\hat{C})$ for any set C in order to establish the additivity property of the new measure.

Definition 3.4 [update product] Let $\mathbf{\Sigma} = (\Sigma, \{\xrightarrow{i}\}_{i \in \mathbf{I}}, \{\mathcal{P}_{i,x}\}, \mathsf{pre})$ be an action model and $\mathbf{M} = (X, \{\xrightarrow{i}\}, \|\cdot\|, \{\mathcal{P}_{i,x}\})$. Let $Y = \{(x, \sigma) : x \in \mathsf{pre}(\sigma)(\mathbf{M})\}$. The update product between $\mathbf{M}$ and $\mathbf{\Sigma}$ is written $\mathcal{M} \otimes \mathbf{\Sigma}$ and is defined as $(\mathbf{M} \otimes_U \mathbf{\Sigma}) \otimes_R Y$. ◁

Returning to the example above, the action of revealing the bit to k in such a way that i knows k learned something is a semi-private announcement. Similarly, k's learning the result of the flip in such a way that i knows k learned something is also a semi-private announcement. The relational part of the action signature for semi-private announcements may be depicted by the following diagram:

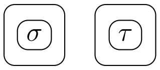


R_1

From each of the two action types, i's probability space is the only probability space where the sample space is that single action type. This action signature may be used for the action models of both stages. For the first stage, the precondition of σ is 1, and the precondition of τ is 0. For the second action signature, the precondition of σ could be H, while the precondition for τ could be T.

But what should be the probability spaces of the action model? Let us assume that one action model will capture both the semi-private announcement of the bit to k and announcement that k plans to do either s or d. Then both i and

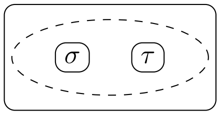
k's probability spaces in the action model could be i and k's probability spaces:



P_1

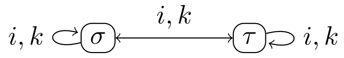
From each of the action types, the probability space is the only space that can be defined over a sample space with one element. Dotted ovals are therefore not needed. That i and k share the same probability spaces agrees with the view that probabilities should be objective.

But let us consider what happens if we break down the transition from M_1 to M_2 into two steps (giving us an intermediate model) and similarly break down the transition from M_2 to M_3 into two steps. The action where k is informed of the bit will still be considered a semi-private announcement, and the relational structure will be the same. The only difference shall be i's probability space, which we change to the following structure:



P_2

The action where i is informed that k plans to do either action s or d will serve the purpose of splitting i's probability space into two. This can be done by using the probability structure P_1. We will also use P_1 for k's probability structure. As k's probability spaces are already split, using P_1 for k as well will not affect k's probability structure in the update model. For the relational structure we use



R_2

As the precondition of σ is the bit 1 and the precondition of τ is the bit 0, the updated model will have the same relational structure as it did right before updating.

We use the same relational and probabilistic structures in the actions model from M_2 to M_3, but we use different preconditions. We may let the precondition for σ be H rather than 1, and we may let the precondition for τ be T rather than 0.

When we consider the language and semantics, we may wish that every formula correspond to a measurable set. But even this might not guarantee that the set Y in definition 3.3 is itself measurable. Consider an action signature for which k's probability space is given by diagram P_2, that is, there are only two measurable sets: the whole set and the empty set. Suppose a probabilistic epistemic model $\mathbf{M}$ has two states: x and y, and k's probability sample space is $\{x, y\}$ and all subsets are measurable. Then in the product measure, the measurable sets are

$$\{\emptyset, \{(x,\sigma),(y,\sigma)\}, \{(x,\tau),(y,\tau)\},\\ \{(x,\sigma),(y,\sigma),(x,\tau),(y,\tau)\}\}.$$

Suppose there were a formula φ for which only x is true, and another formula for which y is true. Then these formulas correspond to measurable sets. Let the function pre reflect these two formulas, by defining $\mathsf{pre}(\sigma)(\mathbf{M}) = x$ and $\mathsf{pre}(\tau)(\mathbf{M}) = y$. Then when taking the full update product, we would be relativizing with respect to the set $Y = \{(x,\sigma),(y,\tau)\}$, which is not measurable. In general, if an action signature has only discrete probability spaces (probability spaces where the σ-algebras are power sets of the sample spaces), then the measurability of the sets $\mathsf{pre}(\sigma)(\mathbf{M})$ for each $\sigma \in \Sigma$, does guarantee that the set Y in definition 3.3 is measurable. It remains to be seen that in an updated model, every formula still corresponds to a measurable set.

4 Conclusion

This paper is a synthesis of two related projects. One is to add a previous time operator to a probabilistic dynamic epistemic logic similar to the one given in [8], and the other is to involve action models and update products in a probabilistic dynamic epistemic logic. Although it appears that these projects are independent, the second project may help support the first. We have so far approached the first project with the initial goal of maintaining simplicity in hope that technical results will be easier to achieve. But sometimes extra structure makes it easier to prove certain results, and we have yet to see if the involvement of unmeasurable sets will facilitate the completeness proof of a probabilistic dynamic epistemic logic with a previous time operator.

The second project explores the possibility and motivation of non-discrete probability measures and updating based on non-measurable sets. We have seen one way to update finite probabilistic epistemic models that are not necessarily discrete upon finite action models that are not necessarily discrete in order to yield a new finite probabilistic epistemic model. Although this updating can guarantee that the updated model is indeed a probabilistic epistemic model, it does not guarantee that in the updated model, all the formulas correspond to measurable sets; we have yet to see which conditions would ensure the updated model does have that property. This is only a concern if we wish to enforce additivity axioms of probability. Otherwise we may have the machinery for a nice inner (or outer) probability dynamic epistemic logic.

As this update product is quite flexible, with non-measurable sets in both probabilistic dynamic epistemic models and action models, questions open up as to how to interpret particular instances of updating. We have looked at an example in [3] to help us with this. While doing so, we distinguished between subjective and objective probabilities and considered breaking down each action into two. The second action for each only affects the probability spaces being updated, and does not affect the structure of the epistemic relations. There so far is no language for this probabilistic dynamic epistemic logic with action models and update products, and in coming up with a language, we should determine what fundamentally is driving this change in probability spaces. May the source of information play an important role, as suggested by the phrasing of the action "k offers i a bet that the result was heads"? What is the essential component of the action phrased "k informs i of his plan to perform action s if the result of the coin matches the bit and d it does not match"? I suggest that in future work, finding more examples will help reveal underlying patterns that will enable us to adequately answer these questions.

References

[1] A. Baltag, L. Moss, and S. Solecki. Logics for epistemic actions: Completeness, decidability, expressivity. ms. Indiana University, 2003.

[2] A. Baltag and L. S. Moss. Logics for epistemic programs. *Synthese*, 139(2, Knowledge, Rationality & Action):165–224, 2004.

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

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

[5] T. French, R. van der Meyden, and M. Reynolds. Axioms for logics of knowledge and past time: Synchrony and unique initial states. *Proceedings of Advances in Modal Logic*, 2005.

[6] D. Gerbrandy. *Bisimulations on Planet Kripke.* PhD thesis, ILLC, Universiteit van Amstardam, 1998.

[7] J. Halpern and M. Tuttle. Knowledge, probability, and adversaries. *Journal of the ACM*, 40(4):917–962, 1993.

[8] B. P. Kooi. Probabilistic dynamic epistemic logic. *J. Logic Lang. Inform.*, 12(4):381–408, 2003. Special issue on connecting the different faces of information.

[9] J. Miller and L. Moss. Undecidability of iterated modal relativization. *Studia Logica*, 79(3):373–407, April 2005.

[10] R. Parikh and R. Ramanujam. A knowledge based semantics of messages. *J. Logic Lang. Inform.*, 12(4):453–467, 2003. Special issue on connecting the different faces of information.

[11] J. Plaza. Logics of public communications. *Synthese*, 158(2, Knowledge, Rationality & Action):165–179, 2007.

[12] J. Sack. Logic for update products and steps into the past. ms., 2007.

[13] J. Sack. Temporal languages for epistemic programs. *Journal of Logic, Language and Information*, 17(2):183–216, April 2008.

[14] A. Yap. Product update and looking backward. ms. www.illc.uva.nl/lgc/papers/bms-temporal.pdf, 2005.

Diachronic Uncertainty and Equivalence Notions for ETL Models of Extensive Form Games

Alistair Isaac
Stanford University

Abstract

Diachronic uncertainty, uncertainty about where an agent falls in time, poses interesting conceptual difficulties. Although the agent is uncertain about where she falls in time, nevertheless, she can only be uncertain at a particular moment in time. This conceptual paradox can be resolved by providing an equivalence notion between models with diachronic uncertainty and models with synchronic uncertainty. The former are interpreted as capturing the causal structure of a situation, while the latter are interpreted as capturing its epistemic structure. We consider some of the properties of models for epistemic temporal logic which make them a suitable formalism for investigating such equivalence notions. We conclude with a simple example.

1. Introduction

Philosophers and Game Theorists have become increasingly interested in problems of diachronic uncertainty. In particular, if the agent knows at one state in a decision problem that at a later state she will forget or otherwise lose awareness of where she is in time, how should the agent compute appropriate actions and / or beliefs? In the game theory literature, the paradigmatic case of such forgetfulness is the Absent-Minded Driver ([Pi97]); in the philosophical literature, much discussion has centered around Sleeping Beauty ([El00]). Conceptually, however, agents can only be uncertain at a point in time; in other words, *all uncertainty is synchronic*. Any realistic model of a decision making agent should describe the succession of epistemic states through which the agent passes. Each one of these states will be synchronic, in the sense that it occurs at a distinct point in time, although these synchronic uncertainties may be uncertainties about where the agent falls in time. Given a specification of a decision problem involving diachronic uncertainty, we may ask: (i) How can we convert this into a problem involving only synchronic uncertainties? (ii) How should probabilities be assigned within the new information partitions? This document will focus on the properties of extensive form games when interpreted as ETL models as considerations towards an answer to (i).

2. Interpreting ETL Models

In order to make these questions more precise, we must work within a unified framework. In the game theory literature, all modeling of such problems uses the formalism of extensive form games. In the philosophical literature, although a vanilla Bayesianism lurks in the background of the debate, no one formalism dominates discussion; a crucial ingredient to the points of contention, however, is the use of propositions which can change truth value through time (in particular, "self-locating" propositions, which refer indexically to the agent's position in the temporal structure of the world). Epistemic temporal logics lie at a happy meeting ground between these two approaches. Syntactically, epistemic temporal languages are powerful enough to express uncertainty about where the agent falls in the temporal order. Semantically, the models of epistemic temporal logics are rich enough to include extensive form games as a special case. Furthermore, epistemic temporal logics have natural probabilistic extensions. In this section, we characterize pertinent subsets of the space of epistemic temporal models.

Epistemic temporal models are forests partitioned into equivalence classes for each agent. The interpretation of these partitions is that the agent is unable to distinguish between worlds in a partition. We refer to these as uncertainty partitions or information sets. Given a set of events Σ, Σ^* is the set of strings over Σ. Elements of Σ^* are called histories, states, or worlds. A set $\Pi \subseteq \Sigma^*$ is a protocol if it is closed under finite prefixes. So, a protocol Π is just a forest, and if Π contains the empty set, it is a tree. Call the set of agents A. With each agent $i \in A$, we identify an equivalence relation $\sim_i$. These equivalence relations partition the nodes of Π into sets of worlds which are indistinguishable for agent i.

DEFINITION 1: An **ETL frame** is a tuple $\langle \Sigma, \Pi, \{\sim_i\}_{i\in A} \rangle$ where Σ is a set of events, Π is a protocol, and for each $i \in A$, $\sim_i$ is an equivalence relation on Π.

DEFINITION 2: An **ETL model** is a tuple $\langle \Sigma, \Pi, \{\sim_i\}_{i\in A}, V \rangle$ where Σ, Π, and $\{\sim_i\}_{i\in A}$ are an ETL frame and *V* is a valuation function from the set of atomic formulae *At* into the power set of Π, $V : At \longrightarrow 2^{\Pi}$.

Conceptually, we can think of an ETL model as a specification of the causal structure of the world (the ordering of possible events), which is then decorated with epistemic relations. We are interested, however, in what will happen if we prioritize epistemic structure rather than causal structure. What happens if we insist that models characterize the sequence of the agent's epistemic states, even when this sequence conflicts with the sequence of events? This is the case with diachronic uncertainty. If an agent *i* is uncertain at t_2 whether the time is t_1 or t_2, then some events which the agent considers possible will not in fact be possible (namely, those events which can only immediately follow t_1). Our goal is to consider this conceptual transformation within the framework of epistemic temporal models by considering equivalence classes of ETL models with respect to intuitively motivated notions of situation equivalence.

The space of ETL models is quite rich, and characteristics of its fine structure have been charted in [vB08] and [vB06]. [vB08] characterizes the subset of ETL models which are equivalent to models for dynamic epistemic logic (DEL). From [vB06] we know that this fragment of ETL preserves some nice computational properties (in particular, so long as we limit ourselves to a future modality which can only see ahead one step in time (this is the essence of DEL), we preserve decidability). [vB08] distinguishes two types of DEL-generated protocols: uniform protocols and state-dependent protocols. These notions are of interest for our purposes in the constraints they place on permissable uncertainty partitions $\sim_i$. In order to emphasize this aspect of the situation, we define four classes of ETL models.

DEFINITION 3: An ETL model $\langle \Sigma, \Pi, \{\sim_i\}_{i\in A}, V \rangle$ is

(i) **state-dependent** *iff* there is no general restriction on the events that can occur after any history

(ii) **agent-dependent** *iff* for any agent i, event e, and histories h, h', if $h \sim_i h'$ and $he \in \Pi$, then $h'e \in \Pi$

(iii) **cardinality-dependent** *iff* for any agent i and histories h, h', if $h \sim_i h'$, then $|\{h'' \in \Pi | \exists e\, (h'' = he)\}| = |\{h'' \in \Pi | \exists e\, (h'' = h'e)\}|$

(iv) **uniform** *iff* if $p \in At$ is a *precondition* of event *e* and $h \in V(p)$, then $he \in \Pi$

In standard DEL models, the events possible at a world are characterized by a function $E : At \longrightarrow \Sigma$. If $E(p) = e$, then the proposition *p* represents a *precondition* of the event e, and e is possible at any world $h \in V(p)$. We write $pre(e)$ for $E^{-1}(e)$, *i.e.* the set of preconditions of e. The notion of a state-dependent DEL protocol generalizes this idea by replacing the function from atomic formulae into the space of events with a function from histories *h* into the space of events (*i.e.* from Π into Σ). This is the appropriate interpretation of a state-dependent ETL model: it is a model in which the events following a given history are not constrained in any systematic way by other features of the model. Agent-dependent, cardinality-dependent, and uniform models are all special cases of state-dependent models where the function from histories into events is somehow constrained. These notions will help us distinguish various definitions of extensive form games in the following section. Before embarking on that discussion, let us expand our repertoire with some further notions from [vB08].

DEFINITION 4: An ETL model $\langle \Sigma, \Pi, \{\sim_i\}_{i\in A}, V \rangle$ satisfies

(i) **strong synchronicity** *iff* for all histories h, h', if for some agent i, $h \sim_i h'$, then $len(h) = len(h')$, where $len(h)$ is just the number of events in h

(ii) **weak synchronicity** *iff* for all histories h, h', if for some agent i, $h \sim_i h'$, then h is not a proper prefix of h'[1]

(iii) **perfect recall** *iff* for all histories h, h' and events e, e', if $he \sim_i h'e'$, then $h \sim_i h'$

(iv) **local no miracles** *iff* for all histories h, h', h'', h''', agents i, and events e, e', if $he \sim_i h'e'$, $h' \sim^* h''$, and $h'' \sim_i h'''$, then $h''e \sim_i h'''e'$, where $\sim^*$ is the reflexive transitive closure of the $\sim_i$ relations

[vB08] shows that the class of ETL models generated from uniform DEL protocols is just that which satisfies strong synchronicity, perfect recall, local no miracles, and local epistemic bisimulation invariance, and the class of ETL models generated from state-dependent DEL protocols is just that which satisfies propositional stability, strong syn-

[1] "Weak synchronicity" does not appear in [vB08], but it will be helpful in our discussion of games below. Strong synchronicity implies weak synchronicity, but not vice versa. In some situations, we can transform a model satisfying weak synchronicity into one satisfying strong synchronicity by simply introducing dummy nodes which bring asynchronous uncertainty partitions into sync (*c.f.* the introduction of "'dummy' chance moves with one alternative" in [Ku53], 51). For an example of a game which cannot be brought into synchrony using this method, see [Pi97], example 6.

chronicity, perfect recall, and local no miracles.[2] Tomohiro Hoshi (this conference) has investigated these distinctions in more detail for the subset of DEL known as public announcement logic (PAL).

3. Uncertainty in Extensive Form Games

Much of the game theory literature on uncertainty has focused on uncertainty about other players' moves. Since one usually assumes that players alternate turns in extensive form games, this uncertainty can be modeled via weakly synchronic equivalence relations for each agent. Furthermore, modeling choices have been constrained by conceptual analysis of the notion of indistinguishability itself: *what constraint appropriately captures the idea that the agent cannot distinguish between two states?*

[Th52]and [Ku53] define information partitions in extensive form games such that two constraints are met. First, no two worlds in the information partition may lie on the same branch. Second, at each world in the partition, the cardinality of the set of potentially occurring events must be the same. [Pi97] drops the first constraint, in order to allow for forgetfulness, yet strengthens the second constraint by stipulating not just that the cardinality of the set of possible events be the same for each world in an information set, but that the set of possible events be *identical* for each world. Thus, [Th52]and [Ku53] define models which are cardinality-dependent and satisfy weak synchronicity, while [Pi97] defines models which are agent-dependent.

These constraints are motivated by the idea that an information set models a situation in which an agent must *act*, although she does not know the current state of the world. Actions are just distinguished events, events caused by some particular agent. If different (or different numbers of) actions are available to an agent at two nodes in the game tree, then the agent can use her knowledge of which actions are available to her to distinguish these histories from each other. Thus, if two states of the world are indistinguishable to an agent, then the agent must have the same actions available to her at each one. Agent-dependency and cardinality-dependency are attempts to capture this intuition.

As noted above, if an agent is uncertain between two worlds at different points in time t_1 and t_2, then it must be the case that different events are possible at the two worlds. However, it may nevertheless be the case that the agent has the same set of available actions. In the Absent Minded Driver example, a man has left a bar drunk and forgets while driving home whether he has already made his turn or not. The problem is usually modeled with an information set including two indistinguishable intersections. The driver must pass through these intersections in sequence, so he will encounter them at different times. Thus, there must be some events possible at one which are not possible at the other. However, in terms of actions, the driver only has two options: turn or go straight. So, if our model only includes the actions available to the agent, excluding other events, it will satsify agent-dependence.

The constraints on modeling in the game theory literature are motivated by extrinsic considerations. In thinking about agents performing actions in a game, and what it might mean for an agent to be uncertain between possible states of play, concept analysis dictates that either agent-dependency or cardinality-dependency constrain permissable models. In state-dependent ETL models, however, we have as modeling tools both a valuation function *V* and an event function *E*. If we retain the notion of preconditions at the conceptual level, then we can characterize a situation in which an agent believes e to be possible, when in fact it is not.

> DEFINITION 5: The **possible events** $E_i^P(h)$ for an agent i at a history h in an uncertainty partition $I = \{w_1, ..., h, ..., w_n\}$ are just those events e such that $\bigcup I \subseteq V(pre(e))$

If we are in a state-dependent ETL model $\left\langle \Sigma, \Pi, \{\sim_i\}_{i \in A}, V \right\rangle$, then there is no constraint that for any $h \in \Pi$, and agent i, $E_i^P(h) = \{e \in \Sigma | he \in \Pi\}$. In other words, the set of possible events *from the agent's perspective* need not equal the *actual* possibilities allowed by the model. Of course, this move deflates the role of preconditions; they no longer play a structural role in constraining the model, but merely act as a bookkeeping device for tracking agent expectations.

4. An Example: the Absent Minded Driver

Perhaps the simplest example of a game with diachronic uncertainty is the Absent Minded Driver (*fig. 1*).

The driver begins at Ø and drives straight. He passes through two intersections, w_1 and w_2. At each intersection, he can either continue to drive straight, or turn. If he turns at the first intersection, he arrives in the bad part of town, *B*. If he turns at the second intersection, he arrives home, *H*, as desired. If the driver continues straight through both intersections, he must stay at a motel, *M*. It is stipulated, however, that the driver cannot distinguish the first and second intersections; in other words, he cannot remember whether he has turned or not.

In light of the considerations raised above, we might ask whether there is a distinct game tree, equivalent to the Absent Minded Driver *in the relevant respects*, but prioritizing epistemic states. Such a game tree would satisfy synchronicity, in line with the analysis of uncertainty as a fundamentally synchronic notion, yet would preserve *in some*

[2] We have omitted definitions of local epistemic bisimulation invariance and propositional stability as they are not discussed in the sequel.

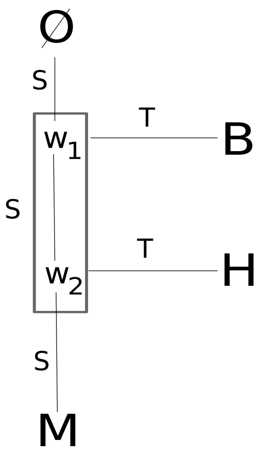


Figure 1. The Absent Minded Driver

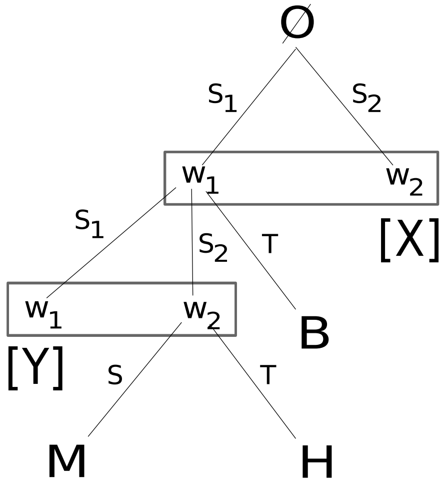


Figure 2. The *Epistemic* Absent Minded Driver

sense the causal structure of the original model. We might call such a pair of models *epistemically equivalent*.

DEFINITION 6: Two ETL models M_1 and M_2 are **epistemically equivalent** *iff*
(i) all agents $i \in A_1 \bigcap A_2$ pass through uncertainty states in the same order (with possible duplications) in M_1 and M_2
(ii) all events $e \in \Sigma_1 \bigcap \Sigma_2$ occur in the same order (with possible duplications) in M_1 and M_2

Consider, for example, the ETL model depicted in *figure 2*. In this model, the agent passes through the same sequence of epistemic states as in the Absent Minded Driver; the uncertainty partition from *figure 1* has merely been duplicated to capture the fact that the agent will experience it at two distinct points in time. There are two questionable modeling choices here, however. First, what is the significance of moves s_1 and s_2? Second, how should one interpret [X] and [Y]? s_1 and $s_2$represent the epistemic disjunct between the choice of a single action (go straight), and the two resulting possibilities, w_1 and w_2. Rather than consider these as two distinct moves, or events, we may instead wish to include separate moves by the driver and a chance player. The driver chooses *s*, but *chance* (or, perhaps it would be better to call him *confusion*) plays to increase the possibilities the agent countenances. This "move" must be interpreted as an *epistemic*, rather than *physical*, event: *the event of forgetting*.

The worlds above [X] and [Y] are those which the agent erroneously believes possible. We have several modeling options available to us here, though we consider only three. First, we might simply leave these as terminal nodes. Second, we may replace [X] and [Y] with unitary chance moves to some distinguished world; this world might be interpreted as an impossible state. Both these strategies would capture the agent's error, or the physical impossibility of any events occurring at [X] and [Y]. Both these options fail to connect with game theoretic models, however. The first, because game theoretic models never allow uncertainties over terminal states; the second, because transitions to an "impossible" world would involve adding a new terminal state, but one without any coherent notion of payoff attached to it.

A third option would connect quite nicely with the game theoretic literature, in particular [Th52]. Thompson defines equivalence classes of extensive form games with respect to the corresponding strategic game. He suggests four transformations on game trees which preserve strategic form. Any two extensive form games which share a strategic form can be transformed into each other via some sequence of these four transformations. Since Thompson only considers models which satisfy weak synchronicity, the Absent Minded Driver does not fall within his paradigm. One strategy for dealing with [X] and [Y] suggested by Thompson's transformations is simply to copy the game tree from under the other node in the uncertainty partition to the position under the "impossible" node. Conceptually, we might interpret this as capturing the fact that the *actual* possibilities are the same from both states in the driver's uncertainty partition as he is only *actually* at one of them. If we apply this strategy plus that described above for adding moves by a chance player, we derive *figure 3*. In *figure 3*, the sequence of epistemic states in *figure 1* is preserved, as is the sequence of actual events. We have had to add a chance move, interpreted as the epistemic process of forgetting, but doing so has allowed us to produce a model which is sus-

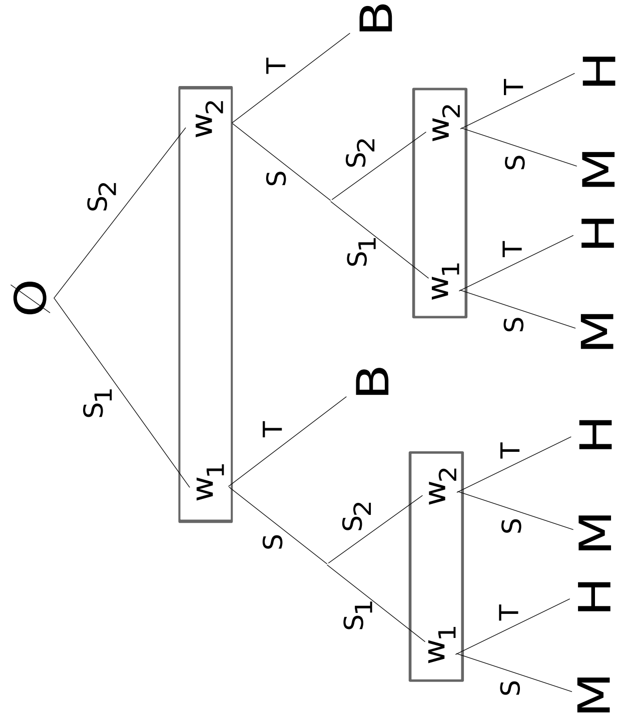


Figure 3. The *Epistemic* **Absent Minded Driver (***final***)**

ceptible to the transformations described in [Th52].

Final remark: the strategy described for transforming the model in *figure 1* to the model in *figure 3* will not work in all cases. Consider, for example, *crossed* ETL models.

> DEFINITION 7: An ETL model $\left\langle \Sigma, \Pi, \{\sim_i\}_{i \in A}, V \right\rangle$ is **crossed** *iff* there exists an agent $i \in A$ and histories h, h', h'', and h''' with $h \neq h''$ such that $hh', h''h''' \in \Pi$, $h \nsim_i h'$, $h'' \nsim_i h'''$, $hh' \sim_i h''$, and $h \sim_i h''h'''$

Example 6 of [Pi97] is a crossed model. If one attempts to implement the transformation strategy described above on a crossed ETL model, one will produce an infinite tree which cycles through the two partitions $\{h, h''h'''\}$ and $\{hh', h''\}$. It remains to be seen what precise class of constraints on ETL models characterizes just those susceptible to the above described transformation. At the very least, such models must be *uncrossed*.

References

[El00] Elga, Adam (2000) “Self-locating Belief and the Sleeping Beauty Problem,” in *Analysis* 60.2; 143-7.

[Ku53] Kuhn, Harold W. (1953) “Extensive Games and the Problem of Information,” reprinted from *Contributions to the Theory of Games II* in *Classics in Game Theory* (1997), Harold W. Kuhn (ed.); 46-68.

[Pi97] Piccione, Michele and Ariel Rubinstein (1997) “On the Interpretation of Decision Problems with Imperfect Recall,” in *Games and Economic Behavior* 20; 3-24.

[Th52] Thompson, F. B. (1952) “Equivalence of Games in Extensive Form,” RAND Memo RM-759, in *Classics in Game Theory* (1997), Harold W. Kuhn (ed.); 36-45.

[vB06] van Benthem, Johan and Eric Pacuit (2006) “The Tree of Knowledge in Action: Towards a Common Perspective,” in *Proceedings of Advances in Modal Logic* (AiML 2006), King’s College Press, Guido Governatori, Ian Hodkinson and Yde Venema (eds.).

[vB08] van Benthem, Johan, Jelle Gerbrandy, Tomohiro Hoshi, and Eric Pacuit (2008) “Merging Frameworks for Interaction,” unpublished manuscript available online at staff.science.uva.nl/˜johan/deletl-march2008.pdf

Multi-agent belief dynamics: bridges between dynamic doxastic and doxastic temporal logics

Johan van Benthem
ILLC Amsterdam
Stanford University
johan@science.uva.nl

Cédric Dégremont*
ILLC Amsterdam
cdegremo@science.uva.nl

Abstract

In this paper, we compare two modal frameworks for multi-agent belief revision: dynamic doxastic logics computing stepwise updates and temporal doxastic logics describing global system evolutions, both based on plausibility pre-orders. We prove representation theorems showing under which conditions a doxastic temporal model can be represented as the stepwise evolution of a doxastic model under successive 'priority updates'. We define these properties in a suitable doxastic-temporal language, discuss their meaning, and raise some related definability issues.

Analyzing the behavior of agents in a dynamic environment requires describing the evolution of their knowledge as they receive new information. Moreover agents entertain beliefs that need to be revised after learning new facts. I might be confident that I will find the shop open, but once I found it closed, I should not crash but rather make a decision on the basis of new consistent beliefs. Such beliefs and information may concern ground-level facts, but also beliefs about other agents. I might be a priori confident that the price of my shares will rise, but if I learn that the market is rather pessimistic (say because the shares fell by 10%), this information should change my higher-order beliefs about what other agents believe.

Tools from modal logic have been successfully applied to analyze knowledge dynamics in multi-agent contexts. Among these, Temporal Epistemic Logic [23], [19]'s Interpreted Systems, and Dynamic Epistemic Logic [2] have been particularly fruitful. A recent line of research [11, 10, 9] compares these alternative frameworks, and [10] presents a representation theorem that shows under which conditions a temporal model can be represented as a dynamic one. Thanks to this link, the two languages also become comparable, and one can merge ideas: for example, a new line of research explores the introduction of protocols into the logic of public announcements PAL, as a way of modeling informational processes (see [9]).

To the best of our knowledge, there are no similar results yet for multi-agent belief revision. One reason is that dynamic logics of belief revision have only been well-understood recently. But right now, there is work on both dynamic doxastic logics [5, 3] and on temporal frameworks for belief revision, with [14] as a recent example. The exact connection between these two frameworks is not quite like the case of epistemic update. In this paper we make things clear, by viewing belief revision as priority update over plausibility *pre-orders*. This correspondence allows for similar language links as in the knowledge case, with similar precise benefits.

We start in the next section with background about earlier results and basic terminology. In section 2 we give the main new definitions needed in the paper. Section 3 presents the key temporal doxastic properties that we will work with. In section 4 we state and prove our main result linking the temporal and the dynamic frameworks, first for the special case of *total* pre-orders and then in general. We also discuss some variations and extensions. In section 5 we introduce formal languages, providing an axiomatization for our crucial properties, and discussing some related definability issues. We state our conclusions and mention some further applications and open problems in the last section.

1 Introduction: background results

Epistemic temporal trees and dynamic epistemic logics with product update are complementary ways of looking at multi-agent information flow. Representation theorems linking both approaches were proposed for the first time in [6]. A nice presentation of these early results can be found in [21, ch5]. We start with one recent version from [9], referring the reader to that paper for a proof, as well as generalizations and variations.

[*] The second author was supported by a GLoRiClass fellowship of the European Commission (Research Training Fellowship MEST-CT-2005-020841).

Definition 1.1 [Epistemic Models, Event Models and Product Update]

- An *epistemic model* $\mathcal{M}$ is of the form $\langle W, (\sim_i)_{i\in N}, V\rangle$ where $W \neq \emptyset$, for each $i \in N$, $\sim_i$ is a relation on W, and $V : Prop \to \wp(H)$.
- An *event model* $\epsilon = \langle E, (\sim_i)_{i\in N}, \mathtt{pre}\rangle$ has $E \neq \emptyset$, and for each $i \in N$, $\sim_i$ is a relation on W. Finally, there is a precondition map $\mathtt{pre} : E \to \mathcal{L}_{EL}$, where $\mathcal{L}_{EL}$ is the usual language of epistemic logic.
- The *product update* $\mathcal{M}\otimes\epsilon$ of an epistemic model $\mathcal{M} = \langle W, (\sim'_i)_{i\in N}, V\rangle$ with an event model ϵ is the model $\langle E, (\sim_i)_{i\in N}, \mathtt{pre}\rangle$, whose worlds are pairs (w, e) with the world w satisfying the precondition of the event e, and accessibilities defined as:

$$(w, e) \sim'_i (w', e') \text{ iff } e \sim_i e', w \sim_i w'$$

◁

Intuitively epistemic models describe what agents currently know while the product update describe the new multi-agent epistemic situation after some epistemic event has taken place. Nice intuitive examples are in [1].

Next we turn to the epistemic temporal models introduced by [23]. In what follows, Σ^* is the set of finite sequences on any set Σ, which forms a branching 'tree'.

Definition 1.2 [Epistemic Temporal Models] An *epistemic temporal model* (ETL model for short) $\mathcal{H}$ is of the form $\langle \Sigma, H, (\sim_i)_{i\in N}, V\rangle$ where Σ is a finite set of events, $H \subseteq \Sigma^*$ and H is closed under non-empty prefixes. For each $i \in N$, $\sim_i$ is a relation on H, and there is a valuation $V : Prop \to \wp H$. ◁

The following epistemic temporal properties drive [9]'s main theorem.

Definition 1.3 Let $\mathcal{H} = \langle \Sigma, H, (\sim_i)_{i\in N}, V\rangle$ be an ETL model. $\mathcal{H}$ satisfies:

- **Propositional stability** if, whenever h is a finite prefix of h', then h and h' satisfies the same proposition letters.
- **Synchronicity** if, whenever $h \sim h'$, we have $len(h) = len(h')$.

Let $\sim^*$ be the reflexive transitive closure of the relation $\bigcup_{i\in N} \sim_i$:

- **Local Bisimulation Invariance** if, whenever $h \sim^* h'$ and h and h' are epistemically bisimilar[1], we have $h'e \in H$ iff $he \in H$.
- **Perfect Recall** if, whenever $ha \sim_i h'b$, we also have $h \sim_i h'$.
- **Local No Miracles** if, whenever $ga \sim_i g'b$ and $g \sim^* h \sim_i h'$, then for every $h'a, hb \in H$, we also have $h'a \sim_i hb$.

◁

These properties describe the idealized epistemic agents that are presupposed in dynamic epistemic logic:

Theorem 1.4 (van Benthem et al. [9]) *Let $\mathcal{H}$ be an ETL model, $\mathcal{M}$ an epistemic model, and the 'protocol' P a set of finite sequences of pointed events models closed under prefixes. We write $\otimes$ for* product *update. Let $Forest(M, P) = \bigcup_{\vec{\epsilon}\in P} M \otimes \vec{\epsilon}$ be the 'epistemic forest generated by' $\mathcal{M}$ and sequential application of the events in P.* [2] *The following are equivalent:*

- *$\mathcal{H}$ is isomorphic to $Forest(M, P)$.*
- *$\mathcal{H}$ satisfies propositional stability, synchronicity, local bisimulation invariance, Perfect Recall, and Local No Miracles.*

Thus, epistemic temporal conditions describing idealized epistemic agents characterize just those trees that arise from performing iterated product update governed by some protocol. [9] and [21, ch5] have details.

Our paper extends this analysis to the richer case of belief revision, where plausibility orders of agents evolve as they observe possibly surprising events. We prove two main results, with variations and extensions:

Theorem 1.5 *Let $\mathcal{H}$ be a doxastic temporal model, $\mathcal{M}$ a plausibility model, $\vec{\epsilon}$ a sequence of event models, and $\otimes$* priority *update. The following are equivalent, where the notions will of course be defined later:*

1. *$\mathcal{H}$ is isomorphic to the forest generated by $\mathcal{M} \otimes \vec{\epsilon}$*
2. *$\mathcal{H}$ satisfies propositional stability, synchronicity, invariance for bisimulation, as well as principles of Preference Propagation, Preference Revelation and Accommodation.*

Theorem 1.6 *Preference Propagation, Preference Revelation and Accommodation are definable in an extended doxastic modal language.*

[1] The reader is referred to Subsection 3.1 for a precise definition of bisimulation invariance.

[2] For a more precise definition of this notion, see Section 2 below.

2 Definitions

We now turn to the definitions needed for the simplest version of our main representation theorem, postponing matching formal languages to Section 5. In what follows, let $N = \{1, \dots, n\}$ be a finite set of agents.

2.1 Plausibility models, event models and priority update

As for the epistemic case, we first introduce static models that encode the current prior (conditional) beliefs of agents. These carry a pre-order $\leq$ between worlds encoding a plausibility relation. Often this relation is taken to be total, but when we think of elicited beliefs as *multi-criteria decisions*, a pre-order allowing for incomparable situations may be all we get [18]. We will therefore assume reflexivity and transitivity, but not totality.

As for notation: we write $a{\simeq}b$ ('indifference') if $a \leq b$ and $b \leq a$, and $a < b$ if $a \leq b$ and $b \not\leq a$.

Definition 2.1 [Doxastic Plausibility Models] A *doxastic plausibility model* $\mathcal{M} = \langle W,$ $(\preceq_i)_{i\in N}, V\rangle$ has $W \neq \emptyset$, for each $i \in N$, $\preceq_i$ is a pre-order on W, and $V : Prop \to \wp H$. ◁

We now consider how such models evolve as agents observe events.

Definition 2.2 [Plausibility Event Model] A *plausibility event model* (event model, for short) ϵ is a tuple $\langle E, (\preceq_i)_{i\in N}, \mathtt{pre}\rangle$ with $E \neq \emptyset$, each $\preceq_i$ a pre-order on E, and $\mathtt{pre} : E \to \mathcal{L}$, where $\mathcal{L}$ is a doxastic language. [3] ◁

Definition 2.3 [Priority Update; [3]]
Priority update of a plausibility model $\mathcal{M} = \langle W, (\preceq_i)_{i\in N}, V\rangle$ and an event model $\epsilon = \langle E, (\preceq_i)_{i\in N}, \mathtt{pre}\rangle$ is the plausibility model $\mathcal{M} \otimes \epsilon = \langle W', (\preceq'_i)_{i\in N}, V'\rangle$ defined as follows:

- $W' = \{(w, e) \in W \times E \mid \mathcal{M}, w \Vdash \mathtt{pre}(e)\}$
- $(w, e) \preceq'_i (w', e')$ iff either $e{\prec_i}e'$, or $e{\simeq_i}e'$ and $w \preceq_i w'$
- $V'((s, e)) = V(s)$

Here, the new plausibility relation is still a pre-order. ◁

The idea behind priority update is that beliefs about the last event override prior beliefs. If the agent is indifferent, however, the old plausibility order applies. More motivation can be found in [3, 8].

2.2 Doxastic Temporal Models

Definition 2.4 [Doxastic Temporal Models] A *doxastic temporal model* ($DoTL$ model for short) $\mathcal{H}$ is of the form $\langle \Sigma, H, (\leq_i)_{i\in N}, V\rangle$, where Σ is a finite set of events, $H \subseteq \Sigma^*$ is closed under non-empty prefixes, for each $i \in N$, $\leq_i$ is a pre-order on H, and $V : Prop \to \wp H$. ◁

Our task is to identify just when a doxastic temporal model is isomorphic to the 'forest' generated by a sequence of priority updates:

2.3 Dynamic Models Generate Doxastic Temporal Models

Definition 2.5 [$DoTL$ model generated by updates]
Each initial plausibility model $\mathcal{M} = \langle W, (\preceq_i)_{i\in N}, V\rangle$ and sequence of event models $\epsilon_j = \langle E_j, (\preceq^j_i)_{i\in N}, \mathtt{pre}_j\rangle$ yields a *generated DoTL plausibility model* $\langle \Sigma, H, (\leq_i)_{i\in N}, \mathbf{V}\rangle$ as follows:

- Let $\Sigma := \bigcup_{i=1}^m e_i$.
- Let $H_1 := W$ and for any $1 < n \leq m$ let $H_{n+1} := \{(we_1 \dots e_n) | (we_1 \dots e_{n-1}) \in H_n$ and $\mathcal{M} \otimes \epsilon_1 \otimes \dots \otimes \epsilon_{n-1} \Vdash \mathtt{pre}_n(e_n)\}$. Finally let $H = \bigcup_{1\leq k\leq m} H_k$.
- If $h, h' \in H_1$, then $h \leq_i h'$ iff $h \preceq^{\mathcal{M}}_i h'$.
- For $1 < k \leq m$, $he \leq_i h'e'$ iff 1. $he, h'e' \in H_k$, and 2. either $e{\prec^k_i}e'$, or $e{\simeq^k_i}e'$ and $h \leq_i h'$.
- Let $wh \in \mathbf{V}(p)$ iff $w \in V(p)$.

This is a temporal doxastic model as above. ◁

Now come the key doxastic temporal properties of our idealized agents.

3 Frame Properties for Priority Updaters

We first introduce the notion of bisimulation, modulo a choice of language.

3.1 Bisimulation Invariance

Definition 3.1 [$\leq$-Bisimulation]
Let $\mathcal{H}$ and $\mathcal{H}'$ be two $DoTL$ plausibility models $\langle H, (\leq_1, \dots, \leq_n), V\rangle$ and $\langle H', (\leq'_1, \dots, \leq'_n), V'\rangle$ (for simplicity, assume they are based on the same alphabet Σ). A relation $Z \subseteq H \times H'$ is a $\leq$-Bisimulation if, for all $h \in H$, $h' \in H'$, and all $\leq_i$ in $(\leq_1, \dots, \leq_n)$,

(prop) h and h' satisfy the same proposition letters,

[3] This definition is incomplete without specifying the relevant language, but all that follows can be understood by considering the formal language as a 'parameter'.

(zig) If hZh' and $h \leq_i j$, then there exists $j' \in H'$ such that jZj' and $h' \leq'_i j'$,

(zag) If hZh' and $h' \leq'_i j'$, then there exists $j \in H$ such that jZj' and $h \leq_i j$.

If Z is a $\leq_n$-bisimulation and hZh', we will say that h and h' are $\leq$-bisimilar. ◁

Definition 3.2 [$\leq$-Bisimulation Invariance] A $DoTL$ model $\mathcal{H}$ satisfies *$\leq$-bisimulation invariance* if, for all $\leq$-bisimilar histories $h, h' \in H$, and all events e, $h'e \in H$ iff $he \in H$. ◁

3.2 Agent-Oriented Frame Properties

In the following we drop agent labels and the "for each $i \in N$" for the sake of clarity. Also, when we write ha we will always assume that $ha \in H$. We will make heavy use of the following notion:

Definition 3.3 [Accommodating Events]

Two events $a, b \in \Sigma$ are *accommodating* if, for all $ga, g'b$, $(g \leq g' \leftrightarrow ga \leq g'b)$ and similarly for $\geq$, i.e., a, b preserve and anti-preserve plausibility.

◁

Definition 3.4 Let $\mathcal{H} = \langle \Sigma, H, (\leq_i)_{i \in N}, V \rangle$ be a $DoTL$ model. $\mathcal{H}$ satisfies:

- **Propositional stability** if, whenever h is a finite prefix of h', then h and h' satisfy the same proposition letters.
- **Synchronicity** if, whenever $h \leq h'$, we have $len(h) = len(h')$.

The following three properties trace the belief revising behavior of agents in doxastic trees.

- **Preference Propagation** if, whenever $ja \leq j'b$, then $h \leq h'$ implies $ha \leq h'b$.
- **Preference Revelation** if, whenever $jb \leq j'a$, then $ha \leq h'b$ implies $h \leq h'$.
- **Accommodation** if, a and b are *accommodating* whenever both $ja \leq j'b$ and $ha \not\leq h'b$.

◁

These properties - and in particular the last one - are somewhat trickier than in the epistemic case, reflecting the peculiarities of priority update in settings where incomparability is allowed. But we do have:

Fact 3.5 *If $\leq$ is a* total *pre-order and $\mathcal{H}$ satisfies Preference Propagation and Preference Revelation, then $\mathcal{H}$ satisfies Accommodation.*

Proof. From left to right. Assume that $g \leq g'$ and $ja \leq j'b$. By Preference Propagation, $ga \leq g'b$. Now assume that $ha \not\leq h'b$. Then by totality, $h'b \leq ha$. Since $g \leq g'$, it follows by Preference Propagation that $gb \leq g'a$.

From right to left, assume that $gb \leq g'a$ and that $ja \leq j'b$. It follows by Preference Revelation that $g \leq g'$. Now assume that $ga \leq g'b$ (1) and $ha \not\leq h'b$ (2). From (2), it follows by totality that $h'b \leq ha$ (3). But if (3) and (1), then by Preference Revelation we have $g \leq g'$. QED

We can also prove a partial converse without totality:

Fact 3.6 *If $\mathcal{H}$ satisfies Accommodation, it satisfies Preference Propagation.*

Proof. Let $ja \leq j'b$ (1) and $h \leq h'$ (2). Assume that $ha \not\leq h'b$. Then by Accommodation, for every $ga, g'b, g \leq g' \leftrightarrow ga \leq g'b$. So, in particular, $h \leq h' \leftrightarrow ha \leq h'b$. But since $h \leq h'$, we get $ha \leq h'b$: a contradiction. QED

No similar result holds for Preference Revelation. An easy counter-example shows that, even when $\leq$ is total:

Fact 3.7 *Accommodation does not imply Preference Revelation.*

4 The Main Representation Theorem

We start with a warm-up case, taking plausibility to be a *total* pre-order.

4.1 Total pre-orders

Theorem 4.1 *Let $\mathcal{H}$ be a total doxastic-temporal model, $\mathcal{M}$ a total plausibility model, $\vec{\epsilon}$ a sequence of total event models, and let $\otimes$ stand for* priority *update. The following are equivalent:*

- *$\mathcal{H}$ is isomorphic to the forest generated by $\mathcal{M} \otimes \vec{\epsilon}$.*
- *$\mathcal{H}$ satisfies propositional stability, synchronicity, bisimulation invariance, Preference Propagation, and Preference Revelation.*

Proof.

Necessity We first show that the given conditions are indeed satisfied by any $DoTL$ model generated through successive priority updates along some given protocol sequence. Here, *Propositional stability* and *Synchronicity* are straightforward from the definition of generated forests.

Preference Propagation Assume that $ja \leq j'b$ (1). It follows from (1) plus the definition of priority update that $a \leq b$ (2). Now assume that $h \leq h'$ (3). It follows from (2), (3) and priority update that $ha \leq h'b$.

Preference Revelation Assume that $jb \leq j'a$ (1). It follows from (1) and the definition of priority update that $b \leq a$ (2). Now assume $ha \leq h'b$ (3). By the definition of priority update, (3) can happen in two ways. Case 1: $a < b$ (4). It follows from (4) by the definition of $<$ that $b \not\leq a$ (5). But (5) contradicts (2). We are therefore in Case 2: $a \simeq b$ (6) and $h \leq h'$ (7). But (7) is precisely what we wanted to show.

Note that we did not make use of totality here.

Sufficiency Given a $DoTL$ model $\mathcal{M}$, we first show how to construct a DDL model, i.e., a plausibility model and a sequence of event models.

Construction Here is the initial plausibility model $\mathcal{M} = \langle W, (\preceq_i)_{i \in N}, \hat{V}\rangle$:

- $W := \{h \in H \mid len(h) = 1\}$.
- Set $h \preceq_i h'$ iff $\leq_i$.
- For every $p \in Prop$, $\hat{V}(p) = V(p) \cap W$.

Now we construct the j-th event model $\epsilon_j = \langle E_j, (\preceq_i^j)_{i \in N}, \mathtt{pre}_j\rangle$:

- $E_j := \{e \in \Sigma \mid$ there is a history $he \in H$ with $len(h) = j\}$
- For each $i \in N$, set $a \preceq_i^j b$ iff there are $ha, h'b \in H$ such that $len(h) = len(h) = j$ and $ha \leq_i h'b$.
- For each $e \in E_j$, let $\mathtt{pre}_j(e)$ be the formula that characterizes the set $\{h \mid he \in H$ and $len(h) = j\}$. By general modal logic, *bisimulation invariance* guarantees that there is such a formula, though it may be an infinitary one in general.

Now we show that the construction is correct in the following sense:

Claim 4.2 (Correctness) *Let $\leq$ be the plausibility relation in the given doxastic temporal model. Let $\preccurlyeq^F_{DDL}$ be the plausibility relation in the forest induced by priority update over the just constructed plausibility model and matching sequence of event models. We have:*

$$h \leq h' \text{ iff } h \preccurlyeq^F_{DDL} h'.$$

Proof of the claim The proof is by induction on the length of histories. The base case is obvious from the construction of our initial model $\mathcal{M}$. Now for the induction step. As for notation we will write $a \leq b$ for $a \preceq_i^n b$ with n the length for which the claim has been proved, and i an agent.

From $DoTL$ to $Forest(DDL)$ Assume that $h_1 a \leq h_2 b$ (1). It follows that in the constructed event model $a \leq b$ (2). Case 1: $a < b$. By priority update we have $h_1 a \preccurlyeq^F_{DDL} h_2 b$. Case 2: $b \leq a$ (3). This means that there are $h_3 b, h_4 a$ such that $h_3 b \leq h_4 a$. But then by *Preference Revelation* and (1) we have $h_1 \leq h_2$ (in the doxastic temporal model). It follows by the inductive hypothesis that $h_1 \preccurlyeq^F_{DDL} h_2$. But then by priority update, since by (2) and (3) a and b are indifferent, we have $h_1 a \preccurlyeq^F_{DDL} h_2 b$.

From $Forest(DDL)$ to $DoTL$ Next let $h_1 a \preccurlyeq^F_{DDL} h_2 b$. The definition of priority update has two clauses. Case 1: $a < b$. By definition, this implies that $b \not\leq a$. But then by the above construction, for all histories $h_3, h_4 \in H$ we have $h_3 b \not\leq h_4 a$. In particular we have $h_2 b \not\leq h_1 a$. But then by *totality*[4], $h_1 a \leq h_2 b$. Case 2: $a \simeq b$ (4) and $h_1 \preccurlyeq^F_{DDL} h_2$. For a start, by the inductive hypothesis, $h_1 \leq h_2$ (5). By (4) and our construction, there are $h_3 a, h_4 b$ with $h_3 a \leq h_4 b$ (6). But then by *Preference Propagation*, (5) and (6) imply that we have $h_1 a \leq h_2 b$. QED

Next, we turn to the general case of pre-orders, allowing incomparability.

4.2 The general case

While the argument went smoothly for *total* pre-orders, it gets somewhat more interesting when incomparability enters the stage. In the case of pre-orders we need the additional axiom of Accommodation as stated below:

Theorem 4.3 *Let $\mathcal{H}$ be a doxastic-temporal model, $\mathcal{M}$ a plausibility model, $\vec{\epsilon}$ be a sequence of event models while $\otimes$ is* priority *update. The following assertions are equivalent:*

- *$\mathcal{H}$ is isomorphic to the forest generated by $\mathcal{M} \otimes \vec{\epsilon}$,*
- *$\mathcal{H}$ satisfies bisimulation invariance, propositional stability, synchronicity, Preference Revelation and Accommodation.*

By Fact 3.6, requiring Accommodation also gives us Preference Propagation.

Proof.
Necessity of the conditions The verification of the conditions in the preceding subsection did not use totality. So we concentrate on the new condition:

[4] Note that this is the only place in which we make use of totality.

Accommodation Assume that $ja \leq j'b$ (1). It follows by the definition of priority update that $a \leq b$ (2). Now let $ha \not\leq h'b$ (3). This implies by priority update that $a \not< b$ (4). By definition, (2) and (4) means that $a {\simeq} b$ (5). Now assume that $g \leq g'$ (6). It follows from (5), (6) and priority update that $ga \leq g'b$. For the other direction of the consequent assume instead that $g \not\leq g'$ (7). It follows from (5), (7) and priority update that $ga \not\leq g'b$.

Sufficiency of the conditions Given a $DoTL$ model, we again construct a DDL plausibility model plus sequence of event models:

Construction The plausibility model $\mathcal{M} = \langle W, (\preceq_i)_{i \in N}, \hat{V} \rangle$ is as follows:

- $W := \{h \in H \mid len(h) = 1\}$,
- Set $h \preceq_i h'$ whenever $\leq_i$,
- For every $p \in Prop$, $\hat{V}(p) = V(p) \cap W$.

We construct the j-th event model $\epsilon_j = \langle E_j, (\preceq_i^j)_{i \in N}, \mathtt{pre}_j \rangle$ as follows:

- $E_j := \{e \in \Sigma \mid$ there is a history of the form $he \in H$ with $len(h) = j\}$
- For each $i \in N$, define $a {\preceq_i^j} b$ iff either (a) there are $ha, h'b \in H$ such that $len(h) = len(h) = j$ and $ha \leq_i h'b$, or (b) [a new case] a and b are accommodating, and we put $a \simeq b$ (i.e. $a \leq b$ and $b \leq a$).
- For each $e \in E_j$, let $\mathtt{pre}_j(e)$ be the formula that characterizes the set $\{h \mid he \in H$ and $len(h) = j\}$. *Bisimulation invariance* guarantees that there is always such a formula (maybe involving an infinitary syntax).

Again we show that the construction is correct in the following sense:

Claim 4.4 (Correctness) *Let $\leq$ be the plausibility relation in the doxastic temporal model. Let $\preccurlyeq_{DDL}^F$ be the plausibility relation in the forest induced by successive priority updates of the plausibility model by the sequence of event models we constructed. We have:*

$$h \leq h' \text{ iff } h \preccurlyeq_{DDL}^F h'.$$

Proof of the claim We proceed by induction on the length of histories. The base case is clear from our construction of the initial model $\mathcal{M}$. Now for the induction step, with the same simplified notation as earlier.

From $DoTL$ to $Forest(DEL)$ There are two cases:

Case 1. $ha \leq h'b, h \leq h'$**.** By the inductive hypothesis, $h \leq h'$ implies $h \preccurlyeq_{DDL}^F h'$ (1). Since $ha \leq h'b$, it follows by construction that $a \leq b$ (2). It follows from (1) and (2) that by priority update $ha \preccurlyeq_{DDL}^F h'b$.

Case 2. $ha \leq h'b, h \not\leq h'$**.** Clearly, then, a and b are not *accommodating* and thus the special clause has not been used to build the event model, though we do have $a \leq b$ (1). By the contrapositive of Preference Revelation, we also conclude that for all $ja, j'b \in H$, we have $j'b \not\leq ja$ (2). Therefore, our construction gives $b \not\leq a$ (3), and we conclude that $a < b$ (4). But then by priority update, we get $ha \preccurlyeq_{DDL}^F h'b$.

From $Forest(DEL)$ to $DoTL$ We distinguish again two relevant cases.

Case 1. $ha \preccurlyeq_{DDL}^F h'b, h \preccurlyeq_{DDL}^F h'$**.** By definition of priority update, $ha \preccurlyeq_{DDL}^F h'b$ implies that $a \leq b$ (1). There are two possibilities. Case 1: The special clause of the construction has been used, and a, b are accommodating (2). By the inductive hypothesis, $h \preccurlyeq_{DDL}^F h'$ implies $h \leq h'$ (3). But (2) and (3) imply that $ha \leq h'b$. Case 2: Clause (1) holds because for some $ja, j'b \in H$, in the $DoTL$ model, $ja \leq j'b$ (4). By the inductive hypothesis, $h \preccurlyeq_{DDL}^F h'$ implies $h \leq h'$ (5). Now, it follows from (4), (5) and Preference Propagation that $ha \leq h'b$.

Case 2. $ha \preccurlyeq_{DDL}^F h'b, h \not\preccurlyeq_{DDL}^F h'$**.** Here is where we put our new accommodation clause to work. Let us label our assertions: $h \not\preccurlyeq_{DDL}^F h'$ (1) and $ha \preccurlyeq_{DDL}^F h'b$ (2). It follows from (1) and (2) by the definition of priority update that $a < b$ (3), and hence, by definition $b \not\leq a$ (4). Clearly, a and b are not accommodating (5): for otherwise, we would have had $a {\simeq} b$, and hence $b \leq a$, contradicting (4). Therefore, (3) implies that there are $ja, j'b \in H$ with $ja \leq j'b$ (6). Now assume for *contradictio* that (in the $DoTL$ model) $ha \not\leq h'b$ (7). It follows from (6) and (7) by Accommodation that a and b *are* accommodating, contradicting (5). Thus we have $ha \leq h'b$. QED

Given a doxastic temporal model describing the evolution of the beliefs of a group of agents, we have determined whether it could have been generated by successive 'local' priority updates of a plausibility model. Of course, further scenarios are possible, e.g., bringing in knowledge as well. We discuss some extensions in the next subsection.

4.3 Extensions and variations

4.3.1 Unified plausibility models

There are two roads to merging epistemic indistinguishability and doxastic plausibility. The first works with a plausibility order *and* an epistemic indistinguishability relation, explaining the notion of *belief* with a mixture of the two. Baltag and Smets [3] apply product update to epistemic indistinguishability and priority update to the plausibility relation. A characterization for the doxastic epistemic temporal models induced in this way follows from van Benthem et al. [9] Theorem 1.4 plus Theorem 4.3 of previous subsection (or its simpler counterpart for total orders). All this has the flavor of working with *prior* beliefs and information partitions, taking the *posteriors* to be computed from them.

However there are also reasons for working with (*posterior*) beliefs only (see e.g. [22]). Indeed, Baltag and Smets [3] take this second road, using *unified* 'local' plausibility models with just one explicit relation $\trianglelefteq$. We briefly show how our earlier results transform to this setting. In what follows, we write $a \cong b$ iff $a \trianglelefteq b$ and $b \trianglelefteq a$.

Definition 4.5 The priority update of a *unified* plausibility model $\mathcal{M} = \langle W, (\trianglelefteq_i)_{i \in N}, V\rangle$ and a $\trianglelefteq$-event model $\epsilon = \langle E, (\trianglelefteq_i)_{i\in N}, \mathtt{pre}\rangle$ is the unified plausibility model $\mathcal{M} \otimes \epsilon = \langle W', (\trianglelefteq'_i)_{i\in N}, V'\rangle$ constructed as follows:

- $W' = \{(w, e) \in W \times E \mid \mathcal{M}, w \Vdash \mathtt{pre}(e)\}$,
- $(w, a) \trianglelefteq'_i (w', b)$ iff either 1. $a \trianglelefteq_i b, b \not\trianglelefteq a$ and $w \trianglelefteq w' \vee w' \trianglelefteq w$ or 2. $a \trianglelefteq_i b, b \trianglelefteq a$ and $w \trianglelefteq w'$,
- $V'((s, e)) = V(s)$.

$\lhd$

Here are our familiar key properties in this setting:

Agent revision properties in terms of $\trianglelefteq_i$

- $\trianglelefteq$-Perfect Recall if, whenever $ha \trianglelefteq h'b$ we have $h \trianglelefteq h' \vee h' \trianglelefteq h$.
- $\trianglelefteq$-Preference Propagation if, whenever $h \trianglelefteq h'$ and $ja \trianglelefteq j'b$ then $ha \trianglelefteq h'b$.
- $\trianglelefteq$-Preference Revelation if, whenever $ha \trianglelefteq h'b$ and $jb \trianglelefteq j'a$, also $h \trianglelefteq h'$.
- $\trianglelefteq$-*Accommodation* if, whenever ($ja \trianglelefteq j'b, h' \trianglelefteq h$ and $ha \not\trianglelefteq h'b$), for all $ga, g'b \in H$ $(g \trianglelefteq g' \leftrightarrow ga \trianglelefteq g'b)$, and for all $g'a, gb \in H$ $(g \trianglelefteq g' \leftrightarrow gb \trianglelefteq g'a)$.

The last axiom is slightly weaker than Accommodation. The following result is proved in the extended version of this paper.

Theorem 4.6 *Let $\mathcal{H}$ be a unified doxastic-temporal model, $\mathcal{M}$ a unified plausibility model, $\vec{\epsilon}$ be a sequence of unified event models, while $\otimes$ is* priority *update. The following assertions are equivalent:*

- *$\mathcal{H}$ is isomorphic to the forest generated by $\mathcal{M} \otimes \vec{\epsilon}$,*
- *$\mathcal{H}$ satisfies bisimulation invariance, propositional stability, synchronicity, $\trianglelefteq$-Perfect Recall, $\trianglelefteq$-Preference Propagation, $\trianglelefteq$-Preference Revelation and $\trianglelefteq$-$Accommodation$.*

Our next source of variation is an issue that we have left open throughout our analysis so far, which may have bothered some readers.

4.3.2 Bisimulations and pre-condition languages

Our definition of event models presupposed a language for the preconditions, and correspondingly, the right notion of bisimulation in our representation results should matching (at least, on finite models) the precondition language used. For instance, if the precondition language contains a belief operator scanning the *intersection* of a plausibility $\leq_i$ relation and an epistemic indistinguishability relation $\sim$, then the *zig* and *zag* clauses should not only apply to $\leq_i$ and $\sim_i$ separately, but also to $\leq_i \cap \sim_i$. And things get even more complicated if we allow temporal operators in our languages (cf. [10]). We do not commit to any specific choice here, since the choice of a language seems orthogonal to our main concerns. But we will discuss formal languages in the next section, taking definability of our major structural constraints as a guide.

Finally, our results can be generalized by including one more major parameter in describing processes:

4.3.3 Protocols

So far we have assumed that the same sequences of events were executable uniformly anywhere in the initial doxastic model, provided the worlds fulfilled the preconditions. This strong assumption is lifted in [10, 9], who allow the *protocol*, i.e., the set of executable sequences of events forming our current informational process, to vary from state to state. Initially, they still take the protocol to be *common knowledge*, but eventually, they allow for scenarios where agents need not know which protocol is running. These variations change the complete dynamic-epistemic logic of the system. It would be of interest to extend this work to our extended doxastic setting.

5 Dynamic Languages and Temporal Doxastic Languages

Our emphasis so far has been on structural properties of models. To conclude, we turn to the logical languages that can express these, and hence also, the type of doxastic reasoning our agents can be involved with.

5.1 Dynamic doxastic language

We first look at a core language that matches dynamic belief update.

5.1.1 Syntax

Definition 5.1 [Dynamic Doxastic-Epistemic language] The language of dynamic doxastic language $DDE\mathcal{L}$ is defined as follows:

$$\phi := p \mid \neg\phi \mid \phi \vee \phi \mid \langle \leq_i \rangle \phi \mid \langle i \rangle \phi \mid \mathtt{E}\phi \mid \langle \epsilon, \mathbf{e} \rangle \phi$$

where i ranges over over N, p over a countable set of proposition letters $Prop$, and $(\epsilon, \mathbf{e})$ ranges over a suitable set of symbols for event models. ◁

All our dynamic doxastic logics will be interpreted on the following models.

5.1.2 Models

Definition 5.2 [Epistemic Plausibility Models] An *epistemic plausibility model* $\mathcal{M} = \langle W, (\preceq_i)_{i \in N}, (\sim_i)_{i \in N}, V \rangle$ has $W \neq \emptyset$, and for each $i \in N$, $\preceq_i$ is a pre-order on W, and $\sim_i$ any relation, while $V : Prop \to \wp H$. ◁

Definition 5.3 [$\sim$,$\preceq$-event model] An *epistemic plausibility event model* ($\sim$,$\preceq$-event model for short) ϵ is of the form $\langle E, (\preceq_i)_{i \in N}, (\sim_i)_{i \in N}, \mathtt{pre} \rangle$ where $E \neq \emptyset$, for each $i \in N$, $\preceq_i$ is a pre-order on E and $\sim_i$ is a relation on W. Also, there is a precondition function $\mathtt{pre} : E \to DDE\mathcal{L}$

◁

Definition 5.4 [Priority update] The *priority update* of an epistemic plausibility model $\mathcal{M} = \langle W, (\preceq_i)_{i \in N}, (\sim_i)_{i \in N}, V \rangle$ and a $\sim$,$\prec$-event model $\epsilon = \langle E, (\preceq_i)_{i \in N}, (\sim_i)_{i \in N}, \mathtt{pre} \rangle$ is the plausibility model $\mathcal{M} \otimes \epsilon = \langle W', (\preceq'_i)_{i \in N}, V' \rangle$ whose structure is defined as follows:

- $W' = \{(w, e) \in W \times E \mid \mathcal{M}, w \Vdash \mathtt{pre}(e)\}$
- $(w, e) \preceq'_i (w', e')$ iff $e {\prec_i} e'$, or $e {\simeq_i} e'$ and $w \preceq_i w'$
- $(w, e) \sim'_i (w', e')$ iff $e \sim_i e'$ and $w \sim_i w'$
- $V'((s, e)) = V(s)$.

The result of the update is an epistemic plausibility model. ◁

5.1.3 Semantics

Here is how we interpret the $DDE(L)$ language. A pointed event model is an event model plus an element of its domain. To economize on notation we use event symbols in the semantic clause. We write $\mathtt{pre}(e)$ for $\mathtt{pre}_\epsilon(e)$ when it is clear from context.

Definition 5.5 [Truth definition]
Let $K_i[w] = \{v \mid w \sim_i v\}$.

$$\begin{array}{lll}
\mathcal{M}, w \Vdash p & \text{iff} & w \in V(p) \\
\mathcal{M}, w \Vdash \neg\phi & \text{iff} & \mathcal{M}, w \nVdash \phi \\
\mathcal{M}, w \Vdash \phi \vee \psi & \text{iff} & \mathcal{M}, w \Vdash \phi \text{ or } \mathcal{M}, w \Vdash \psi \\
\mathcal{M}, w \Vdash \langle \leq_i \rangle \phi & \text{iff} & \exists v \text{ such that } w \preceq_i v \text{ and } \mathcal{M}, v \Vdash \phi \\
\mathcal{M}, w \Vdash \langle i \rangle \phi & \text{iff} & \exists v \text{ such that } v \in K_i[w] \text{ and } \mathcal{M}, v \Vdash \phi \\
\mathcal{M}, w \Vdash \mathtt{E}\phi & \text{iff} & \exists v \in W \text{ such that } \mathcal{M}, v \Vdash \phi \\
\mathcal{M}, w \Vdash \langle \epsilon, \mathbf{e} \rangle \phi & \text{iff} & \mathcal{M}, w \Vdash \mathtt{pre}(e) \text{ and } \mathcal{M} \times \epsilon, (w, e) \Vdash \phi
\end{array}$$

◁

The knowledge operator K_i and the universal modality $\mathtt{A}$ are defined as usual.

5.1.4 Reduction axioms

The methodology of dynamic epistemic and dynamic doxastic logics revolves around *reduction* axioms. On top of some complete static base logic, these fully describe the dynamic component. Here is well-known $Action - Knowledge$ reduction axiom of [2]:

$$[\epsilon, \mathbf{e}] K_i \phi \leftrightarrow (\mathtt{pre}(e) \rightarrow \bigwedge \{K_i[\epsilon, \mathbf{f}]\phi \ : \ e \sim_i f\}) \quad (1)$$

Similarly, here are the key reduction axioms for $\langle \epsilon, \mathbf{e} \rangle \langle \leq_i \rangle$ with priority update:

Proposition 5.6 *The following dynamic-doxastic principle is sound for plausibility change:*

$$\begin{array}{c} \langle \epsilon, \mathbf{e} \rangle \langle \leq_i \rangle \phi \leftrightarrow \\ (\mathtt{pre}(e) \wedge (\langle \leq_i \rangle \bigvee \{\langle \mathbf{f} \rangle \phi \ : \ e {\simeq_i} f\} \vee \\ \mathtt{E} \bigvee \{\langle \mathbf{g} \rangle \phi \ : \ e <_i g\})) \end{array} \quad (2)$$

The crucial feature of such a dynamic 'recursion step' is that the order between *action* and *belief* is reversed. This works because, conceptually, the current beliefs already *pre-encode* the beliefs after some specified event. In the epistemic setting, principles like this also reflect agent properties of Perfect Recall and No Miracles [11]. Here, they rather encode radically 'event-oriented' revision policies, and the same point applies to the principles we will find later in a doxastic temporal setting.

Finally for the existential modality $\langle \epsilon, \mathbf{e} \rangle \mathtt{E}$ we note the following fact:

Proposition 5.7 *The following axiom is valid for the existential modality:*

$$\langle \epsilon, \mathbf{e} \rangle \mathtt{E}\phi \leftrightarrow (\mathtt{pre}(e) \wedge (\mathtt{E} \bigvee \{\langle \mathbf{f} \rangle \phi \ : \ f \in Dom(\epsilon)\})) \quad (3)$$

We do not pursue further issues of axiomatic completeness here, since we are just after the model theory of our dynamic and temporal structures.

5.2 Doxastic epistemic temporal language

Next *epistemic-doxastic temporal models* are simply our old doxastic temporal models $\mathcal{H}$ extended with epistemic accessibility relations $\sim_i$.

5.2.1 Syntax

Definition 5.8 [Doxastic Epistemic Temporal Languages]
The language of $DET\mathcal{L}$ is defined by the following inductive syntax:

$$\phi := p \mid \neg\phi \mid \phi \vee \phi \mid \langle e \rangle \phi \mid \langle e^{-1} \rangle \phi \mid \langle \leq_i \rangle \phi \mid \langle i \rangle \phi \mid \mathtt{E}\phi$$

where i ranges over N, e over Σ, and p over proposition letters $Prop$. ◁

5.2.2 Semantics

The language $DET\mathcal{L}$ is interpreted over nodes h in our trees (cf. [11]):

Definition 5.9 [Truth definition]
Let $K_i[h] = \{h' \mid h \sim_i h'\}$.

$$\begin{array}{lll}
\mathcal{H}, h \Vdash p & \text{iff} & h \in V(p) \\
\mathcal{H}, h \Vdash \neg\phi & \text{iff} & \mathcal{H}, h \nVdash \phi \\
\mathcal{H}, h \Vdash \phi \vee \psi & \text{iff} & \mathcal{H}, h \Vdash \phi \text{ or } \mathcal{H}, h \Vdash \psi \\
\mathcal{H}, h \Vdash \langle e \rangle \phi & \text{iff} & \exists h\,' \in H \text{ s.t. } h' = he \text{ and } \mathcal{H}, h' \Vdash \phi \\
\mathcal{H}, h \Vdash \langle e^{-1} \rangle \phi & \text{iff} & \exists h\,' \in H \text{ s.t. } h'e = h \text{ and } \mathcal{H}, h' \Vdash \phi \\
\mathcal{H}, h \Vdash \langle \leq_i \rangle \phi & \text{iff} & \exists h\,' \text{ s.t. } h \leq_i h' \text{ and } \mathcal{H}, h' \Vdash \phi \\
\mathcal{H}, h \Vdash \langle i \rangle \phi & \text{iff} & \exists h\,' \text{ s.t. } h' \in K_i[h] \text{ and } \mathcal{H}, h' \Vdash \phi \\
\mathcal{H}, h \Vdash \mathtt{E}\phi & \text{iff} & \exists h\,' \in H \text{ s.t. } \mathcal{H}, h' \Vdash \phi
\end{array}$$

◁

Now we have the right syntax to analyze our earlier structural conditions.

5.3 Defining the frame conditions

We will prove semantic *correspondence results* (cf. [13]) for our crucial properties using somewhat technical axioms that simplify the argument. Afterwards, we present some reformulations whose meaning for belief-revising agents may be more intuitive to the reader:

5.3.1 The key correspondence result

Theorem 5.10 (Definability) *Preference Propagation, Preference Revelation and Accommodation are definable in the doxastic-epistemic temporal language $DET\mathcal{L}$.*

- *$\mathcal{H}$ satisfies Preference Propagation iff the following axiom is valid:*

$$\begin{array}{l}
\mathtt{E}\langle a \rangle \langle \leq_i \rangle \langle b^{-1} \rangle \top \ \rightarrow \\
(((\langle \leq_i \rangle \langle b \rangle p \ \wedge \ \langle a \rangle q) \\
\rightarrow \ \langle a \rangle (q \ \wedge \ \langle \leq_i \rangle p)
\end{array} \qquad (PP)$$

- *$\mathcal{H}$ satisfies Preference Revelation iff the following axiom is valid:*

$$\begin{array}{r}
\mathtt{E}\langle b \rangle \langle \leq_i \rangle \langle a^{-1} \rangle \top \ \rightarrow \\
((\langle a \rangle \langle \leq_i \rangle (p \ \wedge \ \langle b^{-1} \rangle \top) \ \rightarrow \ \langle \leq_i \rangle \langle b \rangle p)
\end{array} \qquad (PR)$$

- *$\mathcal{H}$ satisfies Accommodation iff the following axiom is valid:*

$$\begin{array}{r}
\mathtt{E}\langle a \rangle \langle \leq_i \rangle \langle b^{-1} \rangle \top \\
\wedge\, \mathtt{E}\, [\langle a \rangle \ (p_1 \ \wedge \ \mathtt{E}\, (p_2 \ \wedge \ \langle b^{-1} \rangle \top)\ ) \\
\wedge \ [a] \ (p_1 \ \rightarrow \ [\leq_i] \neg p_2)] \\
\rightarrow \ (\ ((\langle \leq_i \rangle \langle b \rangle q \ \rightarrow \ [a] \langle \leq_i \rangle q) \\
\wedge \ ((\langle a \rangle \langle \leq_i \rangle (r \ \wedge \ \langle b^{-1} \rangle \top) \ \rightarrow \ \langle \leq_i \rangle \langle b \rangle r)
\end{array} \qquad \text{(AC)}$$

Proof. We only prove the case of *Preference Propagation*, the other two are in the extended version of the paper. We drop agent labels for convenience.

(PP) **characterizes Preference Propagation** We first show that (PP) is valid on all models $\mathcal{H}$ based on preference-propagating frames. Assume that $\mathcal{H}, h \Vdash \mathtt{E}\langle a \rangle \langle \leq_i \rangle \langle b^{-1} \rangle \top$ (1). Then there are $ja, j'b \in H$ such that $ja \leq j'b$ (2). Now let $\mathcal{H}, h \Vdash ((\langle \leq \rangle \langle b \rangle p \ \wedge \ \langle a \rangle q)$ (3). Then there is $h' \in H$ such that $h \leq h'$ (4) and $\mathcal{H}, h' \Vdash \langle b \rangle p$ (5), while also $\mathcal{H}, ha \Vdash q$ (6). We must show that $\mathcal{H}, h \models \langle a \rangle (q \ \wedge \ \langle \leq_i \rangle p)$ (7). But, from (2),(4),(6) and *Preference Propagation*, we get $ha \leq h'b$, and the conclusion follows by the truth definition.

Next, we assume that axiom (PP) is valid on a doxastic temporal frame, that is, true under any interpretation of its proposition letters. So, assume that $ja \leq j'b$ (1), and also $h \leq h'$ (2). Moreover, let $ha, h'b \in H$ (3). First note that (1) automatically verifies the antecedent of (PP) in any node of the tree. Next, we make the antecedent of the second implication in (PP) true at h by interpreting the proposition letter p as just the singleton set of nodes $h'b$, and q as just ha (4). Since (PP) is valid, its consequent will also hold under this particular valuation V. Explicitly we have $\mathcal{H}, V, h \Vdash \langle a \rangle (q \wedge \langle \leq_i \rangle p)$. But spelling out what p, q mean there, we get just the desired conclusion that $ha \leq h'b$. QED

The preceding correspondence argument is really just a *Sahlqvist* substitution case (cf. [13]), and so are the other two. We do not prove a further completeness result, but will show one nice derivation, as a syntactic counterpart to our earlier Fact 3.5.

$$\texttt{E}\,[\langle a\rangle\,(\psi\ \wedge\ \texttt{E}\,(\phi\wedge\langle b^{-1}\rangle\top)\,)\ \wedge\ [a]\,(\psi\ \to [\leq_i]\neg\phi)] \\ \to\ ((\langle a\rangle\langle\leq_i\rangle(\phi\wedge\langle b^{-1}\rangle\top)\to\langle\leq_i\rangle\langle b\rangle\phi) \tag{F}$$

Here is an auxiliary correspondence observation:

Fact 5.11 *On total doxastic temporal models the following axiom is valid:*

$$\langle a\rangle(\psi\ \wedge\ \texttt{E}\,(\phi\wedge\langle b^{-1}\rangle\top))\ \to \\ (\,\langle a\rangle(\psi\ \wedge\ \langle\leq_i\rangle\phi)\ \vee\ \texttt{E}\langle b\rangle(\phi\wedge\ \langle\leq_i\rangle(\psi\wedge\langle a^{-1}\rangle\top)) \tag{Tot}$$

Now we can state an earlier semantic fact in terms of axiomatic derivability in some obvious minimal system for the language $DET\mathcal{L}$:

Fact 5.12

- $\vdash ((PP)\wedge(F))\to$ (AC)
- $\vdash ((PR)\wedge(Tot))\to(F)$

We leave the simple combinatorial details to the extended version of this paper. We now get an immediate counterpart to Fact 3.5:

Corollary 5.13

$$\vdash ((PP)\wedge(PR)\wedge(Tot))\to \text{(AC)} \tag{4}$$

5.3.2 Two intuitive explanations

Here are two ways to grasp the intuitive meaning of our technical axioms.

Reformulation with safe belief. An intermediate notion of knowledge first considered by [24] has been argued for doxastically as *safe belief* by [3] as describing those beliefs we do not give up under true new information. The safe belief modality $\Box^{\geq}$ is just the universal dual of the existential modality $\langle\geq\rangle$ scanning the converse of $\leq$. Without going into details of its logic (e.g., safe belief is positively, but not negatively introspective), here is how we can rephrase our earlier axiom:

- $\mathcal{H}$ satisfies Preference Propagation iff the following axiom is valid on $\mathcal{H}$:

$$\texttt{E}\langle a\rangle\langle\geq\rangle\langle b^{-1}\rangle\top\to((\langle a\rangle\Box^{\geq_i}p\to\Box^{\geq_i}[b]p) \tag{PP'}$$

A similar reformulation is easy to give for Preference Revelation. These principles reverse action modalities and safe belief much like the better-known Knowledge-Action interchange laws in the epistemic-temporal case. We invite the reader to check their intuitive meaning in terms of acquired safe beliefs as informative events happen.

Analogies with reduction axioms Another way to understand the above axioms in their original format with existential modalities is their clear analogy with the reduction axiom for priority update. Here are two cases juxtaposed:

$$\langle\epsilon,\mathbf{e}\rangle\langle\leq_i\rangle p\ \leftrightarrow \\ (\texttt{pre}(e)\wedge((\langle\leq_i\rangle\bigvee\{\langle f\rangle p\ :\ e{\simeq_i}f\} \\ \vee\ \texttt{E}\bigvee\{\langle g\rangle p\ :\ e<_i g\})) \tag{2}$$

$$\texttt{E}\langle a\rangle\langle\leq_i\rangle\langle b^{-1}\rangle\top\ \to \\ ((\langle\leq_{\texttt{i}}\rangle\langle\texttt{b}\rangle p\to[\texttt{a}]\langle\leq_{\texttt{i}}\rangle p) \tag{PP}$$

$$\texttt{E}\langle b\rangle\langle\leq_i\rangle\langle a^{-1}\rangle\top\ \to \\ ((\langle\texttt{a}\rangle\langle\leq_{\texttt{i}}\rangle(p\wedge\langle b^{-1}\rangle\top)\to\langle\leq_{\texttt{i}}\rangle\langle\texttt{b}\rangle p) \tag{PR}$$

Family resemblance is obvious, and indeed, (PP) and (PR) may be viewed as the two halves of the reduction axiom, transposed to the more general setting of arbitrary doxastic-temporal models.

5.4 Variations and extensions of the doxastic temporal language

5.4.1 Weaker languages

The above doxastic-temporal language is by no means the only reasonable one. Weaker forward-looking modal fragments also make sense, dropping both converse and the existential modality. But they do not suffice for the purpose of our correspondence.

Proposition 5.14 (Undefinability)
Preference Propagation, Preference Revelation and Accommodation are not *definable in the forward looking fragment of* **DET**$\mathcal{L}$.

Proof. The reason is the same in all cases: we show that these properties are not preserved under taking *bounded p-morphic images*. The Figure gives an indication how this works concretely. QED

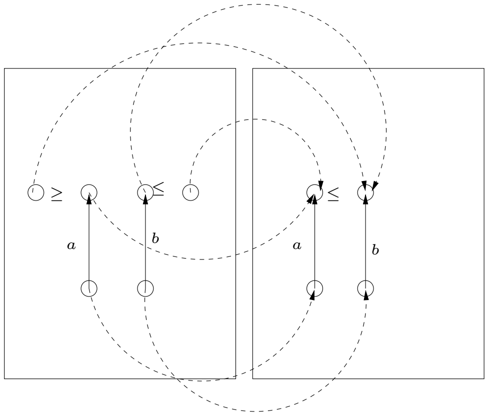


Figure 1. Preference Propagation is not preserved under p-morphic images

5.4.2 Richer languages

But there is also a case to be made for richer languages. For instance, if we want to define the frame property of *synchronicity*, we must introduce an *equilevel relation* in our models, with a corresponding modality for it. While expressing synchronicity then becomes easy, this move is dangerous in principle. Van Benthem and Pacuit [11] point at the generally high complexity of tree logics when enriched with this expressive power.

Likewise, finer epistemic and doxastic process descriptions require further temporal modalities, such as "Since" and "Until", beyond the basic operators we used for matching the needs of dynamic doxastic logic directly.

Finally, there may be even more urgent language extensions for doxastic temporal logic, having to do with our very notion of belief. We have emphasized the notion of *safe belief*, which scans the plausibility relation $\geq$ as an ordinary modality. This notion can be used to define the more standard notion of belief as truth in all most plausible worlds: cf. [15]. But it has been argued recently by [3], and also by [16] that we really want a more 'entangled' version of the latter notion as well, referring to the most plausible worlds *inside the epistemically accessible ones*. Such a notion of 'posterior belief' has the following semantics:

$$\mathcal{H}, h \Vdash B_i\phi \quad \text{iff} \quad \forall h\,' \in Min(K_i[h], \leq_i) \text{ we have } \mathcal{H}, h' \Vdash \phi$$

Technically, expressing this requires an additional intersection modality. While this extension loses some typical modal properties, it does satisfy reduction axioms in the format discussed here: cf. [21].

6 Conclusion

Agents that update their knowledge and revise their beliefs can behave very differently over time. We have determined the special constraints that capture agents operating with the 'local updates' of dynamic doxastic logic. This took the form of some representation theorems that state just when a general doxastic temporal model is equivalent to the forest model generated by successive priority updates of an initial doxastic model by a protocol sequence of event models. We have also shown how these conditions can be defined in an appropriate extended modal language, making it possible to reason formally about agents engaged in such updates and revisions. Our methods are like those of existing epistemic work, but the doxastic case came with some interesting new notions.

As for open problems, the paper has indicated several technical issues along the way, e.g., concerning the expressive power of different languages over our models and their complexity effects (cf. [11] for the epistemic case). In particular, we have completely omitted issues of common knowledge and common belief, even though these are known to generate complications [12].

But from where we are standing now, we see several larger directions to pursue:

- A systematic "protocol logic" of axiomatic completeness for constrained revision processes, analogous to the purely epistemic theory of observation and conversation protocols initiated in [9],
- A comparison of our 'constructive' DDL-inspired approach to DTL universes with the more abstract AGM-style postulational approach of [14],
- A theory of variation for different sorts of agents with different abilities and tendencies, as initiated in [21],
- An analysis of knowledge and belief dynamics in games [7, 17, 4]
- Connections with formal learning theory over epistemic-doxastic temporal universes (cf. [20]).

References

[1] A. Baltag and L. S. Moss. Logics for Epistemic Programs. *Synthese*, 139(2):165–224, 2004.

[2] A. Baltag, L. S. Moss, and S. Solecki. The logic of public announcements, common knowledge, and private suspicions. In *TARK '98: Proceedings of the 7th conference on Theoretical aspects of rationality and knowledge*, pages 43–56, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[3] A. Baltag and S. Smets. Dynamic belief revision over multi-agent plausibility models. In G. Bonanno, W. van der Hoek, and M. Wooldridge, editors, *Logic and the Foundations of Game and Decision Theory: Proceedings of LOFT'06*, Texts in Logic and Games, pages 11–24. Amsterdam University Press, 2006.

[4] A. Baltag, S. Smets, and J. Zvesper. Dynamic Rationality in Extensive Games. 2008. *(in this volume)*.

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

[6] J. v. Benthem. Games in Dynamic Epistemic Logic. *Bulletin of Economic Research*, 53(4):219–248, 2001.

[7] J. v. Benthem. Rational dynamics and epistemic logic in games. *International Game Theory Review*, 9(1):377–409, 2007. Erratum reprint, Volume 9(2), 2007, 377 - 409.

[8] J. v. Benthem. Priority product update as social choice. *ILLC Prepublications Series*, PP-2008-09, 2008.

[9] J. v. Benthem, J. Gerbrandy, T. Hoshi, and E. Pacuit. Merging frameworks for interaction: DEL and ETL. Working paper, 2008.

[10] J. v. Benthem, J. Gerbrandy, and E. Pacuit. Merging frameworks for interaction: DEL and ETL. In D. Samet, editor, *Proceedings of Theoretical Aspects of Rationality and Knowledge (TARK 2007)*, 2007.

[11] J. v. Benthem and E. Pacuit. The Tree of Knowledge in Action: Towards a Common Perspective. In I. H. G. Governatori and Y. Venema, editors, *Advances in Modal Logic*, volume 6. College Publications, 2006.

[12] J. v. Benthem, J. van Eijck, and B. P. Kooi. Logics of communication and change. *Information and Computation*, 204(11):1620–1662, 2006.

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

[14] G. Bonanno. Belief revision in a temporal framework. In G. Bonanno, W. van der Hoek, and M. Wooldridge, editors, *Logic and the Foundations of Game and Decision Theory: Proceedings of LOFT'06*, Texts in Logic and Games, pages 43–50. Amsterdam University Press, 2006.

[15] C. Boutilier. Conditional logics of normality: A modal approach. *Artificial Intelligence*, 68(1):87–154, 1994.

[16] D. de Jongh and F. Liu. Optimality, belief and preference. In S. Artemov and R. Parikh, editors, *Proceedings of the Workshop on Rationality and Knowledge*. ESSLLI, Malaga, 2006.

[17] C. Dégremont and J. Zvesper. Logique dynamique pour le raisonnement stratégique dans les jeux extensifs. In *Journées Francophones sur les Modèles Formels de l'Interaction (MFI'07)*, pages 61–73, 2007.

[18] K. Eliaz and E. A. Ok. Indifference or indecisiveness? Choice-theoretic foundations of incomplete preferences. *Games and Economic Behavior*, 56(1):61–86, July 2006.

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

[20] K. Kelly. Ockham's razor, truth, and information. In J. van Behthem and P. Adriaans, editors, *Handbook of the Philosophy of Information*. 2008.

[21] F. Liu. *Changing for the Better: Preference Dynamics and Agent Diversity*. Phd dissertation, ILLC Amsterdam, 2008.

[22] S. Morris. The common prior assumption in economic theory. *Economics and Philosophy*, 11:227–253, 1995.

[23] R. Parikh and R. Ramanujam. A knowledge based semantics of messages. *Journal of Logic, Language and Information*, 12(4):453–467, 2003.

[24] R. Stalnaker. A theory of conditionals. In R. Stalnaker, W. L. Harper, and G. Pearce, editors, *Ifs: Conditionals, Belief, Decision, Chance and Time*. D. Reidel, Dordrecht, 1981.

When all is done but not (yet) said: Dynamic rationality in extensive games

Alexandru Baltag
Oxford
baltag@comlab.ox.ac.uk

Sonja Smets
Brussels & Oxford
sonsmets@vub.ac.be

Jonathan A. Zvesper
Amsterdam
jonathan@illc.uva.nl

The jury is still out concerning the epistemic conditions for backward induction, the "oldest idea in game theory" ([2, p. 635]). Aumann [2] and Stalnaker [31] take contradictory positions in the debate: Aumann claims that common 'knowledge' of 'rationality' in a game of perfect information entails the backward-induction solution; Stalnaker that it does not.[1] Of course there is nothing wrong with any of their relevant formal proofs, but rather, as pointed out by Halpern [22], there are differences between their interpretations of the notions of knowledge, belief, strategy and rationality. Moreover, as pointed out by Binmore [14, 15], Bonanno [17], Bicchieri [13], Reny [26], Brandenburger [18] and others, the reasoning underlying the backward induction method seems to give rise to a *fundamental paradox* (the so-called "BI paradox"): in order to even start the reasoning, a player assumes that (common knowledge, or some form of common belief in) Rationality holds at all the last decision nodes (and so the obviously irrational leaves are eliminated); but then, in the next reasoning step (going backward along the tree), some of these (last) decision nodes are eliminated, as being incompatible with (common belief in) Rationality! Hence, the assumption behind the previous reasoning step is now undermined: the reasoning player can now see, that *if* those decision nodes that are now declared "irrational" were ever to be reached, then the only way that this could happen is if (common belief in) Rationality failed. Hence, she was *wrong* to assume (common belief in) Rationality when she was reasoning about the choices made at those last decision nodes. This whole line of arguing seems to undermine itself!

In this paper we use as a foundation the relatively standard and well-understood setting of Conditional Doxastic Logic (CDL, [16, 5, 7, 6]), and its "dynamic" version (obtained by adding to CDL operators for *truthful public announcements* $[!\varphi]\psi$): the logic PAL-CDL, introduced by Johan van Benthem [11]. In fact, we consider a slight extension of this last setting, namely the logic APAL-CDL, obtained by further adding dynamic operators for *arbitrary* announcements $[!]\psi$, as in [3]). We use this formalism to capture a novel notion of "dynamic rationality" and to investigate its role in decision problems and games. As usual in these discussions, we take a *deterministic stance*, assuming that the initial state of the world at the beginning of the game already fully determines the future play, and thus the unique outcome, irrespective of the players' (lack of) knowledge of future moves. We do *not*, however, require that the state of the world determines what *would* happen, if that state were *not* the actual state. That is, *we do not need to postulate the existence of any "objective counterfactuals"*. But instead, we only need "subjective counterfactuals": in the initial state, not only the future of the play is specified, but also the players' *beliefs* about each other, as well as their *conditional beliefs*, pre-encoding their *possible revisions of belief*. The players' conditional beliefs express what one may call their "propensities", or "dispositions", to revise their beliefs in particular ways, if given some particular pieces of new information.

Thus at the outset of a game, all is "done", including the future. But all is not necessarily said. In a deterministic model, as time progresses the only thing that changes are the pictures of the world in the minds of the players: the information states of the players. This is *"on-line" learning*: *while* the game is being played, the players *learn* the played moves, and so they may change their minds about the situation. We can *simulate* this on-line learning (and its effect on the players' beliefs) via *off-line "public announcements"*: if, *before* the start of the game, the agents were *publicly told* that the game will reach some node u, then they would be in the *same epistemic state* as they would have been by (not having any such public announcement but instead) playing the game until node u was reached.

So in this paper we stress the importance of the *dynamics* of beliefs and rationality during a play of an extensive game, and we use dynamic operators in order to simulate the play of the game. Since we focus on games of *perfect information*, we only need *public* announcements to simu-

[1] Others agree with Stalnaker in disagreeing with Aumann: for example, Samet [29] and Reny [26] also put forwards arguments against Aumann's epistemic characterisation of subgame-perfect equilibrium. Section 5 is devoted to a discussion of related literature.

late the moves of the game. The idea of adding modalities for public announcements to epistemic logic was introduced and developed in [24, 20]. Dynamic epistemic logic [4] provides for much richer dynamic modalities than just public announcements, capturing the effects of more complex and more "private" forms of learning. We think these could be applied to the case of games with *imperfect information*. However, for simplicity, we leave these developments for future work and consider for now only perfect information, and so only public announcements.

Using the terminology of Brandenburger [18], ours is a *belief-based approach* to game theory (in the same category as the work of Battigalli and Siniscalchi [9, 10]), in contrast to the *knowledge-based approach* of Aumann [2] and others. This means that we take the players' *beliefs* (including conditional beliefs) as basic, instead of their *knowledge*. However, there is a notion of knowledge that naturally arises in this context: the "irrevocable knowledge", consisting of the beliefs that are *absolutely unrevisable*, i.e. believed under any conditions. This notion of knowledge is meant to apply only to the players' "hard information", obtained by observation or by undoubtable evidence. This is a much stronger condition than "certain belief" (subjective probability 1) or even "true belief", and as a result it may happen that very few things are "known" in this sense. One of the things we assume to be irrevocably known is the *structure of the game*: the possible outcomes, the players' preferences etc; also, in a game of perfect information, the *played moves* are observed, and thus known, after they are played; finally, another thing irrevocably known to a player is *her own beliefs*: by introspection, she knows what she believes and what not. Besides this, we do not assume much else to be known, although our setting is definitely consistent with (common) knowledge of all the players' beliefs, their strategies, their rationality etc.

One thing we do *not* assume as known is the *future of the game*: no outcomes that are consistent with the structure of the game are to be excluded at the outset of the game. In fact, we make the opposite assumption: that it is common knowledge that nobody knows the future, i.e. nobody knows that some outcome will not be reached. This "open future" assumption seems to contradict common knowledge of rationality; but in fact, it is consistent with it, if by rationality we only mean "rational planning", leaving open the possibility that players may make mistakes or may change their minds. The players may certainly *believe* their rational plans will be faithfully carried out, but they have no way to "know" this in advance. We think of our "open future" assumption as being a realistic one, and moreover one that embodies the agents' "freedom of choice", as well as the "possibility of error", that underly a correct notion of rationality. An agent's rationality can be assessed only if she is given some options to freely choose from. There are certainly cases in which the future can be known, e.g. when it is determined by a known natural law. But it is an essential feature of rational agents that their own choices are not known to them to be thus determined; or else, they would have no real choices, and thus no rational choice. Any natural determinism is assumed to be absorbed in the definition of the game structure, which does pose absolute limits to choices. In a sense, this simply makes precise the meaning of our "knowledge" as "hard information", and makes a strict delimitation between the past and the future choices, delimitation necessary to avoid the various paradoxes and vicious circles that plague the notions of rational decision and freedom of choice: the agents may have "hard information" about the past and the present, but not about their own future free choices (although they may have "soft" information, i.e. "certain" beliefs, with probability 1, about their future choices).

Our notion of *"dynamic" rationality* takes into account the dynamics of beliefs, as well as the dynamics of knowledge. On the one hand, following Stalnaker, Reny, Battigalli and Siniscalchi etc. (and in contrast with Aumann), we assess the rationality of a player's move at a node against the beliefs held *at the moment when the node is reached*. On the other hand, we incorporate the above-mentioned epistemic limitation to rationality: the rationality of an agent's move only makes sense when that move is *not already known* (in an irrevocable manner) to her. Agents cannot be held responsible for moves that they cannot choose or change any more. Since the agents' knowledge increases during a game of perfect information, their set of available options decreases: passed options/nodes, or nodes that were bypassed, cannot be the objects of choice any more. As a result, our notion of rationality is *future-oriented*: it only concerns her plans concerning current and future decisions. An agent can be rational *now* even if in the past she has made some "irrational" moves. So in a sense, the meaning of "rationality" *changes in time*, synchronous to the change of beliefs and the change of (known) set of options. This concept of rationality, developed on purely a priori grounds, solves in one move the "BI-paradox": the first reasoning step in the backward-induction argument (dealing with the last decision nodes of the game) is *not undermined* by the result of the second reasoning step, since the notion of "Rationality" assumed in the first step is *not* the same as the "Rationality" disproved in the second step! The second step only shows that some counterfactual nodes cannot be reached by rational play, and thus it implies that some agent must have been irrational (or must have had some doubts about the others' rationality, or must have made some "mistake") *before* such an "irrational" node was reached; but this doesn't contradict in any way the assumption that the agents will be rational *at* that node (and further in the future).

Since dynamic rationality is only about *rational plan-*

ning, we need to strengthen it in order to capture *rational playing* of the game. We do this by adding to dynamic rationality a condition requiring that players actually play in accordance with their beliefs. The resulting condition is called "rational play".

Dynamics cannot really be understood without its correlative: *invariance* under change. Certain truths, or beliefs, *stay true* when everything else changes. We have already encountered an "absolute" form of invariance: "irrevocable knowledge", i.e. belief that is invariant under *any* possible information change. Now, we need a second, weaker form of invariance: "stability". A truth, or a belief, is *stable* if it remains true, or continues to be believed, after any *(joint) learning of "hard" information* (via some truthful public announcement). In fact, in the case of an "ontic" (non-doxastic) fact p, Stalnaker's favourite notion of "knowledge" of p [31, 33] (a modal formalisation of Lehrer and Klein's "defeasibility theory of knowledge"), also called "safe belief" in [7], corresponds precisely to *stable belief* in p. Stability can be or not a property of a belief or a common belief: a proposition P is a "stable (common) belief" if the fact that P is (common) belief is a stable truth, i.e. P continues to be (common) belief after any (joint) learning of "hard" information.

We can now give an informal statement of the main theorem of this paper:

> *Common knowledge of the game structure, of "open future" and of stable (common*[2]*) belief in rational play entails common belief in the backward induction outcome.*

Overview of the Paper To formalise stability and "stable common belief", we introduce in the next section Conditional Doxastic Logic CDL and its dynamic version APAL-CDL. Section 2 recalls the definition of extensive games and shows how to build models of those games in which the structure of the game is common knowledge, in our strong sense of "knowledge". In Section 3 we define "rationality" and "rational play", starting from more general decision-theoretic considerations, and arriving at a definition of *dynamic rationality* in extensive (aka "dynamic") games, which is in some sense a special case of the more general notion. Section 4 gives a formal statement of our main results. Section 5 discusses connections between our work and some existing literature on the epistemic foundations of backward induction.

1 Conditional Doxastic Logic

CDL models, also called "plausibility models" are essentially the "belief revision structures" in Board [16], simplified by incorporating structurally the assumption of Full Introspection of Beliefs (which allows us to use binary plausibility relations on worlds for each agent, instead of ternary relations). But since we will also want to talk about the *actual change* under the effects of actions, like moves in a game, rather than just the *static* notion that is in effect captured by Board's models, we will enrich the language of CDL with model-changing *dynamic* operators for "public announcements", in the spirit of Dynamic Epistemic Logic (cf. [4, 11, 12]).

The models are "possible worlds" models, where the worlds will usually be called *states*. Grove [21] showed that the AGM postulates [1] for rational belief change are equivalent to the existence of a suitable pre-order over the state space.[3] The intended interpretation of the pre-order $\leq_i$ of some agent i is the following: $s \leq_i t$ means that, in the event $\{s, t\}$, i considers *s at least as plausible as t.*

In interactive situations, where there are several players, each player i has a doxastic pre-order $\leq_i$. In addition to having different *beliefs*, any two players might have different *knowledge*. We follow the mainstream in game theory since Aumann and model interactive knowledge using a partitional structure. However, as in Board [16], we will derive i's partition from i's pre-order $\leq_i$. Let us be more precise: fix a set S and a relation $\leq_i \subseteq S \times S$; then we define the *comparability class* of $s \in S$ for $\leq_i$ to be the set $[s]_i = \{t \in S \mid s \leq_i t \text{ or } t \leq_i s\}$ of states $\leq_i$-comparable to s. Now we want the set of comparability classes to form a partition of S, so we will define a *plausibility frame* to be a sequence $(S, \leq_i)_{i \in N}$ in which S is a non-empty set of states, and each $\leq_i$ a pre-order on S such that for each $s \in S$, the restriction of $\leq_i$ to $[s]_i$ is a "complete" (i.e. "total" or "connected") preorder.

Fact 1.1 *In any plausibility frame, $\{[s]_i \mid s \in S\}$ forms a partition of S. We will interpret this as the* information partition *for player i (in the sense of "hard" information, to be explained below).*

So we can define player i's *knowledge* operator in the standard way, putting for any "proposition" $P \subseteq S$:

$$K_i P := \{s \in S \mid [s]_i \subseteq P\}$$

As explained below, this captures a notion of indefeasible, absolutely unrevisable knowledge. But we also want a notion of *belief B*, describing "soft" information, which might

[2] Adding the word "common" to this condition doesn't make a difference: common knowledge that everybody has a stable belief in P is the same as common knowledge of common safe belief in P.

[3] A pre-order is any reflexive transitive relation. In Grove's representation theorem the pre-order must also be total and converse-well-founded.

be subject to *revision*. So we want *conditional belief operators* B^P, in order to capture the *revised beliefs* given some new information P. If S is *finite*, let $min_{\leq_i}(P)$ denote the $\leq_i$-minimal P elements $\{s \in P \mid \forall t \in P, s \leq_i t\}$. So $min_{\leq_i}(P)$ denotes the set of states which i considers most plausible given P. Then $min_{\leq_i}(P \cap [s]_i)$ denotes the set of that states which i considers most plausible given both P and i's knowledge at state s. Thus we define player i's *conditional belief* operator as:

$$B_i^Q P := \{s \in S \mid min_{\leq_i}(Q \cap [s]_i) \subseteq P\}.$$

There is a standard way to extend this definition to total preorders on *infinite* sets of states, but we skip here the details, since we are mainly concerned with finite models. $B_i^Q P$ is the event that *agent i believes P conditional on Q*. Conditional belief should be read carefully: $B_i^Q P$ does *not* mean that after learning that Q, i will believe P; rather it means that after learning Q, i will believe that P *was the case before the learning*. This is a subtle but important point: the conditional belief operators do not directly capture the dynamics of belief, but rather as van Benthem [11] puts it, they 'pre-encode' it. We refer to [11, 7] for more discussion. The usual notion of *(non-conditional) belief* can be defined as a special case of this, by putting $B_i P := B_i^S P$. The notions of *common knowledge* CkP and *common belief* CbP are defined in the usual way: first, one introduces *general knowledge* $EkP := \bigcap_i K_i P$ and *general belief* $EbP := \bigcap_i B_i P$, then one can define $CkP := \bigcap_n (Ek)^n P$ and $CbP := \bigcap_n (Eb)^n P$.

It will be useful to associate with the states S some non-epistemic content; for this we use a *valuation function*. Assume given some finite set Φ of symbols, called *basic (or atomic) sentences*, and meant to describe ontic (non-epistemic, non-doxastic) "facts" about the (current state of the) world. A *valuation on* Φ is a function V that associates with each $p \in \Phi$ a set $V(p) \subseteq S$: V specifies at which states p is true. A *plausibility model* for (a given set of atomic sentences) Φ is a plausibility frame equipped with a valuation on Φ.

Interpretation: 'hard' and 'soft' information Information can come in different flavours. An essential distinction, due to van Benthem [11], is between 'hard' and 'soft' information. *Hard information* is absolutely "indefeasible", i.e. *unrevisable*. Once acquired, a piece of 'hard' information forms the basis of the strongest possible kind of knowledge, one which might be called *irrevocable knowledge* and is denoted about by K_i. For instance, the principle of Introspection of Beliefs states that (introspective) agents possess 'hard' information about their own beliefs: they know, in an absolute, irrevocable sense, what they believe and what not. *Soft information*, on the other hand, may in principle be defeated (even if it happens to be correct). An agent usually possesses only soft information about other agents' beliefs or states of mind: she may have beliefs about the others' states of mind, she may even be said to have a kind of 'knowledge' of them, but this 'knowledge' is defeasible: in principle, it could be revised, for instance if the agent were given more information, or if she receives misinformation.

For a more relevant, game-theoretic example, consider extensive games of perfect information: in this context, it is typically assumed (although usually only in an implicit manner) that, at any given moment, both the *structure of the game* and the players' *past moves* are 'hard' information; e.g. once a move is played, all players *know, in an absolute, irrevocable sense*, that it was played. Moreover, past moves (as well as the structure of the game) are common knowledge (in the same absolute sense of knowledge). In contrast, a player's 'knowledge' of other players' rationality, and even a player's 'knowledge' of her own future move at some node that is not yet reached, are not of the same degree of certainty: in principle, they might have to be revised; for instance, the player might make a mistake, and fail to play according to her plan; or the others might in fact play "irrationally", forcing her to revise her 'knowledge' of their rationality. So this kind of defeasible knowledge should better be called 'belief', and is based on players' "soft" information.[4]

In the 'static' setting of plausibility models given above, soft information is captured by the "belief" operator B_i. As already mentioned, this is defeasible, i.e. *revisable*, the revised beliefs after receiving some new information φ being pre-encoded in the conditional operator B_i^{φ}. Hard information is captured by the "knowledge" operator K_i; indeed, this is an absolutely unrevisable form of belief, one which can never be defeated, and whose negation can never be accepted as truthful information. This is witnessed by the following valid identities:

$$K_i P = \bigcap_{Q \subseteq S} B_i^Q P = B_i^{\neg P} \emptyset.$$

Special Case: Conditional Probabilistic Systems If, for each player i, we are given a *conditional probabilistic system* a la Renyi [27] over a common set of states S (or if alternatively we are given a *lexicographic probability system* in the sense of Blume et al), we can define subjective conditional probabilities $\text{Prob}_i(P|Q)$ for events of zero probability. When *S is finite and the system is discrete* (i.e., $\text{Prob}(P|Q)$ is defined for all non-empty events Q), we can use this to define conditional belief operators for arbitrary events, by putting $B_i^Q P := \{s \in S : \text{Prob}_i(P|Q) = 1\}$.

[4] By looking at the above probabilistic interpretation, one can see that the fact that an event or proposition has (subjective) probability 1 corresponds only to the agent having "soft" information (i.e. *believing* the event). "Hard" information corresponds to the proposition being true in *all* the states in the agent's information cell.

It is easy to see that these are special cases of finite plausibility frames, by putting: $s \leq_i t$ iff $\text{Prob}_i(\{s\}|\{s,t\}) \neq 0$. Moreover, the notion of conditional belief defined in terms of the plausibility relation is *the same* as the one defined probabilistically as above.

Dynamics and Information: 'hard' public announcements Dynamic epistemic logic is concerned with the "origins" of hard and soft information: the "epistemic actions" that can appropriately inform an agent. In this paper, we will focus on the simplest case of hard-information-producing actions: *public announcements*. These actions model the *simultaneous joint learning of some 'hard' piece of information* by a group of agents; this type of learning event is perfectly "transparent" to everybody: there is nothing hidden, private or doubtful about it. But dynamic epistemic logic [4] also deals with other, more complex, less transparent and more private, forms of learning and communication.

Given a plausibility model $\mathcal{M} = (S, \leq_i, V)_{i \in N}$ and a "proposition" $P \subseteq S$, the *updated model* $\mathcal{M} \upharpoonright P$ produced by a public announcement of P is given by *conditionalisation*: $(P, \leq_i \upharpoonright P, V \upharpoonright P)$, where $\leq \upharpoonright P$ is the restriction of $\leq$ to P and $(V \upharpoonright P)(p) = V(p) \cap P$. Notice that public announcements can change the knowledge and the beliefs of the players. So far we have, for readability, been writing events without explicitly writing the frame or model in question. However, since we are now talking about model-changing operations it is useful to be more precise; for this we will adopt a modal logical notation.

APAL-CDL: Language and Semantics Our language $APAL - CDL$ is built recursively, in the usual manner, from atomic sentences in Φ, using the Boolean connectives $\neg\varphi$, $\varphi \wedge \psi$, $\varphi \vee \psi$ and $\varphi \Rightarrow \psi$, the epistemic operators $K_i\varphi$, $B_i^{\varphi}\psi$, $Ck\varphi$ and $Cb\varphi$ and the dynamic modalities $[!\varphi]\psi$ and $[!]\varphi$. (The language CDL of conditional doxastic logic consists only of the formulas of $APAL - CDL$ that can be formed *without* using the dynamic modalities.)

For any formula φ of this language, we write $[\![\varphi]\!]_{\mathcal{M}}$ for the *interpretation* of φ, the event denoted by φ, in $\mathcal{M}$. We write $\mathcal{M}^{\varphi}$ for the updated model $\mathcal{M} \upharpoonright [\![\varphi]\!]_{\mathcal{M}}$ after the public announcement of φ. The interpretation map is defined recursively: $[\![p]\!]_{\mathcal{M}} = V(p)$; Boolean operators behave as expected; and the definitions given above of the epistemic operators in terms of events give the interpretation of epistemic formulae. Then the interpretation of the dynamic formulae, which include *public announcement modalities* $[!\varphi]\psi$, goes as follows:

$$[\![[!\varphi]\psi]\!]_{\mathcal{M}} = \{s \in S \mid s \in [\![\varphi]\!]_{\mathcal{M}} \Rightarrow s \in [\![\psi]\!]_{\mathcal{M}^{\varphi}}\}$$

Thus $[!\varphi]\psi$ means that after any true public announcement of φ, ψ holds. The *arbitrary (public) announcement modality* $[!]\varphi$ is to be read: after *every* (public) announcement, φ holds. Intuitively, this means φ is a *"stable" truth*: not only it is true, but it continues to stay true when any new (true) information is (jointly) learned (by all the players). There are some subtleties here: do we require that the new information/announcement be expressible in the language for example? This is the option taken in [3], where the possible announcements are restricted to *epistemic* formulas, and a *complete axiomatisation* is given for this logic. In the context of *finite models* (as the ones considered here), this definition is actually equivalent to allowing *all formulas of our language* $\mathcal{L}$ as announcements. As a result, we can safely use the following apparently circular definition:

$$[\![[!]\varphi]\!]_{\mathcal{M}} = \{s \in S \mid \forall\psi \in \mathcal{L}\ s \in [\![[!\psi]\varphi]\!]_{\mathcal{M}}\}$$

Dynamic epistemic logic captures the "true" dynamics of (higher-level) beliefs after some learning event: in the case of public announcements, the beliefs of an agent i after a joint simultaneous learning of a sentence φ are fully expressed by the operator $[!\varphi]B_i$, obtained by composing the dynamic and doxastic operators. Note that this is *not* the same as the conditional operator B_i^{φ}, but the two are related via the following "Reduction Law", introduced in [11]:

$$[!\varphi]B_i\psi \ \Leftrightarrow\ (\varphi \Rightarrow B_i^{\psi}[!\varphi]\psi).$$

This is the precise sense in which the conditional belief operators are said to "pre-encode" the dynamics of belief.

Special Case: Bayesian Conditioning In the case of a conditional probability structure, the update $\mathcal{M} \upharpoonright P$ by a public announcement $!P$ corresponds to *Bayesian update (conditionalisation)*: the state space is reduced to the event P, and the updated probabilities are given by $\text{Prob}_i'(Q) := \text{Prob}_i(Q|P)$. So a dynamic modality $[!P]Q$ corresponds to the event that, after conditionalising with P, event Q holds. Similarly, the arbitrary announcement modality $[!]P$ is the event that P *stably holds*, i.e. it holds after conditionalising with any true event.

2 Models and languages for games

The notion of *extensive game with perfect information* is defined as usual (cf. [23]): Let N be a set of 'players', and G be a finite tree of 'decision nodes', with terminal nodes (leaves) $\mathcal{O}$ (denoting "possible outcomes"), such that at each non-terminal node $v \in G - \mathcal{O}$, some player $i \in N$ is the decision-maker at v. We write $G_i \subseteq G$ for the set of nodes at which i is the decision-maker. Add to this a payoff function h_i for each player i, mapping all the leaves $o \in \mathcal{O}$ into real numbers, and you have an extensive game. We write 'G' to refer both to the game and to the corresponding set of nodes. We also write $u \to v$ to mean that

v is an immediate successor of u, and $u \rightsquigarrow v$ to mean that there is a path from u to v. A *subgame* of a game G is any game G', having a subset $G' \subseteq G$ as the set of nodes and having the immediate successor relation $\rightarrow'$, the set of decision nodes G'_i and the payoff function h'_i (for each player i) being given by restrictions to G' of the corresponding components of the game G (e.g. $G'_i = G_i \cap G'$ etc). For $v \in G$, we write G^v for the subgame of G in which v is the root. A *strategy* σ_i for player i in the game G is defined as usually as a function from G_i to G such that $v \rightarrow \sigma_i(v)$ holds for all $v \in G_i$. Similarly, the notions of *strategy profile*, of *the (unique) outcome determined by a strategy profile* and of *subgame-perfect equilibrium* are defined in the standard way. Finally, we define as usually a *backward induction outcome* to be any outcome $o \in \mathcal{O}$ determined by some subgame-perfect equilibrium. We denote by BI_G the set of all backward-induction outcomes of the game G.

Consider as an example the "centipede" game G (cf. [14]) given in Figure 1. This is a two-player game for a (Alice) and b (Bob).

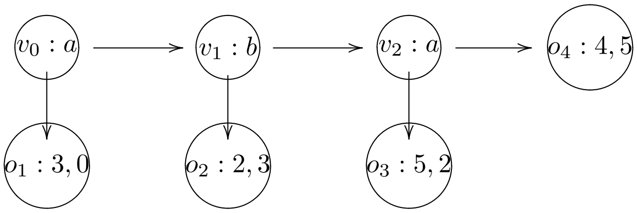


Figure 1. The "centipede" game G

Here, we represent the nodes of the game by circles and the possible moves by arrows. In each circle we write first the name of the node that the circle represents; then, if the node is non-terminal, we write the name of the player who decides the move at that node; while in the terminal nodes (outcomes) o_1, o_2, o_3, o_4, we write the payoffs as pairs (p_a, p_b), with p_a being Alice's payoff, and p_b Bob's. Note that in this game there is one backward induction outcome, o_1, and furthermore that the unique backward induction strategy profile assigns to each v_m the successor o_{m+1}.

Language for Games For any given game G, we define a set of *basic (atomic) sentences* Φ_G from which to build a language. First, we require Φ_G to contain a sentence for each leaf: for every $o \in \mathcal{O}$, there is a basic sentence $\overline{o}$. For simplicity, we often just write o, instead of $\overline{o}$. In addition Φ_G contains sentences to express the *players' preferences over leaves*: for each $i \in N$ and $\{o, o'\} \subseteq \mathcal{O}$, Φ_G has a basic sentence $o \prec_i o'$. Our formal language for games G is simply the language $APAL - CDL$ defined above, where the set of atomic sentences is the set Φ_G. To talk about the non-terminal nodes, we introduce the following abbreviation:

$$\overline{v} := \bigvee_{v \rightsquigarrow o} o \,,$$

for any $v \in G - \mathcal{O}$. As for terminal nodes, we will often denote this sentence by v for simplicity, instead of $\overline{v}$.

Plausibility Models for Games We now turn to defining game models. A *plausibility model for game G* is just a plausibility model $(S, \leq_i, V)_{i \in N}$ for the set Φ_G. We interpret every state $s \in S$ as an *initial state in a possible play* of the game. Intuitively, the sentence $\overline{o}$ is true at a state s if *outcome o will be reached during the play* that starts at s; and the sentence $o \prec_i o'$ says that *player i's payoff at o is strictly smaller than her payoff at o'*.

Observe that nothing in our definition of models for G guarantees that states come with a unique outcome or that the players know the set of outcomes! To ensure this (and other desirable constraints), we later focus on a special class of plausibility models for a game, called "game models".

Examples Figures 2 and 3 represent two different plausibility models $\mathcal{M}_1$ and $\mathcal{M}_2$ for the centipede game G. Here, we use *labelled arrows for the converse plausibility relations* $\geq_a$ (going from less plausible to more plausible states), but for convenience *we skip all the loops*.

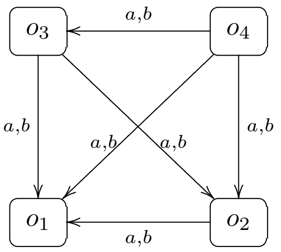


Figure 2. A game model $\mathcal{M}_1$ for the centipede game G

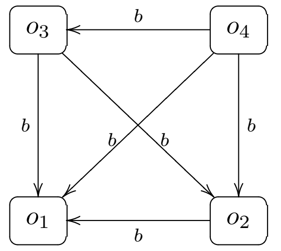


Figure 3. A plausibility model $\mathcal{M}_2$ for G, which is *not* a game model

Note that in the model M_2, Alice (player a) *knows the state of the world*: in each state, she knows both the outcome and Bob's beliefs (and belief revision policy), i.e. the sentence $\bigwedge_{o\in\mathcal{O}}(\mathbf{o} \Rightarrow K_a\mathbf{o})$ holds at all states of M_2. But this is *not true* in model M_1: on the contrary, in M_1 (it is common knowledge that) nobody knows the outcome of the game, and moreover nobody can exclude any outcome. Intuitively, the future is "epistemically open" in M_1, but not in M_2. However, we can also intuitively see that, in both models, (it is common knowledge that) all the players *know the (structure of the) game*: the available outcomes, the structure of the tree, the payoffs etc.

We now want to formalise our intuitions about open future and about having common knowledge of the structure of the game. To do this, we will focus on a special class of models, that we call "game models". Intuitively, each state of a game model comes with a complete play of the game, and hence it should have a uniquely determined outcome, and the set of possible outcomes as well as the players' preferences over them should be common knowledge. However, the players in this (initial) state should not have non-trivial knowledge about the outcome of the play. Indeed, they should have "freedom of choice" during the play, which means they can in principle play any move, so that at the outset of the play they cannot exclude a priori any outcomes.

Game Models The class of *game models for G*, denoted by $\mathfrak{M}_G$, as the class of all plausibility model for G satisfying the following conditions (for all players $i \in N$):

1. $\forall s \in S\, \exists! o \in \mathcal{O} : s \in V(o)$
2. $V(o \prec_i o') = \begin{cases} S & \text{if } h_i(o) < h_i(o') \\ \emptyset & \text{otherwise} \end{cases}$
3. $\forall s \in S\, \forall o \in \mathcal{O} : V(o) \cap [s]_i \neq \emptyset$

The first condition entails that there is common knowledge of the set of possible outcomes, as well as of the fact that each state is associated *a unique actual outcome*. This reflects the fact that the future, for each particular play (state), is determined. The second condition entails that the preferences over outcomes are commonly known. Finally, the third condition says that (it is common knowledge that) the *future is epistemically open*: in the initial state of any play, no player has "knowledge" (in the strong sense of "irrevocable", absolutely unrevisable knowledge) that any outcome is impossible. This is meant to apply even to the states that are incompatible with that player's plan of action.

Open Future We take condition (3) to embody the players' *freedom of choice*, as well as the *possibility of error*: in principle, players might always change their minds or make mistakes, hence any belief excluding some of the outcomes may have to be revised later. Even if we would assume (as usually is assumed) that players *(irrevocably) know their own strategy*, i.e. even if they are not allowed to change their minds, and even if we assume (as postulated by Aumann) that *they have common knowledge of "rationality"* (and so that they can exclude some obviously irrational choices), it still would not follow that they can completely exclude any outcome: mistakes can always happen, or players may always lose their rationality and become temporarily insane; so a rational plan does not necessarily imply a rational play, and hence the future still remains open.

Condition (3) is natural given our interpretation of the "knowledge" operator K as representing *hard information*, that is absolutely certain and irrevocable. If any node is "known" (in this sense) to be unreachable, then that node should simply be deleted from the game tree: this just corresponds to playing a *different game*. So if a player i would irrevocably know that a node is unreachable, then the structure of the game is not "really" common knowledge: i would in fact know that she is playing another game than G. Thus, one can consider the "open future" postulate as a natural strengthening of the "common knowledge of the game" assumption.

A different way to proceed would be to impose the above conditions only locally, at the "real" (initial) state of the play. Let Struct_G be the following sentence, describing the "structure of the game" G:

$$\bigvee_{o\in\mathcal{O}} o \wedge \bigwedge_{o\neq o'\in\mathcal{O}} \neg(o \wedge o') \wedge$$
$$\bigwedge_{\substack{i\in N, o, o'\in\mathcal{O} \\ \text{s.t. } h_i(o)<h_i(o')}} o \prec_i o' \wedge \bigwedge_{\substack{i\in N, o, o'\in\mathcal{O} \\ \text{s.t. } h_i(o)\geq h_i(o')}} \neg o \prec_i o'$$

Similarly, let $\text{F}_G := \bigwedge_{o\in\mathcal{O}, i\in N} \neg K_i \neg o$ be the sentence saying that at the outset of game G the future is epistemically open. Then our proposed "local" requirement is that in the initial state s we have "common knowledge of the structure of the game and of open future", i.e. s satisfies the sentence $Ck(\text{Struct}_G \wedge \text{F}_G)$. Then it is easy to see that this "local" requirement is equivalent to the above global requirement of having a "game model": for every state s in any plausibility model $\mathcal{M}$ for G, s satisfies $Ck(\text{Struct}_G \wedge \text{F}_G)$ iff it is bisimilar[5] to a state in some game model $\mathcal{M}' \in \mathfrak{M}_G$.

Examples Note that the model $\mathcal{M}_1$ from Figure 2 is a *game model*, while $\mathcal{M}_2$ from Figure 3 is *not*: indeed, in $\mathcal{M}_2$ it is common knowledge that Alice always *knows the outcome*, which contradicts the "Open future" assumption.

Encoding Strategies as Conditional Beliefs If a player adopts a particular (pure) *strategy*, our language can encode

[5] Here, "bisimilarity" is the standard notion used in modal logic, applied to plausibility models viewed as Kripke models with atomic sentences in Φ and with relations $\leq_i$. The important point is that our language $APAL - CDL$ cannot distinguish between bisimilar models and states.

this in terms of *the player's conditional beliefs about what she would do at each of her decision nodes*. For instance, we say that Alice "adopts the backward induction strategy" in a given state s of a model for the Centipede Game in Figure 1 iff the sentences $B_a o_1$ and $B_a^{v_2} o_3$ hold at state s. Similarly, we can express the fact that Bob adopts a particular strategy, and by putting these together we can capture *strategy profiles*. A given profile is realized in a model if the correspondent sentence is true at a state of that model.

Note that, in our setting, *nothing forces the players to adopt (pure) strategies*. Strategies are "complete" plans of action prescribing a unique choice (a belief that a particular move will be played) for each decision node of the player. But the players might simply consider all their options as equi-plausible, which essentially means that they do not have a strategy.

Examples In (any state of) model $\mathcal{M}_1$ from Figure 2 it is common knowledge that *both players adopt their backward induction strategies*. In contrast, in the model $\mathcal{M}_3$ from Figure 4, it is common knowledge that *no player has a strategy* (at any node):

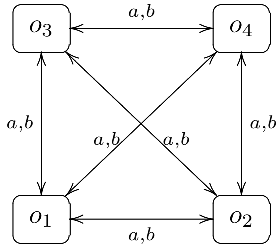


Figure 4. A game model $\mathcal{M}_3$ in which players don't have strategies

So the assumption that *players have (pure) "strategies"* is an extremely strong assumption, which *we will not need*. There is no a priori reason to assume (and there are good empirical reasons to reject) that players play according to fully-determined strategies. Our models are general enough to dispense with this assumption; indeed, our work shows that *this assumption is not needed for proving (common belief) that the backward induction strategy is played*.

Intentions as Beliefs In the above discussion, we identified an agent's *intentions* with *her beliefs about what she is going to do*, and so we represented the decision maker's plan of action as a belief about her (future) action. This identification is philosophically debatable, since agents may be aware of the possibility of mistakes, and so they may doubt that their intentions will be realized. But one can also argue that, in the context of Game Theory, such distinctions will be of very limited significance: indeed, an intention that is not believed to be enforceable is irrelevant for strategic planning (though see [28] for a discussion of intentions in game theory). The players only need to know each others' beliefs about their future actions and about each others' beliefs etc., in order to make their own rational plans; whether or not they are being informed about each others' (completely unenforceable and not believed to be enforceable) "intentions" will not make any difference. So, for the purposes of this paper, we can safely adopt the simplifying assumption that the *agents believe that they will be able to carry out their plans*. Given this assumption, an agent's "intentions" can be captured by her beliefs about her (future) actions.

Representing Players' Evolving Beliefs Recall that we think of every state of a game model $\mathcal{M}_G \in \mathfrak{M}_G$ as an *initial state (of a possible play)* of the game G. As the play goes on, the players' hard and soft information, their knowledge and beliefs, *evolve*. To represent this evolution, we will need to successively change our model, so that e.g. when a node v is reached, we want to obtain a corresponding model of the subgame G^v. That is precisely, in this perfect information setting, what is achieved by *updating the model with public announcements*: indeed, in a game of perfect information, every move, say from a node u to one of its immediate successors u', can be "simulated" by a public announcement $!u'$. In this way, for each subgame G^v of the original model $\mathcal{M}$, we obtain a model $\mathcal{M}^v$, that *correctly describes the players' knowledge and beliefs at the moment when node v is reached during a play*. This is indeed a model of the corresponding subgame G^v:

Proposition 2.1 *If $\mathcal{M} \in \mathfrak{M}_G$ then $\mathcal{M}^v \in \mathfrak{M}_{G^v}$.*

Example Consider a play of the Centipede game G that starts in the initial situation described by the model $\mathcal{M}_1$ in Figure 2, and in which the real state of the world is the one having outcome o_2: so Alice first plays "right", reaching node v_1, and the Bob plays "down", reaching the outcome o_2. The model $\mathcal{M}_1$ from Figure 2 gives us the initial situation, the model $\mathcal{M}_1^{v_1}$ in Figure 5 describes the epistemic situation after the first move, and then the model $\mathcal{M}_1^{o_2}$ in Figure 6 gives the epistemic situation at the end of the play:

In this way, for each given initial state s (of a given play $v_0, v_1, \ldots, o$ of the game, where o is the unique outcome such that $s \in V(o)$), we obtain a *sequence of evolving game models*

$$\mathcal{M} = \mathcal{M}^{v_0}, \mathcal{M}^{v_1}, \ldots, \mathcal{M}^{o} ,$$

describing *the evolving knowledge and beliefs of the players* during any play. Each model $\mathcal{M}^v$ accurately captures the players' beliefs at the moment when node v is reached. Note also that every such sequence ends with a model $\mathcal{M}^o$ consisting of only one node (a leaf o); this reflects the fact

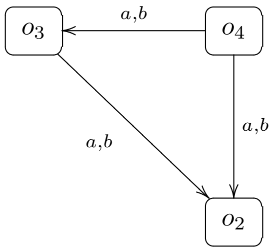


Figure 5. The model $\mathcal{M}_1^{v_1}$

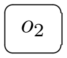


Figure 6. The model $\mathcal{M}_1^{o_2}$

that *at the end of the game, there is no uncertainty left*: the outcome, as well as the whole history of the game, are now common knowledge.

Simulating Moves by Public Announcements Using the dynamic "public announcement" modalities in constructs such as $[!v]B_i$, we can talk, *at the initial state* $s \in \mathcal{M}$ and *without leaving the original model* $\mathcal{M} \in \mathfrak{M}_G$, about all these future, evolving beliefs of the players at nodes v other than the initial node v_0. Indeed, in a game of perfect information, all the moves are *public*. So the epistemic effect of a move to node v is the same as that of a truthful public announcement $!v$ (saying that the node v is reached during the play). In other words, *we can "simulate" moves in games of perfect information by truthful public announcements*.[6]

3 Rationality in Decisions and Games

We now define our fundamental notions of *dynamic rationality* and *rational play*. First we will look at *single-agent (one-step) decision situations*, and then at interactive decision situations, i.e. *games*.

3.1 Single Agent Decision Problems

Given a *one-step decision problem* $\mathcal{P}$ with a set of *outcomes* $\mathcal{O}$, the *decision-maker* i selects one of the outcomes $o \in \mathcal{O}$. The decision-maker may have various hard and soft information about which outcomes can actually be realized and which not. This will determine her knowledge and her beliefs. We assume that *her "hard" knowledge restricts her possible choices*: she *can* only select outcomes that she doesn't *know* to be *impossible*.

What this amounts to is the following: for the decision maker i, the "true" set of possible outcomes is $\{o \in \mathcal{O} \mid \neg K_i \neg o\}$, i.e. the set of all the *"epistemically possible" outcomes*. So her selected option must satisfy: $o \in \{o \in \mathcal{O} \mid \neg K_i \neg o\}$. This allows us to capture the "selection" problem using epistemic operators.

To assess whether the decision is "rational" or not, one considers the decision-maker's subjective preferences, modelled as a total preorder $\preccurlyeq_i$ on $\mathcal{O}$. We assume that *agents know their preferences*; indeed, these are interpreted as "doxastic" preferences: beliefs about what's best. Given this interpretation, the CDL postulation of Full Introspection (of beliefs) implies that agents know their preferences.

Rational Choice Rationality, in this case, corresponds to requiring that the selected option is *not worse than* any other (epistemically) possible alternative. In other words, i's solution of the decision problem $\mathcal{P}$ is *rational* if she does not choose any option that is strictly less preferable than an option she doesn't know to be impossible:

$$\mathrm{R}_i^{\mathcal{P}} := \bigwedge_{o,o' \in \mathcal{O}} (o \prec_i o' \wedge \neg K_i \neg o' \Rightarrow \neg o).$$

The main difference between our definition and the standard definition of rational decision-making is the *epistemic limitation of the choice set*. The epistemic operators are used here to delimit *what is currently known about the availability of options*: i's choice should only be compared against options that are not known to be unavailable. This is an *important difference*, and its importance will become clear when we generalise our definition to extensive games.

3.2 Extensive Games

We now aim to extend the above definitions to the case of *multi-agent many-stage decisions*, i.e. "extensive games" (of perfect information). Recall that in an extensive game we are given the players' subjective preferences $\preccurlyeq_i$ only over the leaves. However, at all the intermediate stages of the game, players have to make local choices, not between "final" outcomes, but between "intermediary" outcomes, that is: between other nodes of the game tree.

So, in order to assess players' rationality, we need to *extend* the subjective preference relations to *all the nodes* of the game tree. Fortunately, given the above doxastic interpretation of preferences, there is an obvious (and natural) way to define these extensions. Namely, a player considers a node u to be *strictly less preferable* to a node u' if she *believes* the first to be *strictly dominated* by the second. More precisely, if every outcome that she believes to be achievable given that u is reached is worse than every outcome that she believes to be achievable given that u' is reached:

$$u \prec_i u' := \bigwedge_{o,o' \in \mathcal{O}} (\neg B_i^{u} \neg o \wedge \neg B_i^{u'} \neg o' \Rightarrow o \prec_i o').$$

[6] We believe that the more general case, of games of imperfect information, can also be handled by using other kinds of epistemic actions proposed in Dynamic Epistemic Logic [4]. But we leave this development for future work.

By the Full Introspection of beliefs (a postulate of the logic CDL), it follows that we still have that *players know their extended preferences* over all the nodes of the game.

Rationality at a Node Each node $v \in G_i$ can be considered as a (distinct) decision problem, in which the decision-maker is i, the set of outcomes is the set $\{u \in G : v \to u\}$ of all immediate successors of v, and the subjective preference relation is given by the (restriction of the) extended relation $\prec_i$ defined above (to the set $\{u \in G : v \to u\}$). So we can define the rationality of a player i at a node $v \in G_i$ as rationality for the corresponding decision problem, i.e. the player's selection at each decision node consists only of "best answers". Note that, as before, *the player's choice is epistemically limited*: if she has "hard knowledge" excluding some successors (for instance, because those nodes have already been bypassed), then those successors are excluded from the set of possible options. The only difference is that the "knowledge" involved is the one the agent *would have at that decision node*, i.e. it is *conditional on that node being reached*. Formally, we obtain:

$$\mathrm{R}_i^v := \bigwedge_{u,u' \leftarrow v} (u \prec_i u' \wedge \neg K_i^v \neg u' \Rightarrow \neg u)$$

where $K_i^\varphi \psi := K_i(\varphi \Rightarrow \psi)$.

Dynamic Rationality Let R_i be the sentence

$$\mathrm{R}_i = \bigwedge_{v \in G_i} \mathrm{R}_i^v.$$

If R_i is true, we say that player i satisfies *dynamic rationality*. By unfolding the definition, we see it is equivalent to:

$$\mathrm{R}_i = \bigwedge_{v \in G_i} \bigwedge_{u,u' \leftarrow_i v} (u \prec_i u' \wedge \neg K_i^v \neg u' \Rightarrow \neg u).$$

As we'll see, asserting this sentence at a given moment is a way of saying that *the player will play rationally from that moment onwards*, i.e. she will make the best move at any *current or future* decision node.

In the following, "Dynamic Rationality" denotes the sentence

$$R := \bigwedge_i R_i$$

saying that all players are dynamically rational.

Comparison with Substantive Rationality To compare our notion with Aumann's concept of "substantive rationality", we have to first adapt Aumann's definition to a belief-revision context. This has already been done by a number of authors e.g. Battigalli and Siniscalchi [9, 10], resulting in a definition of "rationality at a node" that differs from ours only by *the absence of epistemic qualifications to the set of available options* (i.e. the absence of the term $\neg K_i^v \neg u'$). The notion of *substantive rationality* is then obtained from this in the same way as dynamic rationality, by quantifying over all nodes, and it is thus equivalent to the following definition:

$$\mathrm{SR}_i = \bigwedge_{v \in G_i} \bigwedge_{u,u' \leftarrow_i v} (u \prec_i u' \Rightarrow \neg u).$$

It is obvious that *substantive rationality implies dynamic rationality*

$$\mathrm{SR}_I \Rightarrow \mathrm{R}_i,$$

but *the converse is in general false*. To better see the difference between SR_i and R_i, recall that a formula being true in a model $\mathcal{M} \in \mathfrak{M}_G$ means that it is true at the first node (the root) of the game tree G. However, we will later have to evaluate the formulas R_i and SR_i at *other* nodes w, i.e. in *other models* of the form M^w (models for *subgames* G^w). Since *the players' knowledge and beliefs evolve during the game*, what is (not) known/believed conditional on v *in model* M^w differs from was (not) known/believed conditional on v *in the original model* (i.e. at the outset of the game). In other words, the meaning of both dynamic rationality R_i and substantive rationality SR_i *will change during a play*. But they *change in different ways*. At the initial node v_0, the two notions are equivalent. But, once a node v has been bypassed, or once the move at v has already been played by a player i, that player is counted as *rational at node* v according to our definition, while according to the usual (non-epistemically qualified) definition the player may have been irrational at v.

In other words, the epistemic limitations we imposed on our concept of dynamic rationality make it into a *future-oriented* concept. At any given moment, the rationality of a player depends only on her current beliefs and knowledge, and so only on the options that she *currently considers possible*: past, or by-passed, options are irrelevant. Dynamic Rationality simply expresses the fact that the player's decision in any future contingencies is rational (given her future options and beliefs). Unlike substantive rationality, our concept has nothing to do with the past or with contingencies that are known to be impossible: a player i may still be "rational" in our sense at a given moment/node v even when v could only have been reached if i has already made some "irrational" move. The (knowledge of some) past mistake(s) may of course affect the others' *beliefs* about this player's rationality; but it doesn't directly affect her rationality, and in particular it doesn't automatically render her irrational.

Solving the BI Paradox As explained above, our concept is very different from (and, arguably, more realistic than) Aumann's and Stalnaker's substantive rationality, but also from other similar concepts in the literature (for example Rabinowicz's [25] "habitual" or "resilient" rationality, etc). The difference becomes more apparent if we consider the assumption that *"rationality" is common belief, in the strongest possible sense,* including *common "strong" belief* (in the sense of Battigalli and Siniscalchi [10]), common persistent belief, or even common "knowledge" in the sense of Aumann. As correctly argued by Stalnaker and Reny, these assumptions, if applied to the usual notions of rationality in the literature, bear no relevance for what the players would do (or believe) at the nodes that are incompatible with these assumptions! The reason is that, if these counterfactual nodes were to be reached, then by that time the belief in "rationality" would have already been *publicly disproved*: we cannot even entertain the possibilities reachable by irrational moves except by suspending our belief (or "knowledge") in rationality. Hence, the above assumptions cannot tell us anything about the players' behaviour or rationality at such counterfactual nodes, and thus they cannot be used to argue for the plausibility of the backward induction solution (even if they logically imply it)! In contrast, our notion of dynamic rationality is *not* automatically disproved when we reach a node excluded by common belief in it: a player *may* still be rational with respect to her current and future options and decisions *even after* making an "irrational" move. Indeed, the player may have been playing irrationally in the past, or may have had a moment of temporary irrationality, or may have made some mistakes in carrying out her rational plan; but she may have recovered now and may play rationally thereafter. Since our notion of rationality is future-oriented, no information about past moves will necessarily and automatically shatter belief in rationality (although of course it *may* still shatter it, or at least weaken it). So it is perfectly consistent (although maybe not always realistic) to assume that players maintain their common belief in dynamic rationality despite all past failures of rationality. In fact, *this is our proposed solution to the BI paradox*: we will show that such a "stable" common belief in dynamic rationality (or more precisely, common *knowledge* of the stability of the players' common belief in rationality) is exactly what is needed to ensure common belief in the backward induction outcome!

Rational Planning A weaker condition requires only that, for each decision node v, the option that the decision-maker *is planning at v to select (at v)* is the best, given the other (epistemically) possible alternatives. By identifying as above the players' plans of actions with their beliefs about their actions, we can thus say that a decision maker is a *rational planner* in the game G if at each decision node she *believes* that she will take "the best decision", even if in the end she may accidentally make a wrong choice:

$$\mathrm{RP}_i := \bigwedge_{v\in G_i} B_i^v \mathrm{R}_i^v .$$

By unfolding the definition, we see it is equivalent to:

$$\mathrm{RP}_i = \bigwedge_{v\in G_i} \bigwedge_{u,u'\leftarrow_i v} (u \prec_i u' \wedge \neg K_i^v \neg u' \Rightarrow B_i^v \neg u).$$

No Mistakes As noted above, $RP_i^{\mathcal{P}}$ only states that the decision maker i has a *rational plan* for current and future contingencies. But mistakes can happen, so if we want to ensure that the decision that is actually taken is rational we need to require the player makes *no mistakes* in carrying out her plan:

$$\text{No-Mistakes}_i := \bigwedge_{v\in G_i} \bigwedge_{u\leftarrow v} (B_i^v \neg u \Rightarrow \neg u)$$

The sentence No-Mistakes$_i$ says that player i's decision are always consistent with her "plan": she never plays a move that, at the moment of playing, she believed won't be played.

As expected, the conjunction of "rational planning" and "no mistakes" entails "rational playing":

$$\mathrm{RP}_i \wedge \text{No-Mistakes}_i \Rightarrow \mathrm{R}_i .$$

4 Backward Induction in Games of Perfect Information

It is easy to see that Aumann's theorem can be strengthened to the following

Proposition 4.1 *In any state of any plausibility model for a game of perfect information, common knowledge of dynamic rationality implies the backward induction outcome.*

Unfortunately, common knowledge of (either dynamic or substantive) rationality *can never hold in a game model*: it is simply incompatible with the "Epistemically-Open Future" condition. By requiring that players have "hard" information about the outcome of the game, Aumann's assumption *does not allow them to reason hypothetically or counterfactually about other possible outcomes*, at least not in a consistent manner.[7] This undermines the intuitive rationale behind the backward induction solution, and it is thus open to Stalnaker's criticism.

So in this section, we are looking for natural conditions that can be satisfied on game models, but that still imply the backward induction outcome (or at least common belief in it). One such condition is *common knowledge of (general) stable belief* in (dynamic) rationality: $Ck[!]EbR$. This is in fact a "strong" form of common belief, being equivalent to $Ck[!]CbR$, i.e. to *common knowledge of stable common belief* in rationality.

[7] Indeed, if o is the backward induction outcome, then the above Proposition entails $K_i o$ for all players i, and thus for every other outcome $o' \neq o$ and every proposition P, we have $B_i^{o'} P$: the players *believe everything* (including inconsistencies) conditional on o'.

Theorem 4.2 *The following holds in any state s of any game model $\mathcal{M} \in \mathfrak{M}_G$:*

$$Ck[!]Eb\mathrm{R} \;\Rightarrow\; Cb(\mathit{BI})\,,$$

where BI $:= \bigvee\{o \mid o \in BI_G\}$ *is the sentence saying that the current state determines a backward-induction outcome. Equivalently, the following formula is valid over plausibility frames for the game G:*

$$Ck(\mathit{Struct}_G \wedge F_G \wedge [!]Cb\mathrm{R}) \;\Rightarrow\; Cb(\mathit{BI}).$$

In English: assuming common knowledge of the game structure and of open future, if it is common knowledge that, no matter what new (truthful) information the players may (jointly) learn during the game (i.e. *no matter what is played*), general belief in rationality will be maintained, then it is common belief that the backward induction outcome will be reached. If we define "stable common belief" in a proposition P as $[!]CbP$, then we can give a more concise English formulation of the above theorem: *common knowledge of the game structure, of open future and of stable common belief in dynamic rationality implies common belief in the backward-induction outcome.*

Although rationality cannot be common knowledge in a game model, *rational planning can be*. When this is the case, we obtain the following

Corollary 4.3 *In a game model, common knowledge of "rational planning" and of stable belief in "no mistakes" implies the backward-induction outcome; i.e. the formula*

$$Ck(\mathrm{RP} \wedge [!]Eb\text{No-Mistakes}) \;\Rightarrow\; Cb(\mathit{BI})$$

is valid on game models.

The above results only give us *common belief* in the backward-induction outcome, but nothing ensures that this belief is correct. If we want to ensure that the backward-induction outcome is actually played, we need to add the requirement that the (stable common) belief in rational play assumed in the premise is correct, i.e. that *players actually play rationally*:

Theorem 4.4 *The following holds in any state s of any game model $\mathcal{M} \in \mathfrak{M}_G$:*

$$\mathrm{R} \wedge Ck[!]Eb\mathrm{R} \;\Rightarrow\; \mathit{BI}$$

No strategies! Observe that *we did not assume that the players have complete (pure) "strategies"* (fully determined plans of action, uniquely specifying one move for at each decision node), but only that they have *partial* plans, i.e. (incomplete) beliefs about what moves should they play: at each decision node they choose a *set* of moves rather than one unique move. So an important side-result of our work is that the assumption that players have (complete, pure) strategies is *not necessary* for proving backward-induction results.

Ensuring Backward-Induction Strategy Profile If, however, we want to postulate that *every player does have a (complete, pure) strategy*, we need to say that, for each node v of her choice, there exists a (unique) immediate successor u that she *believes* will be played if v is reached (i.e. she *plans to play* u at v):

$$\text{Strategies} := \bigwedge_i \bigwedge_{v \in G_i} \bigvee_{u \leftarrow_i v} B_i^v u.$$

In cases where Str is common knowledge as well, we can strengthen the Theorem 4.2 to:

Corollary 4.5 *The following holds in any state s of any game model $\mathcal{M} \in \mathfrak{M}_G$:*

$$Ck(\mathit{Strategies} \wedge [!]Eb\mathrm{R}) \;\Rightarrow\; Cb(\mathit{BI\text{-}Profile})$$

where BI-Profile is the sentence saying that the strategies given by each player's conditional beliefs in the initial state s form a backward-induction profile.

Finally, the following theorem ensures that above results are not vacuous:

Theorem 4.6 *For every extensive game G, there is a game model $\mathcal{M} \in \mathfrak{M}_G$ and a state $s \in \mathcal{M}$ satisfying the sentence*

$$\text{No-Mistakes} \wedge Ck(\mathrm{RP} \wedge \mathit{Strategies} \wedge [!]Eb\text{No-Mistakes}).$$

As a consequence, the sentence $R \wedge Ck[!]Eb\mathrm{R} \wedge Ck Strategies$ is also satisfied.

The *proofs* of these theorems are in **Appendix 1**. Alternative (weaker) conditions ensuring the backward induction outcome are given in **Appendix 2**.

5 Comparison with Other Work

The game-theoretic issues that we deal with in this paper originate in the work of Aumann [2], Stalnaker [30, 31, 32] and Reny [26], and have been investigated by a number of authors [14, 15, 13, 8, 9, 10, 17, 18, 22, 29, 19] etc. Our work obviously owes a great deal to these authors for their illuminating discussions of the topic.

The logic CDL of conditional belief was first introduced and axiomatised by Board [16], in a slightly more complicated form. The version presented here is due to Baltag and Smets [5, 7]. The dynamic extension of CDL obtained by adding the public announcements modalities (coming from the public announcement logic PAL, originally developed by Plaza [24]) has been developed by van Benthem [11] and, independently, by Baltag and Smets [5]. The extension of PAL with arbitrary announcement modalities $[!]\varphi$ is due to Balbiani et al [3]. The belief-revision-friendly version of APAL presented here (obtained by combining APAL with CDL) is an original contribution of our paper.

The work of Battigalli and Siniscalchi [10] is the closest to ours, both through their choice of the basic setting for the "static logic" (also given by conditional belief operators) and through the introduction of a strengthened form of common belief ("common strong belief") as an epistemic basis for a backward-induction theorem. Strong belief, though different from our "stable" belief, is another version of persistent belief: belief that continues to be maintained unless and until it is contradicted by new information. However, their notion of rationality is only "partially dynamic": although taking into account the dynamics of beliefs (using conditional beliefs given node v to assess the rationality of players' choices at v), it does not fully take into account the limitations posed to the set of possible options by the dynamics of "hard knowledge". In common with most other previous notions of rationality, it requires agents to make rational choices at all nodes, including the past ones and the ones that have already been bypassed. As a result, it is enough for a player to make only one "irrational" move to completely shatter the (common) belief (however strong) in rationality; and as a consequence, common strong belief in rationality does not by itself imply backward induction. To obtain their theorem, Battigalli and Siniscalchi have to add another assumption: that the game model is a *complete type structure*, i.e. it contains, in a certain sense, every possible epistemic-doxastic "type" for each player. This means that the players are assumed to have *absolutely no "hard" information, not only about the outcomes or about the other players' strategies, but also about the other players' beliefs*, so that they have to consider as epistemically possible *all* consistent (probabilistic) belief assignments for the other players! This is an extremely extremely strong (and, in our opinion, unrealistic) "completeness" assumption, one that can only be fulfilled in an infinite model. In contrast, the analogue completeness assumption in our approach is the much weaker "Open Future" assumption, postulating that (at the beginning of the game) players have no non-trivial "hard" information about the outcomes (except the information given by the structure of the game): they cannot foretell the future, cannot irrevocably know the players' freely chosen future moves (though they *do* irrevocably know the past, and they *may* irrevocably know the present, including all the beliefs and the plans of action of all the players). Our more realistic postulate is weak enough to be realized on finite models. In particular, it can be realized on models as small as the set of terminal nodes of the game tree (having one state for each terminal node), and in which all the plans of action are common knowledge, so that the only uncertainty concerns possible mistakes in playing (and hence the final outcome).

Samet [29] introduces a notion of *hypothetical knowledge*, in order to develop an epistemic characterisation of backward induction. Hypothetical knowledge looks prima facie similar to conditional belief, except that the interpretation of the hypothetical knowledge formula $K_i^{\varphi}\psi$ is different: "Had φ been the case, i would have known ψ" (op. cit., p. 237). This mixture of counterfactual conditionals and knowledge is specifically introduced in [29] only to discuss backward induction, and it has not occurred before or subsequently in the literature. In contrast, our approach is grounded in the relatively standard and well-understood foundations of Conditional Doxastic Logic, independently studied by logicians and philosophers. While Samet does make what we agree is the important point that some form of counterfactual reasoning is of vital importance to the epistemic situation in extensive games, his model and conditions seem to us more complex, less transparent and less intuitive that ours.

We are aware of only one prior work that uses dynamic epistemic logic (more precisely, the logic of public announcements, but in the context of "classical DEL", i.e. dealing only with knowledge update and not with belief revision) for the analysis of solution concepts in extensive games: van Benthem's work [12]. That work takes Aumann's "static" notion of rationality as given, and accepts Aumann's classical result as valid, and so it does not attempt to deal with the cases in which Aumann's assumptions do not apply, nor to address the criticism and the issues raised by Stalnaker, Reny and others. Instead, van Benthem's contribution focuses on the *sources* of knowledge, on explaining *how* complex epistemic conditions of relevance to Game Theory (such as Aumann's common knowledge of rationality) can *be brought about*, via repeated public announcements of rationality. So van Benthem does *not* use public announcements in order to simulate a play of the game. Public announcements in van Benthem's approach represent *off-line learning*, i.e. *pre-play* or *inter-play* learning, whereas the public announcements in our present approach represent *on-line learning*, i.e. learning that takes place *during the play* of the game. A very interesting open question is to address the same issue answered by van Benthem, but for the case of the dynamic-epistemic condition proposed here, instead of Aumann's condition: find some off-line communication or learning protocol that can achieve common knowledge of stable common belief in rational play.

References

[1] Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

[2] Robert Aumann. Backward induction and common knowledge of rationality. *Games and Economic Behavior*, (8):6–19, 1995.

[3] Philippe Balbiani, Alexandru Baltag, Hans P. van Ditmarsch, Adreas Herzig, Tomohiro Hoshi, and Tiago de Lima. 'knowable' as 'known after an announcement'. 2008. To appear in Review of Symbolic Logic.

[4] Alexandru Baltag, Lawrence S. Moss, and Slawomir Solecki. The logic of public announcements, common knowledge and private suspicions. Technical Report SEN-R9922, Centrum voor Wiskunde en Informatica, 1999.

[5] Alexandru Baltag and Sonja Smets. Conditional doxastic models: A qualitative approach to dynamic belief revision. *Electronic Notes in Theorerical Computer Science*, 165:5–21, 2006.

[6] Alexandru Baltag and Sonja Smets. The logic of conditional doxastic actions. To appear in Texts in Logic and Games, 2008.

[7] Alexandru Baltag and Sonja Smets. A qualitative theory of dynamic interactive belief revision. To appear in Texts in Logic and Games, 2008.

[8] Pierpaolo Battigalli. On rationalizability in extensive games. *Journal of Economic Theory*, 74(1):40–61, May 1997.

[9] Pierpaolo Battigalli and Marciano Siniscalchi. Hierarchies of conditional beliefs and interactive epistemology in dynamic games. *Journal of Economic Theory*, 88(1):188–230, September 1999.

[10] Pierpaolo Battigalli and Marciano Siniscalchi. Strong belief and forward induction reasoning. *Journal of Economic Theory*, 106(2):356–391, October 2002.

[11] Johan van Benthem. Dynamic logic for belief revision. *ILLC Prepublication*, PP-2006(11), 2006.

[12] Johan van Benthem. Rational dynamics and epistemic logic in games. *International Game Theory Review*, 9(1):13–45, 2007. (Erratum reprint, 9(2), 377–409).

[13] Cristina Bicchieri. Self-refuting theories of strategic interaction: a paradox of common knowledge. *Erkenntnis*, 30:69–85, 1989.

[14] Ken Binmore. Modeling rational players, part I. *Economics and Philosophy*, 3:179–214, 1987.

[15] Ken Binmore. A note on backward induction. *Games and Economic Behavior*, 17(1):135–137, November 1996.

[16] Oliver Board. Dynamic interactive epistemology. *Games and Economic Behavior*, 49:49–80, 2002.

[17] Giacomo Bonanno. The logic of rational play in games of perfect information. *Economics and Philosophy*, 7:37–65, 1991.

[18] Adam Brandenburger. The power of paradox: some recent developments in interactive epistemology. *International Journal of Game Theory*, 35(4):465–492, April 2007.

[19] Thorsten Clausing. Doxastic conditions for backward induction. *Theory and Decision*, 54:315–336, 2003.

[20] Jelle D. Gerbrandy and Willem Groeneveld. Reasoning about information change. *Journal of Logic, Language, and Information*, 6:147–169, 1997.

[21] Adam Grove. Two modellings for theory change. *Journal of Philosophical Logic*, 17(2):157–170, 1988.

[22] Joseph Y. Halpern. Substantive rationality and backward induction. 1998.

[23] Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*. MIT Press, 1994.

[24] Jan A. Plaza. Logics of public communications. In M. L. Emrich, M. S. Pfeifer, M. Hadzikadic, and Z. W. Ras, editors, *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems*, pages 201–216, 1989.

[25] Wlodek Rabinowicz. Grappling with the centipede: defense of backward induction for bi-terminating games. *Philosophy and Economics*, 14:95–126, 1998.

[26] Philip Reny. Rationality in extensive form games. *Journal of Economic Perspectives*, (6):92–100, 1992.

[27] Alfréd Rényi. On a new axiomatic theory of probability. *Acta Mathematica Academiae Scientiarum Hungaricae*, 6:285–335, 1955.

[28] Olivier Roy. *Thinking before acting: intentions, logic, rational choice*. PhD thesis, ILLC, Amsterdam, 2008.

[29] Dov Samet. Hypothetical knowledge and games with perfect information. *Games and Economic Behavior*, (17):230–251, 1996.

[30] Robert C. Stalnaker. On the evaluation of solution concepts. *Theory and Decision*, 37:49–73, 1994.

[31] Robert C. Stalnaker. Knowledge, beliefs and counterfactual reasoning in games. *Economics and Philosophy*, 12:133–163, 1996.

[32] Robert C. Stalnaker. Belief revision in games: forward and backward induction. *Mathematical Social Sciences*, 36:31–56, 1998.

[33] Robert C. Stalnaker. On logics of knowledge and belief. *Philosophical Studies*, 128:169–199, 2006.

Appendix 1: Some Proofs

Definition 5.1 For a finite set $\mathcal{O}$ of "outcomes" and a finite set P of "players", we denote by $\mathfrak{Games}(\mathcal{O}, P)$ the class of all perfect information games having any subset of $\mathcal{O}$ as their set of outcomes and having any subset of P as their set of players. $\triangleleft$

Definition 5.2 A sentence is *valid on a game G* if it is true at every state s of every game model $\mathcal{M} \in \mathfrak{M}_G$.

A sentence is *valid over* $\mathfrak{Games}(\mathcal{O}, P)$ if it is valid on every game $G \in \mathfrak{Games}(\mathcal{O}, P)$. $\triangleleft$

Lemma 5.3 *For every perfect information game G, if we denote the root of G by v_0, the first player of G (playing at v_0) by i and the first move of i (the successor node played at v_0) by v_1, then the sentence*

$$R_i^{v_0} \wedge \bigwedge_{u \leftarrow v_0} B_i^u [!u] BI \wedge [!v_1] BI \Rightarrow BI$$

is valid on G.

Proof. This follows directly from the definition of rationality at a node and the definition of BI. The assumption that $B_i^u[!u]$BI is true at s means that all the states (deemed as "most plausible by i conditional on u ") in the set $s_i^u := min_{\leq_i}(\overline{u} \cap [s]_i)$ have only outcomes that are backward induction outcomes in the corresponding subgame: i.e. we have $o(t) \in BI_{G^u}$ for all $t \in s_i^u$. Given that all these outcomes $\{u : u \leftarrow v_0\}$ are consistent with i's knowledge (since we are in a game model), the fact that i is rational at v_0 implies that the successor node v_1 chosen by i must be one that maximises her payoff $h_i(o(s_i^u))$ among all the outcomes in $\bigcup_{u \leftarrow v_0} BI_{G^u}$. But, by the definition, such a node v_1 is exactly the choice prescribed at v_0 by the backward induction strategy! Given this backward-induction choice (v_1) of i at node v_0, and given the fact (ensured by the condition $[!v_1]$BI) that starting from node v_1 everybody *will* play the backward induction choices, we can conclude that the outcome $o(s)$ belongs to the backward induction set of outcomes $BI_{G^v} = BI_G$ for the game G. Hence s satisfies BI. QED

The *Main Lemma* underlying our results is the following:

Lemma 5.4 ***("Main Lemma")*** *Fix a finite set $\mathcal{O}$ of outcomes and a finite set P of players. Let ϕ be any sentence in our language $APAL - CDL$ having the following property: for every game $G \in \mathfrak{Games}(\mathcal{O}, P)$, if we denote the root of G by $v := v_0^G$, the first player of G (playing at v_0) by $i := i_0^G$ and the first move of i (the successor node played at v_0^G) by $v_1 := v_1^G$, then the sentence*

$$\phi \Rightarrow R_i^{v_0} \wedge \bigwedge_{u \leftarrow v_0} B_i^u [!u]\phi \wedge [!v_1]\phi$$

is valid on G.

Under this assumption, we have that the sentence

$$\phi \Rightarrow BI$$

is valid over $\mathfrak{Games}(\mathcal{O}, P)$.

Proof. We need to prove that, for every game $G \in \mathfrak{Games}(\mathcal{O}, P)$, the sentence $\phi \Rightarrow$ BI is valid on G. The proof is *by induction on the length of the game G*.

For games of length 0 (only one outcome, no available moves), the claim is trivial (since the only possible outcome is by definition the backward induction outcome).

Let G be now a game of length $n > 0$, and assume the claim is true for all games of smaller length. Let v_0 be the root of G, i be the first player of G, $\mathcal{M} \in \mathfrak{M}_G$ be a game model for G and s be a state in $\mathcal{M}$ such that $s \models_{\mathcal{M}} \phi$.

Let u be any *arbitrary immediate successor* of v_0 (i.e. any node such that $u \leftarrow v_0$). By the property assumed in the statement of this Lemma, we have that $s \models_{\mathcal{M}} B_i^u[!u]\phi$, and so (if s_i^u is the set defined in the proof of the previous Lemma, then) we have $t \models_{\mathcal{M}} [!u]\phi$ for all $t \in s_i^u$. Hence, we have $t \models_{\mathcal{M}^u} \phi$ for all $t \in s_i^u \cap \overline{u}$. By the induction hypothesis, we must have $t \models_{\mathcal{M}^u}$ BI (since $\mathcal{M}^u$ is a game model for G^u, which has length smaller than G, and so the implication $\phi \Rightarrow$ BI is valid on $\mathcal{M}^u$), for all $t \in s_i^u \cap \overline{u}$. From this we get that $t \models_{\mathcal{M}} [!u]$BI for all $t \in s_i^u$, and hence that $s \models_{\mathcal{M}} B_i^u[!u]$BI.

Let v_1 be now the first move of the game in state s (i.e. the unique immediate successor $v_1 \leftarrow v_0$ such that $s \models_{\mathcal{M}} v_1$). By the property assumed in this Lemma, we have that $s \models_{\mathcal{M}} [!v_1]\phi$. By the same argument as in the last paragraph, the induction hypothesis gives us that $s \models_{\mathcal{M}} [!v_1]$BI. Putting together with the conclusion of the last paragraph and with the fact (following from the theorem's assumption) that $\phi \Rightarrow R_i^{v_0}$ is valid on $\mathcal{M}$, we infer that $s \models_{\mathcal{M}} R_i^{v_0} \wedge \bigwedge_{u \leftarrow v_0} B_i^u[!u]\text{BI} \wedge [!v_1]\text{BI}$. The desired conclusion follows now from Lemma 5.3. QED

Lemma 5.5 *The sentence*

$$\phi \ := \ \text{R} \wedge Ck[!]Eb\text{R}$$

has the property assumed in the statement of Lemma 5.4.

Proof. The claim obviously follows from the following three sub-claims:

1. dynamic rationality is a "stable" property, i.e. the implication R $\Rightarrow \bigwedge_u [!u]$R is valid;
2. the implication $Ck[!]Eb\psi \Rightarrow B_i^u[!u]Ck[!]Eb\psi$ is valid, for all formulas ψ and all nodes $u \in G$;
3. the implication $Ck[!]Eb\psi \Rightarrow [!u]Ck[!]Eb\psi$ is valid, for all formulas ψ and all nodes u.

All these claims are easy exercises in dynamic-epistemic logic. The first follows directly from the definition of dynamic rationality.

The second sub-claim goes as follows: assume that we have $Ck[!]Eb\psi$ at some state of a given model; then we also have $Ck[!u][!]Eb\psi$ for any node u (since $[!]\theta$ implies $[!u][!]\theta$), and so also $K_iCk[!u][!]Eb\psi$ (since common knowledge implies knowledge of common knowledge), from which we get $B_i^uCk[!u][!]Eb\psi$ (since knowledge implies conditional belief under any conditions). This is the same as $B_i^u(u \rightarrow Ck[!u][!]Eb\psi)$, which implies $B_i^u(u \rightarrow Ck^u[!u][!]Eb\psi)$ (since common knowledge implies conditional common knowledge). But this last clause is equivalent to $B_i^u[!u]Ck[!]Eb\psi$ (by the Reduction Law for common knowledge after public announcements).

The third sub-claim goes as follows: assume that we have $Ck[!]Eb\psi$ in some state of a given model; then as before we also have $Ck[!u][!]Eb\psi$, and thus $Ck^u[!u][!]Eb\psi$ (since common knowledge implies conditional common knowledge). From this we get $u \to Ck^u[!u][!]Eb\psi$ (by weakening), which is equivalent to $[!u]Ck[!]Eb\psi$ (by the Reduction Law for common knowledge after public announcements). QED

Theorems 4.4 and 4.2

Proof. Theorem 4.4 follows now from Lemma 5.4 and Lemma 5.5. Theorem 4.2 follows from Theorem 4.4, by applying the operator $Ck[!]Eb$ to both its premiss and its conclusion, and noting that the implication

$$Ck[!]Eb\psi \Rightarrow Ck[!]EbCk[!]Eb\psi$$

is valid. QED

Appendix 2

The epistemic condition $R \wedge Ck[!]EbR$ that was given in this paper (to ensure backward induction) is *not* the weakest possible condition (ensuring this conclusion). Any property ϕ satisfying the assumption of our Main Lemma (Lemma 5.4) would do it. In particular, there exists a *weakest* such condition (the smallest event $E \subseteq S$ such that $E \subseteq R_i^{v_0} \cap \bigcap_{u \leftarrow v_0} B_i^u[!u]E \cap [!v_1]E$), but it is a very complicated and unnatural condition. The one given in the paper seems to be simplest such condition expressible in our language $APAL - CDL$.

However, one can give weaker simple conditions if one is willing to go a bit beyond the language $APAL - CDL$, by adding fixed points for other (definable) epistemic operators.

Let *stable true belief* be a belief that is known to be a stable belief and it is also a stably true belief. Formally, we define:

$$Stb_i\varphi \ := \ K_i[!]B_i\varphi \wedge [!]\varphi.$$

Stable true belief is a form of "knowledge", since *it implies truth and belief:*

$$Stb_i\varphi \Rightarrow \varphi \wedge B_i\varphi$$

(and in fact *it implies stable truth*: $Stb_i\varphi \Rightarrow [!]\varphi$). *Knowledge that something is stably true implies stable true belief in it*:

$$K_i[!]\varphi \Rightarrow Stb_i\varphi.$$

Stable true belief is inherently a "positively introspective" attitude, i.e.

$$Stb_i\varphi \Rightarrow Stb_iStb_i\varphi,$$

but it is not positively introspective with respect to ("hard") knowledge:

$$Stb_i\varphi \not\Rightarrow K_iStb_i\varphi.$$

Stable true belief is *not* negatively introspective, neither inherently nor with respect to knowledge.

We can define *common stable true belief* in the same way as common knowledge: first define *general stable true belief*

$$Estb\varphi = \bigwedge_{i \in P} Stb_i\varphi$$

("everybody has stable true belief"), then put

$$Cstb\varphi = \bigwedge_n (Estb)^n\varphi.$$

Note that this definition, although semantically meaningful, is not a definition in our language $APAL - CDL$, since it uses infinite conjunctions. Indeed, we conjecture that common stable true belief is undefinable in the language $APAL-CDL$, since it doesn't seem to be expressible as a combination of common knowledge, common belief and dynamic operators.

Lemma 5.6 *The sentence $Cstb$R satisfies the assumptions of our Main Lemma (Lemma 5.4).*

As an immediate consequence, we have:

Theorem 5.7 The sentence

$$Cstb\mathrm{R} \Rightarrow \mathrm{BI}$$

is valid over game models. In English: (if we assume common knowledge of the structure of the structure of the game and of open future, then) *common stable true belief in (dynamic) rationality implies the backward induction outcome* .

Strategies made explicit in Dynamic Game Logic

Sujata Ghosh
Center for Soft Computing Research, Indian Statistical Institute, Kolkata
Department of Mathematics, Visva-Bharati, Santiniketan
sujata_t@isical.ac.in

Abstract

We propose an explicit logic of strategies ($SDGL$) in the Dynamic Game Logic (DGL) framework and provide a complete axiomatization for this logic. Some discussions are put forward regarding $SDGL$ and DGL, raising an interesting issue about their combination.

1. Introduction: cudos for strategies

Many events that happen in our daily life can be thought of as games. In fact, besides the 'games' in the literal sense, our day-to-day dialogues, interactions, legal procedures, social and political actions, biological phenomena - all these can be viewed as games together with their goals and strategies. The theory of games has its various applications in the areas of economics, logic, computer science as well as linguistics. Games play a very important role in modelling intelligent interaction. In Rubinstein's words, 'I view game theory as an analysis of concepts used in social reasoning when dealing with situations of conflict' [16].

As evident from the existing literature, much of game theory deals with strategic equilibriums. Various equilibrium theories have been developed till date both for zero-sum as well as non zero-sum games, starting from the initial concept of Nash [12], which have their implications in the studies of the society. They help in providing a 'plan of action' to the agents participating in the state of affairs, which could be articulated as 'games', when faced with strategic decision making in situations of conflict.

Over the past few decades a lot of work has been done in the epistemic foundations of game theory, studying the formal logics of knowledge and belief. The formal systems expressing players' knowledge and beliefs about themselves as well as their competitors were looked at in much details - a tremendous amount of work is still going on. But a very related and relevant issue - players' strategies/plan of actions to play the game, which they base on their epistemic states almost have rarely been looked upon, until very recently. To mention a few, [15] proposes a logic of strategies in games over finite graphs, whereas [17] makes strategies explicit in Alternating-time Temporal Logic. The incorporation of 'strategies' within the logical language would very well aid in the currently popular ventures into social choice mechanism designs.

Strategies of the players playing the game form a basic ingredient of game theory, whether looked upon from the winning point of view or from the best-response one. A lot of other issues like the rationality of the players, their goals and preferences are also very important issues, but they are outside the scope of this work, though we plan to incorporate them in the future.

Our main goal in this work is to incorporate explicit notions of strategies in the framework of Dynamic Game Logic (DGL) [13]. Not unlike other logics talking about game and coalition structures [1, 14], DGL suffers from '$\exists$-sickness' : the detailed level of game structures getting suppressed by existential quantifiers of "having a strategy" [7]. We intend to provide a logic ($SDGL$) that makes the game structures explicit to a great extent.

In general, strategies are partial transition relations and hence dynamic modal logic provides a good framework to talk about them, as mentioned in [4, 5]. But the main challenge here is to combine the strategy calculus together with the game calculus. As one can easily guess, the constructs of Propositional Dynamic Logic [11] play an important role in achieving this amalgamation. In this regard, we should mention that, a lot of discussions and proposals have already been put forward by van Benthem [9, 8]. This effort can be looked upon as a follow-up of one of these proposals.

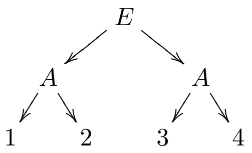
After providing a brief overview of DGL in the next section, we propose a logic for strategizing DGL ($SDGL$) in section 3 with a complete axiomatization. Section 4 provides some discussions over the two logics DGL and $SDGL$, with several pointers towards future work mentioned in the last one.

2. Dynamic Game Logics : an overview

We now give a brief review of DGL, the dynamic game logic of two-person sequential games in this section, which was first proposed in [13], and further developed by [14], [6], [10] and others. DGL talks about 'generic' games which can be played starting from any state s on the 'game boards' and the semantics is based on the 'forcing relations' describing the powers each player has to end in a set of final states, starting from a single initial state.

> $s\rho^i_G X$: player i has a *strategy* for playing game G from state s onwards, whose resulting final states are always in the set X, whatever the other players choose to do.

To exemplify, let us move onto real extensive games for once. Consider the game tree:



In this game, player *E* has two strategies, forcing the sets of end states $\{1, 2\}$, $\{3, 4\}$, while player *A* has four strategies, forcing one of the sets $\{1, 3\}$, $\{1, 4\}$, $\{2, 3\}$, $\{2, 4\}$.

These forcing relations satisfy the following two simple set-theoretic conditions [3]:

(C1) Monotonicity: If $s\rho^i_G X$ and $X \subseteq X'$, then $s\rho^i_G\, X'$.

(C2) Consistency: If $s\rho^E_G Y$ and $s\rho^A_G Z$, then Y, Z overlap.

In the semantics of DGL as proposed in [13, 14], another extra condition is assumed :

(C3) Determinacy: If it is not the case that $s\rho^E_G X$, then, $s\rho^A_G S$-X, and the same for A vis-a-vis E.

Both [13] and [14] talks about determined games. This simplifies things a lot, but fails to express the roles of the players in the games. The dynamic logic for non-determined games was studied extensively in [10] which also introduced the notion of parallel games in the syntax. For the present work the concurrent game construct has not been dealt with. The *iteration* operation for repeated play as present in [13] has also not been considered here. We only consider the following constructs which form new games: *choice* ($G \cup G'$)*, dual* (G^d)*, and sequential composition* $(G; G')$. The readers could easily guess the intuitive meanings of these constructs. For the sake of continuation to the next section, in what follows, the DGL for non-determined games has been briefly discussed. To start with, it should be noted here that the players' powers have a recursive structure in the complex games:

Fact 2.1 *Forcing relations for players in complex sequential two-person games satisfy the following equivalences:*

$$\begin{array}{rcc}
s\rho^E_{G\cup G'}\, X & \textit{iff} & s\rho^E_G\, X \textit{ or } s\rho^E_{G'} X \\
s\rho^A_{G\cup G'}\, X & \textit{iff} & s\rho^A_G X \textit{ and } s\rho^A_{G'} X \\
s\rho^E_{G^d} X & \textit{iff} & s\rho^A_G X \\
s\rho^A_{G^d} X & \textit{iff} & s\rho^E_G X \\
s\rho^i_{G;G'} X & \textit{iff} & \exists Z : s\rho^i_G Z \textit{ and for all } z \in Z,\ z\rho^i_{G'} X.
\end{array}$$

The basic models that play the role of game boards are defined as follows:

Definition 2.2 A game model is a structure $\mathcal{M} = (S, \{\rho^i_g \mid g \in \Gamma\}, V)$, where S is a set of states, V is a valuation assigning truth values to atomic propositions in states, and for each $g \in \Gamma$, $\rho^i_g \subseteq$ S $\times$ $\mathcal{P}$(S). We assume that for each g, the relations are upward closed under supersets (the earlier Monotonicity), while also, the earlier Consistency condition holds for the forcing relations of the players A, E. $\triangleleft$

The language of DGL (without game iteration) is defined as follows:

Definition 2.3 Given a set of atomic games Γ and a set of atomic propositions Φ, game terms γ and formulas ϕ are defined inductively:

$$\gamma := g \mid \phi? \mid \gamma;\gamma \mid \gamma \cup \gamma \mid \gamma^d$$
$$\phi := \bot \mid p \mid \neg\phi \mid \phi \vee \phi \mid \langle\gamma, i\rangle\phi,$$

where $p \in \Phi$, $g \in \Gamma$ and $i \in \{A, E\}$.

$\triangleleft$

The truth definition for formulas ϕ in a model $\mathcal{M}$ at a state s is standard, except for the modality $\langle\gamma, i\rangle\phi$, which is interpreted as follows:

> $\mathcal{M}, s \models \langle\gamma, i\rangle\phi$ iff there exists X : $s\rho^i_\gamma X$ and $\forall x \in X : \mathcal{M}, x \models \phi$.

The complete axiomatization of this logic has been proposed and proved in [10] :

Theorem 2.4 *DGL is complete and its validities are axiomatized by the following axioms:*

a) all propositional tautologies and inference rules

b) if $\vdash \phi \to \psi$ *then* $\vdash \langle g, i\rangle\phi \to \langle g, i\rangle\psi$

c) $\langle g, E\rangle\phi \to \neg\langle g, A\rangle\neg\phi$

d) reduction axioms:

$$\langle\alpha\cup\beta, E\rangle\phi \leftrightarrow \langle\alpha, E\rangle\phi \vee \langle\beta, E\rangle\phi$$
$$\langle\alpha\cup\beta, A\rangle\phi \leftrightarrow \langle\alpha, A\rangle\phi \wedge \langle\beta, A\rangle\phi$$
$$\langle\gamma^d, E\rangle\phi \leftrightarrow \langle\gamma, A\rangle\phi$$
$$\langle\gamma^d, A\rangle\phi \leftrightarrow \langle\gamma, E\rangle\phi$$
$$\langle\alpha;\beta, i\rangle\phi \leftrightarrow \langle\alpha, i\rangle\langle\beta, i\rangle\phi$$
$$\langle\delta?, E\rangle\phi \leftrightarrow (\delta\wedge\phi)$$
$$\langle\delta?, A\rangle\phi \leftrightarrow (\neg\delta\wedge\phi).$$

This logic is also decidable. As can be noticed, the truth definition of the modal game formulas of the form $\langle\gamma, i\rangle\phi$ is given in terms of existence of strategies, without going into their structures. In what follows, the strategy structures have been explicitly dealt together with the game structures.

3. Strategizing DGL

3.1. A logic for strategies

Mentioning strategies explicitly in the dynamic game logic framework prompt us to divert from the usual DGL semantics that takes into consideration 'generic' games on game boards. The whole point is to bring *strategies* within the logical language which till now have their place in giving meaning to the game as well as coalition modalities [14]. Adding explicit strategy terms to DGL, the language of Strategized DGL ($SDGL$) is defined by,

Definition 3.1 Given a set of atomic games Γ, a set of atomic strategies Σ, a *finite* set of atomic actions Π and a set of atomic propositions Φ, game terms γ, strategy terms σ, action terms π and formulas ϕ are defined inductively in the following way:

$$\gamma := g \mid \phi? \mid \gamma^d \mid \gamma;\gamma \mid \gamma\cup\gamma$$
$$\sigma := s \mid \sigma\cup\sigma \mid \sigma;\sigma$$
$$\pi := b \mid \pi\cup\pi \mid \pi^*$$
$$\phi := \bot \mid p \mid \neg\phi \mid \phi\vee\phi \mid [\pi]\phi \mid \langle\pi\rangle\phi \mid \langle\sigma, i, \gamma\rangle\phi$$

where $p\in\Phi$, $s\in\Sigma$, $g\in\Gamma$, $b\in\Pi$, and $i\in\{A, E\}$. ◁

Regarding the intuitive understanding of the strategy terms, '$\cup$' corresponds to the choice of strategies, and ';' to the composition of strategies. It should be mentioned here that, the way the semantics is given later, it would have been enough to use just one combination operation of the strategy terms. The use of both of them aids in understanding the intuition behind their usage.

Moving away from the 'generic' game structures, the models take the form of extensive game trees with a few additional actions. Before going into all these, we need a parent model which is given as follows.

Definition 3.2 A model is a structure $\mathcal{M}$ = $\langle$ S, $\{R_\pi : \pi$'s are actions$\}$, *ref*, **L**, **R**, V $\rangle$, where S is a set of states and V is a valuation assigning truth values to atomic propositions in states. For each π, R_π is a binary relation on S. *ref*, **L**, **R** are all reflexive relations over S, with $\langle$ S, $\{R_\pi : \pi$'s are actions$\}\rangle$ forming a regular action frame. ◁

In this model, atomic and composite games from a specified 'start'-state are defined in the following. It should be mentioned that all these game structures are taken to be finite, defined over finite subsets of S.

Definition 3.3 Game($\mathcal{M}$, s, γ) is a structure defined recursively as follows:

(i) For atomic games g, Game($\mathcal{M}$, s, g) is a structure given as follows: $\langle W\subseteq S,\ s\in W,\ \{R_b\downarrow_W\!: b\in\Pi\}$, $V = V^{\mathcal{M}}\downarrow_W, P : W\to\{E, A, end\}\rangle$.

(ii) For test games $\phi?$, Game($\mathcal{M}$, s, $\phi?$) is a structure given as follows: $\langle\{s\}$, s, $ref\downarrow_{\{s\}}$, $V = V^{\mathcal{M}}\downarrow_{\{s\}}$, P : $\{s\}\to\{\ end\}\rangle$.

(iii) Given Game($\mathcal{M}$, s, γ), Game($\mathcal{M}$, s, γ^d) is the structure, $\langle W\subseteq S, s\in W, \{R_b\downarrow_W\!: b\in\Pi\}, V = V^{\mathcal{M}}\downarrow_W, P : W\to\{E, A, end\}\rangle$, where all the constituents of the structure are the same as the corresponding ones in Game($\mathcal{M}$, s, γ), except for P_{γ^d}, which satisfies the property: P_{γ^d}(w) = E/A, whenever P_γ(w) = A/E, respectively.

(iv) Given Game($\mathcal{M}$, s, γ) and Game($\mathcal{M}$, s, γ'), Game($\mathcal{M}$, s, $\gamma\cup\gamma'$) is the structure given by, $\langle W\subseteq S$, $s\in W$, $\{R_b\downarrow_W\!: b\in\Pi\}$, $\mathbf{L}\downarrow_{\{s,s\}}$, $\mathbf{R}\downarrow_{\{s,s\}}$, $V = V^{\mathcal{M}}\downarrow_W$, $P : W\to\{E, A, end\}\rangle$, where W = $W_\gamma\uplus W_{\gamma'}$, and P extends both P_γ and $P_{\gamma'}$.

(v) Given Game($\mathcal{M}$, s_1, γ) and Game($\mathcal{M}$, s_2, γ'), Game($\mathcal{M}$, s, $\gamma;\gamma'$) is defined if for each $t\in P_\gamma^{-1}(end)$, Game($\mathcal{M}$, t, γ') can be defined. Suppose we have Game($\mathcal{M}$, t_1, γ'), …, Game($\mathcal{M}$, t_n, γ'). In that case, Game($\mathcal{M}$, s, $\gamma;\gamma'$) is the structure $\langle W\subseteq S, s\in W, \{R_b\downarrow_W\!: b\in\Pi\}, V = V^{\mathcal{M}}\downarrow_W, P : W\to\{E, A, end\}\rangle$, where W = $W_\gamma\cup W^1_{\gamma'}\cup\ldots\cup W^n_{\gamma'}$; $s = s_1$; P extends P_γ, P^1_γ, …, P^n_γ, with the restriction that for $w\in P_\gamma^{-1}(end)\cap W$, $P(w) = P_{\gamma'}(s_2)$. ◁

Because of some technical reasons regarding satisfiability, *choice* games can only be defined for the games with the same initial state, which is not really a big issue. The sequential composition game could also be defined under certain restrictions as mentioned above. It is now time to define strategies of the players in a game, which again has a recursive definition. Note that we will only talk about full strategies here and the definition is given likewise.

Definition 3.4 Given Game($\mathcal{M}$, s, γ), a strategy for a player i, given by the relation $\mathcal{R}_i^\gamma$ is defined by,

(i) For Game($\mathcal{M}$, s, g), $\mathcal{R}_E^g[\mathcal{R}_A^g] \subseteq \bigcup\{R_b \downarrow_{W_g} : b \in \Pi\}$ satisfying the following conditions:

(a) $s \in Dom(\mathcal{R}_E^g)[Dom(\mathcal{R}_A^g)]$, and $Ran(\mathcal{R}_E^g)$ $[Ran(\mathcal{R}_A^g)] \cap P_g^{-1}(end) \neq \emptyset$

(b) For each $t \in P_g^{-1}(E, A) - \{s\}$, $t \in Dom(\mathcal{R}_E^g)$ $[Dom(\mathcal{R}_A^g)]$ iff $t \in Ran(\mathcal{R}_E^g)[Ran(\mathcal{R}_A^g)]$.

(c) For each $s \in P^{-1}(E)[P^{-1}(A)], \exists$ unique s' such that $(s, s') \in \mathcal{R}_E^g[\mathcal{R}_A^g]$.

(d) For each $s \in P^{-1}(A)[P^{-1}(E)]$, $(s, s') \in \bigcup\{R_b \downarrow_{W_g} : b \in \Pi\}$ implies $(s, s') \in \mathcal{R}_E^g[\mathcal{R}_A^g]$.

(e) Nothing else is in $\mathcal{R}_E^g[\mathcal{R}_A^g]$.

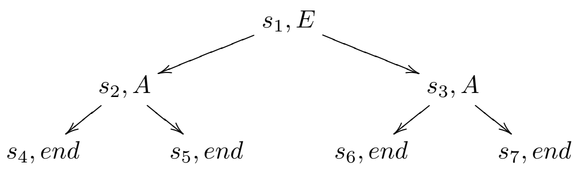
(ii) For Game($\mathcal{M}$, s, $\phi?$), $\mathcal{R}_i^{\phi?}$ = *ref* $\downarrow_{\{s\}}$.

(iii) For Game($\mathcal{M}$, s, γ^d), $\mathcal{R}_E^{\gamma^d} = \mathcal{R}_A^\gamma$, and $\mathcal{R}_A^{\gamma^d} = \mathcal{R}_E^\gamma$.

(iv) For Game($\mathcal{M}$, s, $\gamma \cup \gamma'$),
$\mathcal{R}_E^{\gamma\cup\gamma'}$ = $\mathbf{L} \downarrow_{\{s,s\}} \cup \mathcal{R}_E^\gamma$ or, $\mathbf{R} \downarrow_{\{s,s\}} \cup \mathcal{R}_E^{\gamma'}$,
and, $\mathcal{R}_A^{\gamma\cup\gamma'}$ = $\mathbf{L} \downarrow_{\{s,s\}} \cup \mathbf{R} \downarrow_{\{s,s\}} \cup \mathcal{R}_A^\gamma \cup \mathcal{R}_A^{\gamma'}$.

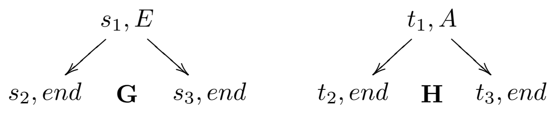
(v) For Game($\mathcal{M}$, s, $\gamma; \gamma'$), $\mathcal{R}_i^{\gamma;\gamma'}$ = $\mathcal{R}_i^\gamma \cup \mathcal{R}_{i_{j_1}}^{\gamma'} \cup \ldots \cup \mathcal{R}_{i_{j_l}}^{\gamma'}$, where the indices correspond to the number of times the 'end'-state is reached in $\mathcal{R}_i^\gamma$. ◁

For an example of the players' strategies, consider the simple extensive game tree:



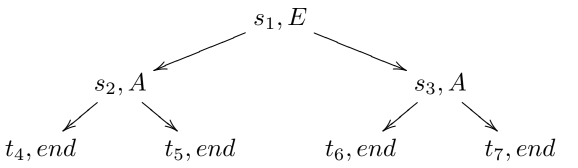
The strategies of E are $\{(s_1, s_2), (s_2, s_4), (s_2, s_5)\}$, and $\{(s_1, s_3), (s_3, s_6), (s_3, s_7)\}$, whereas the strategies for A are $\{(s_1, s_2), (s_1, s_3), (s_2, s_4), (s_3, s_6)\}$, $\{(s_1, s_2), (s_1, s_3), (s_2, s_4), (s_3, s_7)\}$, $\{(s_1, s_2), (s_1, s_3), (s_2, s_5), (s_3, s_6)\}$, and $\{(s_1, s_2), (s_1, s_3), (s_2, s_5), (s_3, s_7)\}$.

If we consider the *choice* operations of two such games, the strategies of the players could be easily computed. For the *sequential composition*, consider the following two games:



The strategies of E in G are $\{(s_1, s_2)\}$, and $\{(s_1, s_3)\}$, and in H is $\{(t_1, t_2), (t_1, t_3)\}$, and similarly, that for A in G is $\{(t_1, t_2), (t_1, t_3)\}$, and in H are $\{(t_1, t_2)\}$, and $\{(t_1, t_3)\}$.

Suppose, the model is such that $G; H$ could be defined and it is as follows:



The readers can notice that it is just the game given as example earlier, and hence could easily verify that the strategies of the players in this complex game conform with the definition given to compute the strategies of the *sequential composition* games, from the simpler ones.

Before going into the truth-definitions of formulas, let us mention a few words about interpreting the strategy terms of the language. The strategy terms are always interpreted corresponding to some game structure Game($\mathcal{M}$, s, γ) and player i. Let $\mathbb{R}_i^\gamma$ denote the set of all strategies for player i in Game($\mathcal{M}$, s, γ).

Definition 3.5 Given Game($\mathcal{M}$, s, γ) and player i, a strategy function $\mathcal{F}_i^\gamma$ is a partial function from the set of all strategy terms to $\mathbb{R}_i^\gamma$, satisfying the following conditions.

(i) For $s \in \Sigma$, $\mathcal{F}_i^\gamma(s)$ is defined, only when γ is an atomic or a test game.

(ii) For the choice game $\alpha \cup \beta$, $\mathcal{F}_i^{\alpha\cup\beta}$ is given by,

$$\mathcal{F}_E^{\alpha\cup\beta}(\sigma \cup \tau) = \mathbf{L} \downarrow_{W_{\alpha\cup\beta}} \cup \mathcal{R}_E^\alpha \quad \text{iff} \quad \mathcal{F}_E^\alpha(\sigma) = \mathcal{R}_E^\alpha,$$

$$\mathcal{F}_E^{\alpha\cup\beta}(\sigma \cup \tau) = \mathbf{R} \downarrow_{W_{\alpha\cup\beta}} \cup \mathcal{R}_E^\beta \quad \text{iff} \quad \mathcal{F}_E^\beta(\tau) = \mathcal{R}_E^\beta,$$

$$\mathcal{F}_A^{\alpha\cup\beta}(\sigma \cup \tau) = \mathcal{R}_A^{\alpha\cup\beta} \quad \text{iff} \quad \mathcal{F}_A^\alpha(\sigma) = \mathcal{R}_A^\alpha \text{ and, } \mathcal{F}_A^\beta(\tau) = \mathcal{R}_A^\beta.$$

(iii) $\mathcal{F}_E^{\gamma^d}(\sigma) = \mathcal{R}_E^{\gamma^d}$ iff $\mathcal{F}_A^\gamma(\sigma) = \mathcal{R}_A^\gamma$, and,
$\mathcal{F}_A^{\gamma^d}(\sigma) = \mathcal{R}_E^{\gamma^d}$ iff $\mathcal{F}_A^\gamma(\sigma) = \mathcal{R}_A^\gamma$.

(iv) For the composition game $\alpha; \beta$, $\mathcal{F}_i^{\alpha;\beta}$ satisfies,

$$\mathcal{F}_i^{\alpha;\beta}(\tau; \eta) = \mathcal{R}_i^{\alpha;\beta} \quad \text{iff} \quad \mathcal{F}_i^\alpha(\tau) = \mathcal{R}_i^\alpha \text{ and, } \mathcal{F}_i^\beta(\eta) = \mathcal{R}_i^\beta.$$

◁

Note that the way these partial functions are given, it takes care of the cases of mismatched syntax (like, $\langle\sigma \cup \tau, E, \alpha; \beta\rangle\phi$), which does not have any corresponding structure in the model. For the semantics of our language, we define the truth of a formula ϕ in $\mathcal{M}$ at a state s in the obvious manner, with the action modalities defined in the usual PDL-style and the following key clause for the game-strategy modality:

> $\mathcal{M}, s \models \langle \sigma, i, \gamma \rangle \phi$ iff for all $s' \in Ran(\mathcal{F}_i^\gamma(\sigma)) \cap P^{-1}(end)$ in Game$(\mathcal{M}, s, \gamma)$, $\mathcal{M}, s' \models \phi$.

Here are some validities of this logic.

- $\langle \sigma, i, \gamma \rangle \phi \to \langle \sigma, i, \gamma \rangle (\phi \vee \psi)$
- $\langle \sigma, i, \gamma \rangle (\phi \wedge \psi) \leftrightarrow \langle \sigma, i, \gamma \rangle \phi \wedge \langle \sigma, i, \gamma \rangle \psi$

3.2. Axioms and completeness

We now provide a complete axiomatization of $SDGL$.

Theorem 3.6 *$SDGL$ is complete and its validities are axiomatized by*

a) all propositional tautologies and inference rules

b) generalization rule for the action modalities

c) axioms for the action constructs:

$$[\pi](\phi \to \psi) \to ([\pi]\phi \to [\pi]\psi)$$
$$\langle \pi \rangle \phi \leftrightarrow \neg[\pi]\neg\phi$$
$$\langle \pi_1 \cup \pi_2 \rangle \phi \leftrightarrow \langle \pi_1 \rangle \phi \vee \langle \pi_2 \rangle \phi$$
$$\langle \pi^* \rangle \phi \leftrightarrow (\phi \vee \langle \pi \rangle \langle \pi^* \rangle \phi)$$
$$[\pi^*](\phi \to [\pi]\phi) \to (\phi \to [\pi^*]\phi)$$

d) $\langle s, i, g \rangle \phi \to \langle b_1 \cup \ldots \cup b_n \rangle \langle (b_1 \cup \ldots \cup b_n)^* \rangle \phi$, where $\Pi = \{b_1, \ldots, b_n\}$

e) $\langle \sigma, i, \gamma \rangle (\phi \to \psi) \to (\langle \sigma, i, \gamma \rangle \phi \to \langle \sigma, i, \gamma \rangle \psi)$

f) if $\vdash \phi \to \psi$ then $\vdash \langle \sigma, i, \gamma \rangle \phi \to \langle \sigma, i, \gamma \rangle \psi$

g) reduction axioms:

$$\langle \sigma \cup \tau, E, \alpha \cup \beta \rangle \phi \leftrightarrow \langle \sigma, E, \alpha \rangle \phi \vee \langle \tau, E, \beta \rangle \phi$$
$$\langle \sigma \cup \tau, A, \alpha \cup \beta \rangle \phi \leftrightarrow \langle \sigma, A, \alpha \rangle \phi \wedge \langle \tau, A, \beta \rangle \phi$$
$$\langle \sigma, E, \gamma^d \rangle \phi \leftrightarrow \langle \sigma, A, \gamma \rangle \phi$$
$$\langle \sigma, A, \gamma^d \rangle \phi \leftrightarrow \langle \sigma, E, \gamma \rangle \phi$$
$$\langle \tau; \eta, i, \alpha; \beta \rangle \phi \leftrightarrow \langle \tau, i, \alpha \rangle \langle \eta, i, \beta \rangle \phi$$
$$\langle \sigma, E, \delta? \rangle \phi \leftrightarrow (\delta \wedge \phi)$$
$$\langle \sigma, A, \delta? \rangle \phi \leftrightarrow (\neg\delta \wedge \phi)$$

h) strategy rules:

for each $X \subseteq \Pi$, the rule below:

if $\vdash \phi \to \langle (\cup X) \rangle \langle (\cup X)^* \rangle \psi$ then $\vdash \phi \to \langle s, i, g \rangle \psi$.

Proof. Soundness of some of the interesting reduction axioms and the strategy rules for the game-strategy modality are shown below. The readers can easily verify the validity of the rest.

1. $\langle \sigma \cup \tau, E, \alpha \cup \beta \rangle \phi \leftrightarrow \langle \sigma, E, \alpha \rangle \phi \vee \langle \tau, E, \beta \rangle \phi$

Suppose $\mathcal{M}, s \models \langle \sigma \cup \tau, E, \alpha \cup \beta \rangle \phi$. Then, for all $s' \in Ran(\mathcal{F}_E^{\alpha\cup\beta}(\sigma \cup \tau)) \cap P_{\alpha\cup\beta}^{-1}(end)$ in Game$(\mathcal{M}, s, \alpha \cup \beta)$, $\mathcal{M}, s' \models \phi$. Now, $\mathcal{F}_E^{\alpha\cup\beta}(\sigma \cup \tau) = \mathbf{L} \downarrow_{W_{\alpha\cup\beta}} \cup \mathcal{R}_E^\alpha$ or, $\mathbf{R} \downarrow_{W_{\alpha\cup\beta}} \cup \mathcal{R}_E^\beta$. W.l.o.g. suppose that $\mathcal{F}_E^{\alpha\cup\beta}(\sigma \cup \tau) = \mathbf{L} \downarrow_{W_{\alpha\cup\beta}} \cup \mathcal{R}_E^\alpha$. Then, for all $s' \in Ran(\mathbf{L} \downarrow_{W_{\alpha\cup\beta}} \cup \mathcal{R}_E^\alpha) \cap P_{\alpha\cup\beta}^{-1}(end)$ in Game$(\mathcal{M}, s, \alpha \cup \beta)$, $\mathcal{M}, s' \models \phi$. By definition of strategies in $\cup$ games, this implies that, for all $s' \in Ran(\mathcal{R}_E^\alpha) \cap P_\alpha^{-1}(end)$ in Game$(\mathcal{M}, s, \alpha)$, $\mathcal{M}, s' \models \phi$. Hence, for all $s' \in Ran(\mathcal{F}_E^\alpha(\sigma)) \cap P_\alpha^{-1}(end)$ in Game$(\mathcal{M}, s, \alpha)$, $\mathcal{M}, s' \models \phi$.So, we have that, $\mathcal{M}, s \models \langle \sigma, E, \alpha \rangle \phi$. Similarly, if $\mathcal{F}_E^{\alpha\cup\beta}(\sigma \cup \tau) = \mathbf{R} \downarrow_{W_{\alpha\cup\beta}} \cup \mathcal{R}_E^\beta$, one can show that, $\mathcal{M}, s \models \langle \tau, E, \alpha \rangle \phi$. So, $\mathcal{M}, s \models \langle \sigma, E, \alpha \rangle \phi$ or $\mathcal{M}, s \models \langle \tau, E, \alpha \rangle \phi$. Hence, $\mathcal{M}, s \models \langle \sigma, E, \alpha \rangle \phi \vee \langle \tau, E, \beta \rangle \phi$.

For the converse, suppose that $\mathcal{M}, s \models \langle \sigma, E, \alpha \rangle \phi$. Then, for all $s' \in Ran(\mathcal{F}_E^\alpha(\sigma)) \cap P_\alpha^{-1}(end)$ in Game$(\mathcal{M}, s, \alpha)$, $\mathcal{M}, s' \models \phi$. So, for all $s' \in Ran(\mathcal{R}_E^\alpha) \cap P_\alpha^{-1}(end)$ in Game$(\mathcal{M}, s, \alpha)$, $\mathcal{M}, s' \models \phi$, which implies that, for all $s' \in Ran(\mathbf{L} \downarrow_{W_{\alpha\cup\beta}} \cup \mathcal{R}_E^\alpha) \cap P_{\alpha\cup\beta}^{-1}(end)$ in Game$(\mathcal{M}, s, \alpha \cup \beta)$, $\mathcal{M}, s' \models \phi$. Hence, reasoning as earlier we have that, $\mathcal{M}, s \models \langle \sigma \cup \tau, E, \alpha \cup \beta \rangle \phi$. The proof for the other disjunct can be dealt with in a similar manner.

2. $\langle \sigma \cup \tau, A, \alpha \cup \beta \rangle \phi \leftrightarrow \langle \sigma, A, \alpha \rangle \phi \wedge \langle \tau, A, \beta \rangle \phi$

Suppose $\mathcal{M}, s \models \langle \sigma \cup \tau, A, \alpha \cup \beta \rangle \phi$. Then, for all $s' \in Ran(\mathcal{F}_A^{\alpha\cup\beta}(\sigma \cup \tau)) \cap P_{\alpha\cup\beta}^{-1}(end)$ in Game$(\mathcal{M}, s, \alpha \cup \beta)$, $\mathcal{M}, s' \models \phi$. Now, $\mathcal{F}_A^{\alpha\cup\beta}(\sigma \cup \tau) = \mathcal{R}_A^{\alpha\cup\beta}$, i.e. $\mathbf{L} \downarrow_{\{s,s\}} \cup \mathbf{R} \downarrow_{\{s,s\}} \cup \mathcal{R}_A^\gamma \cup \mathcal{R}_A^{\gamma'}$. It follows that for all $s' \in Ran(\mathcal{R}_A^\alpha) \cap P_\alpha^{-1}(end)$ in Game$(\mathcal{M}, s, \alpha)$, $\mathcal{M}, s' \models \phi$, and for all $s' \in Ran(\mathcal{R}_A^\beta) \cap P_\beta^{-1}(end)$ in Game$(\mathcal{M}, s, \beta)$, $\mathcal{M}, s' \models \phi$. Then, from the definitions it follows that $\mathcal{M}, s \models \langle \sigma, A, \alpha \rangle \phi \wedge \langle \tau, A, \beta \rangle \phi$. The converse can be proved by retracing the steps backwards.

3. $\langle \sigma, E, \gamma^d \rangle \phi \leftrightarrow \langle \sigma, A, \gamma \rangle \phi$

Suppose $\mathcal{M}, s \models \langle \sigma, E, \gamma^d \rangle \phi$. Then, for all $s' \in Ran(\mathcal{F}_E^{\gamma^d}(\sigma)) \cap P_{\gamma^d}^{-1}(end)$ in Game$(\mathcal{M}, s, \gamma^d)$, $\mathcal{M}, s' \models \phi$. By definition of strategies in the dual game, this implies that, for all $s' \in Ran(\mathcal{F}_A^\gamma(\sigma)) \cap P_\gamma^{-1}(end)$ in Game$(\mathcal{M}, s, \gamma)$, $\mathcal{M}, s' \models \phi$ and so, $\mathcal{M}, s \models \langle \sigma, A, \gamma \rangle \phi$. For the converse proof, retrace back.

4. $\langle \tau; \eta, i, \alpha; \beta \rangle \phi \leftrightarrow \langle \tau, i, \alpha \rangle \langle \eta, i, \beta \rangle \phi$

Suppose $\mathcal{M}, s \models \langle \tau; \eta, i, \alpha; \beta \rangle \phi$. Then, for all

$s' \in Ran(\mathcal{F}_i^{\alpha;\beta}(\sigma;\tau)) \cap P_{\alpha;\beta}^{-1}(end)$ in Game$(\mathcal{M}, s, \alpha;\beta)$, $\mathcal{M}, s' \models \phi$. Hence, for all $s' \in Ran(\mathcal{R}_i^\alpha \cup \mathcal{R}_{i_{j_1}}^\beta \cup \ldots \cup \mathcal{R}_{i_{j_l}}^\beta) \cap P_{\alpha;\beta}^{-1}(end)$ in Game$(\mathcal{M}, s, \alpha;\beta)$, $\mathcal{M}, s' \models \phi$. Then, for k = 1, …, l, for all $t \in Ran(\mathcal{R}_{i_{j_k}}^\beta) \cap P_{\beta_{j_k}}^{-1}(end)$ in Game$(\mathcal{M}, t_{j_k}, \beta)$, $\mathcal{M}, t \models \phi$, and hence for all $t' \in Ran(\mathcal{R}_i^\alpha) \cap P_\alpha^{-1}(end)$, in Game$(\mathcal{M}, s, \alpha)$, $\mathcal{M}, t' \models \langle \eta, i, \beta \rangle \phi$. So, $\mathcal{M}, s \models \langle \tau, i, \alpha \rangle \langle \eta, i, \beta \rangle \phi$.

For the converse part, suppose $\mathcal{M}, s \models \langle \tau, i, \alpha \rangle \langle \eta, i, \beta \rangle \phi$. Then, for all $t' \in Ran(\mathcal{R}_i^\alpha) \cap P_\alpha^{-1}(end)$, in Game$(\mathcal{M}, s, \alpha)$, $\mathcal{M}, t' \models \langle \eta, i, \beta \rangle \phi$, where $\mathcal{F}_i^\alpha(\tau) = \mathcal{R}_i^\alpha$. This can be possible, only when, for each $t \in P_\gamma^{-1}(end)$, Game$(\mathcal{M}$, t, $\gamma')$ can be defined. Hence, Game$(\mathcal{M}, s, \alpha;\beta)$ is defined, and for all $s' \in Ran(\mathcal{R}_i^\alpha \cup \mathcal{R}_{i_{j_1}}^\beta \cup \ldots \cup \mathcal{R}_{i_{j_l}}^\beta) \cap P_{\alpha;\beta}^{-1}(end)$ in Game$(\mathcal{M}, s, \alpha;\beta)$, $\mathcal{M}, s' \models \phi$. So, $\mathcal{M}, s \models \langle \tau;\eta, i, \alpha;\beta \rangle \phi$.

The validity of the strategy rules follows from the fact that, if there is a path from some state s to a state satisfying some formula ϕ, then the Game$(\mathcal{M}, s, g)$ and a corresponding strategy relation $\mathcal{R}_i^g$ can be defined in such a way that $\langle s, i, g \rangle \phi$ holds at s.

The completeness of the axiom system is proved by showing that every consistent formula is satisfiable. Let α be a consistent formula. Let $Cl(\alpha)$ denote the subformula closure of α, satisfying the FL-closure conditions for the action modalities with the following extra conditions:

(i) If $\langle \sigma \cup \tau, E, \alpha \cup \beta \rangle \phi \in Cl(\alpha)$, then $\langle \sigma, E, \alpha \rangle \phi \vee \langle \tau, E, \beta \rangle \phi \in Cl(\alpha)$.
(ii) If $\langle \sigma \cup \tau, A, \alpha \cup \beta \rangle \phi \in Cl(\alpha)$, then $\langle \sigma, A, \alpha \rangle \phi \wedge \langle \tau, A, \beta \rangle \phi \in Cl(\alpha)$.
(iii) If $\langle \tau;\eta, i, \alpha;\beta \rangle \phi \in Cl(\alpha)$, then $\langle \tau, i, \alpha \rangle \langle \eta, i, \beta \rangle \phi \in Cl(\alpha)$.
(iv) $Cl(\alpha)$ is closed under single negations.

Any maximal consistent subset of $Cl(\alpha)$ is said to be an atom. Let $\mathcal{A}$ denote the set of all such atoms. For $T \in \mathcal{A}$, let $\widehat{T}$ denote the conjunction of all the formulas present in T. For $C, D \in \mathcal{A}$, define $C\mathcal{R}_\pi D$ if $\widehat{C} \wedge \langle \pi \rangle \widehat{D}$ is consistent. The *regular* canonical model $\mathcal{C}$ is defined to be the tuple $\langle \mathcal{A}, \{\mathcal{R}_\pi$: π's are actions$\}$, *ref*, $\mathcal{L}, \mathcal{R}, \mathcal{V} \rangle$, where, *ref*, $\mathcal{L}, \mathcal{R}$ are reflexive relations on $\mathcal{A}$, and $\mathcal{V}(p) = \{T \in \mathcal{A} : p \in T\}$, and $\mathcal{R}_\pi$'s satisfy the *regularity* conditions. The existence lemma for the modalities $\langle \pi \rangle$, can be proved in the usual way, and we have that $\mathcal{C}, A \models \phi$ iff $\phi \in A$, for each $\phi \in Cl(\alpha)$, and each $A \in \mathcal{C}$ where ϕ is either an atomic or a boolean or an action modal formula.

It remains to be shown that $\mathcal{C}, A \models \langle \sigma, i, \gamma \rangle \phi$ iff $\langle \sigma, i, \gamma \rangle \phi \in A$. Because of the reduction axioms, it suffices to show that for each $\langle s, i, g \rangle \phi \in Cl(\alpha)$, and each $A \in \mathcal{C}$, $\mathcal{C}, A \models \langle s, i, g \rangle \phi$ iff $\langle s, i, g \rangle \phi \in A$. In other words, we have to show that $\langle s, i, g \rangle \phi \in A$ iff in Game$(\mathcal{C}, A, g)$, $Ran(\mathcal{F}_i^g(s)) \cap P^{-1}(end)$ is the set of all atoms T, such that $\phi \in T$.

Suppose $\langle s, i, g \rangle \phi \in A$. Then because of axiom (d), Game$(\mathcal{C}, A, g)$, and $\mathcal{F}_i^g$ can be defined in such a way, that the implication holds. The converse follows from the fact that if $< b_{i_1} > \ldots < b_{i_m} > \phi$ is consistent, then so is $< b_{i_1} > \ldots < b_{i_m} > \phi \wedge \langle s, i, g \rangle \phi$, which holds because of the strategy rules. QED

4. DGL and $SDGL$ - a comparison

As mentioned earlier, DGL talks about *generic* games played on game boards, and the meaning of the game modalities is given by existence of strategies. $SDGL$ brings out these strategies to the fore. Strategy combinations for playing composite games are talked about in this framework which brings out the extensional nature of strategies, though according to certain views, strategies are inherently intensional. As mentioned by van Benthem [4, 5], strategies of the players in the game tree can be talked about using the program constructs of the dynamic modal logic. Some proposals for combining strategies to achieve a certain goal are also made there.

The task was to combine the strategy constructs together with the game constructs. $SDGL$ proposes a way to do it. As evident from the previous section, one has to resort to the PDL-style action constructs. To make strategies explicit, one can no longer talk about *generic* games. Extensive game trees come into the treatise - games are defined as tree structures, and strategies are defined as subtrees.

In the tradition of DGL semantics, the so-called *forcing* relations satisfy the conditions of upward-monotonicity and consistency (determinacy also, in case of Parikh's and Pauly's DGL). The sets of states forced by these relations have an inherent 'disjunctive' interpretation. A 'conjunctive' interpretation of these sets which is needed when *parallel* game constructs are introduced, has been taken in [10]. It is interesting to note that, the way *strategies* are defined as relations between states, it corroborates with the 'conjunctive' interpretation of the set of 'end'-states reached. Hence, this language rather suggests 'downward monotonicity' at this conjunctive level.

It is clear that there are some sentences which could be expressed in $SDGL$, but not in DGL. But it is also the case that there are certain statements that can be expressed in DGL, but not in $SDGL$: for example, 'player i does not have any strategy in the game g to achieve ϕ' can be expressed in DGL as $\neg\langle g, i\rangle\phi$. Under these circumstances it would be ideal to have a logic that could express both. This gives rise to the following issue:

Question What would be the complete axiomatization of a logic that has both Parikh's original game modalities as well as the game-strategy modalities presented in the earlier section of this paper?

In fact, for the set of strategy relations $\mathbb{R}_i^\gamma$ for player i in Game($\mathcal{M}, s, \gamma$), one can easily define ρ^i_γ (cf.§2), as follows:

$s\rho^i_\gamma X$ iff $X = Ran(\mathcal{R}^\gamma_i) \cap P_\gamma^{-1}(end)$, for some $\mathcal{R}^\gamma_i \in \mathbb{R}^\gamma_i$.

It remains to be seen what conditions have to be imposed on ρ^i_γ to maintain compatibility. This is precisely the same issue as finding joint logics of proofs and provability in arithmetic, on which a lot of effort has been made in the recent past. For a detailed overview, one can have a look at [2]. The most natural analogy that one can think of having both such existential criterion, as well as the witnesses conforming to it could be found in first order logic - $\exists x\phi$ together with term substitutions like $\phi[\sigma/x]$.

5. Conclusions and intentions

This paper proposes a logic which makes strategies explicit in the dynamic game logic framework. The need for the dynamic modal logic syntax for achieving such targets becomes apparent. An interesting issue of getting a joint logic of complex game modalities together with game-strategy modalities emerges. Some possible areas for future investigations are given below.

Explicit strategies for other logics Several other languages talking about game structures and coalition structures like Alternating-time temporal logic and Coalition logic could be investigated so as to add an explicit notion of strategies, which merely occur as an existential notion in the semantics of these logics. This could very well aid in the social choice mechanism designs.

Adding knowledge and preference notions To come closer to the real game scenario which are played by the rational players, one has to incorporate the knowledge/belief as well as preference modalities in the existing framework, i.e. epistemic versions of these game logics with explicit strategies need to be explored.

Games with imperfect information It is evident that the uniform strategies in the imperfect information games do not conform with the compositional analysis that has been done here. That study is inherently different taking into account the knowledge level of the players, which provides a very interesting challenge.

Acknowledgements

I thank Johan van Benthem for introducing me to the logics of games and strategies and providing me with constant support and invaluable suggestions throughout this effort. I have been greatly enriched with the many discussions I had with Fenrong Liu and Cédric Dégremont during my stay in Amsterdam in the year 2006-2007. R. Ramanujam, Sunil Simon and Fernando R. Velázquez-Quesada went through a preliminary draft of this paper and provided me with thoughtful comments.

References

[1] R. Alur, T. Henzinger, and O. Kupferman. Alternating-time temporal logic. *Lecture Notes in Computer Science*, 1536:23–60, 1998.

[2] S. Artemov and L. Beklemishev. Provability logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic, 2nd ed.*, volume 13. Kluwer, Dordrecht, 2004.

[3] J. v. Benthem. Logic in games. Lecture Notes, Amsterdam and Stanford, 1999.

[4] J. v. Benthem. Games in dynamic-epistemic logic. *Bulletin of Economic Research*, 53(4):219–248, 2001.

[5] J. v. Benthem. Extensive games as process models. *Journal of Logic, Language and Information*, 11:289–313, 2002.

[6] J. v. Benthem. Logic games are complete for game logics. *Studia Logica*, 75:183–203, 2003.

[7] J. v. Benthem. Logic games, from tools to models of interaction. In A. Gupta, R. Parikh, and J. v. Benthem, editors, *Logic at the Crossroads.*, pages 283–317. Allied Publishers, Mumbai, 2007.

[8] J. v. Benthem. Strategizing dgl. working paper, ILLC, Amsterdam, 2007.

[9] J. v. Benthem. In praise of strategies. In J. v. Eijck and R. Verbrugge, editors, *Foundations of Social Softare*, pages 283–317. Studies in Logic, College Publications, to appear.

[10] J. v. Benthem, S. Ghosh, and F. Liu. Modelling simultaneous games with concurrent dynamic logic. In *A Meeting of the Minds, Proceedings of the Workshop on Logic, Rationality and Interaction*, pages 243–258, 2007.

[11] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. CUP, 2001.

[12] J. Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36:89–93, 1950.

[13] R. Parikh. The logic of games and its applications. In *Selected papers of the international conference on "foundations of computation theory" on Topics in the theory of computation*, pages 111–139, New York, NY, USA, 1985. Elsevier North-Holland, Inc.

[14] M. Pauly. *Logics for Social Software*. PhD thesis, University of Amsterdam, 2001.

[15] R. Ramanujam and S. Simon. Axioms for composite strategies. *Proceedings of Logic and Foundations of Games and Decision Theory*, 2006.

[16] A. Rubinstein. Comments of the interpretation of game theory. *Econometrica*, 59(4), 1991.

[17] D. Walther, W. v. d. Hoek, and M. Wooldridge. Alternating-time temporal logic with explicit strategies. *Proceedings of XIth Conference (Theoretical Aspects of Rationality and Knowledge)*, pages 269–278, 2007.

Disambiguation Games in Extended and Strategic Form

Sascia Pavan
s.pavan@tele2.it

Abstract

The aim of this paper is to pursue the line of research initiated by Prashant Parikh which gives content and rigour to the intuitive idea that speaking a language is a rational activity. He employs the most promising tool to that end, namely game theory. I consider one of his examples as a sample case, and the model I build is a slight modification of that developed by him. I argue that my account has some advantage, yet many of the key ideas employed are left unchanged. I analyse this model in detail, describing some of its formal features. I conclude raising a problem that has not been yet, sketching a plausible solution.

1. Introduction

The case I want to analyse concerns sentences like

> Every ten minutes a man gets mugged in New York. (1)

This sentence has two readings, one is that there is a certain man in New York, either very unlucky, or reckless, or masochist, that is mugged every ten minutes. The other reading is that every ten minutes, some man or other, not necessarily the same, gets mugged in New York. Imagine an actual conversation where (1) is uttered, the problem is, how can the hearer decide what is the reading originally intended by the speaker? As for (1), we can hardly imagine a situation where the reading intended by the speaker is the first one – namely the unlucky, reckless, masochist interpretation – and where this is the reading selected by the hearer. A relevant feature of (1) is that one of the two possible readings entails the other, in this case the second reading is a logical consequence of the first. We can think of sentences sharing this same feature with (1), but such that they can be employed in a conversation where the intended reading is the logically stronger one. Consider

> All of my graduate students love a Finnish student in my Game-Theory class. (2)

Suppose that (2) is uttered by a professor in Amsterdam. I do not know how many Finnish students studying game theory there are in Amsterdam. Assume there are very few of them. My intuition is that in most situations the hearer would infer that there is a unique Finnish student in the speaker's class that all graduate students love.

I will use game theory to analyse those conversations that involve sentences that, like (1) and (2), can be interpreted in two different ways, such that one reading is a logical consequence of the other. My starting point will be the account proposed by Prashant Parikh in several works [6, 7, 8].

If modelled in game-theoretic terms,[1] conversations like these involve two players, 1 and 2, where the set of 2's possible moves contains two elements, say A and B, corresponding to two alternative interpretations of some ambiguous sentence ϕ. As is customary in game theory, I will imagine that player 1 is male, and player 2 female. In Parikh's model, player 1 has some private information, unknown to player 2. Parikh defines this basic unknown as the speaker's intended meaning. Player 2 has some beliefs about what this private information is, hence about what message player 1 wants to convey, and these beliefs can be expressed as subjective probabilities. Here lies the main shortcoming of Parikh's model. The hearer's task in a conversation is to guess the speaker's intended meaning, or, better, to select a set of possible alternatives, and assess a probability value for these. Therefore, if we model disambiguation as a game, the probability of the intended meaning should not be one of the primitives of the game, rather we have to explain how the hearer can infer this value from the structure of the game. In other terms, if the task of player 2 is to guess what the intended meaning is, and if she already knows which alternative is more likely to be true, then there is not much to be done anymore, she only needs to multiply the subjective probability of each alternative by the payoffs that the moves available to her would yield in each of these alternatives. Suppose that p is the prior probability that player 2 assigns to the belief that player 1 wants to convey the meaning corresponding to A; and that $1-p$ is the probability of the belief that he wants to convey the meaning

[1] Most of the notation and terminology employed is borrowed from Roger Myerson [4], chapters 2-5.

B. Let g_a be the gain for player 2 if she selects the interpretation A when player 1 really wants to convey A, and let m_a be her gain if she selects A when 1's intended meaning is B. Similarly, let g_b be her gain if she correctly selects B, and m_b her gain when she wrongly selects B. If we describe the situation in this way, her task is very simple, she must select A whenever $p \times g_a + (1-p) \times m_a > p \times m_b + (1-p) \times g_b$ and B whenever $p \times g_a + (1-p) \times m_a < p \times m_b + (1-p) \times g_b$. Once we know that she is able to assign a probability value to the belief that 1's intended meaning is A – no matter how she could accomplish this – there is nothing more to be explained, and hence no more need to appeal to game theory to give an account of her behaviour. But, presumably, we need game theory to explain how she could assess this probability.

This is why I claim that the content of player 1's private information has to be something more basic, and therefore that player 2's prior probabilities have to concern what player 1 actually knows. With this modelling of the game, the speaker's intention to convey a given message can be derived from facts with a minor degree of intentionality, namely his knowledge. To paraphrase Willard Van Quine [9], it reduces the grade of *intentional involvement.* Just consider the questions 'What does player 1 know?' and 'What does player 1 want to say?'. We are not always able to provide definite answers to the questions of the first kind, but, at least, we can assess the probability of the answers, just considering what we know about the player's sources of information. Of course, we can also assess the probability of the answers of the questions of the second kind, but the data to be considered include all those relevant for the first kind, and something else, at least this person's goals. In other words, any reasoning behind an answer to a question of the first kind is conceptually simpler than that required by the second kind.

2. Extensive Form

What is now the shape of our model? If A and B are the only legitimate interpretations of an ambiguous utterance ϕ, then either he believes that A or he believes that B. But in the case we are examining, one of the two readings is a logical consequence of the other, for example we can assume that B logically entails A. If this is true, then if 1 believes that B, he necessarily believes that A. Then, as far as player 2 knows, there are two possibilities:

alternative a: 1 knows that A and it is not the case that he knows that B (either because he knows that not B, or because he does not know whether B);

alternative b: he knows that A and B.

This imposes some restrictions on the payoffs of the game. If a is the real situation, then, if 2 selects A when 1 utters ϕ, she will acquire some new and reliable true knowledge, let us name 'g_a' the value that this outcome has for her. But, if in the same situation she chooses B, instead, she gets a false or at least unreliable new belief and hence some bad result, let us name 'm_b' the value of this outcome. If b is the real situation, then the choice of B will yield some new knowledge, and let be g_b the value she puts on it. But since in this situation the information corresponding to A is true and reliable as well, if she chooses A she does not get some bad payoff, her gain should again be g_a. Let us now use 'p' to refer to the prior probability of situation a, so that $1-p$ is the prior probability of b. Which moves are available to player 1? One of them is of course the uttering of the ambiguous sentence ϕ. But, he could also choose to convey the message he has in mind using some longer but unambiguous sentence, μ_a if he is in situation a, μ_b if he is in situation b. When player 1's choice is one of these two, player 2 does not have to consider alternative interpretations, hence, in game-theoretic terms, she has no opportunity to move. In this case, no misunderstanding is possible.

We have to imagine that he is sincere and honest, that she believes what he says, and that this is common knowledge. For simplicity, imagine also that both of them are interested exclusively in the pure flow of information and no further aims. This is unrealistic, of course, but it is just an idealization not more problematic than the physicist's speculations on frictionless planes. Following Parikh, I will construct my model as a coordination game where the players have the same payoffs, which are determined by the net value of the information minus a 'cost' which is proportional to the length of the sentence. Since the two players have identical payoffs, this is a game of 'pure coordination'. The rationale for this choice is that when honest and rational agents communicate, they all aim at successful communication. Of course there are commonly cases where this is not true, most notably when people lie. But we can legitimately focus attention on those benign cases, especially because the very possibility of lying presupposes the existence of honest communication.

The payoff g_a has to be equal to the value of the true information provided by A, call it v_a, minus the cost c involved by ϕ. If player 1 utters μ_a in situation a, there is no possibility of a misunderstanding, but its cost is higher. Hence this combination yields a value $g'_a = v_a - c'$, where $c' > c$. Similarly, if we call 'v_b' the net value of the true information provided by B, we have that $g_b = v_b - c$. And if player 1 utters μ_b in situation b, then the payoff will be $g'b = vb - c'$, if, for the sake of simplicity, we assume that the cost involved by μ_a and μ_b is analogous. Moreover, since B logically entails A, while A does not entail B, we should have that $v_b > v_a$, and this entails that $g_b > g_a$, and $g'_b > g'_a$.

What is the best choice for player 2? Can we say again that she has only to check whether $p \times g_a + (1-p) \times g_a > p \times m_b + (1-p) \times g_b$, i.e. whether $p > (g_b - g_a)/(g_b - m_b)$, or whether $g_a < p \times m_b + (1-p) \times g_b$? Assume that this is the case, and imagine, for example, that $g_a > p \times m_b + (1-p) \times g_b$. What happens if b, not, a, is the real situation, and 1 wants to convey message B? He would probably utter the longer but unambiguous sentence μ_b. This behaviour is not outright irrational, we shall see that it corresponds to an equilibrium in our model, but it is not always the best outcome that player 1 and player 2 can get, in other words it might be inefficient. Moreover, from the mere fact that there is a probability p that 1 knows that A and not that B nothing follows, from a conceptual point of view, about what he means when he says something. Now she really needs to consider also his goals in order to be able to guess which alternative he wants her to choose.

We can conceive of cases where an unambiguous sentence is so much longer than the corresponding ambiguous one, that a cheap misunderstanding can be preferable to an unambiguous but demanding speech act. We can also imagine situations where the speakers choose ambiguous and potentially misleading messages because they do not want other people to acquire some confidential information. Just think of two spies involved in a telephone conversation, both knowing that their line has been tapped. Sometimes a leak can do more harm than a misunderstanding. I will assume that this is not the case in the conversation we are considering, and that in this case the cost of an utterance is relatively small when compared to the net value of information. Hence, the model employed here requires the following ordering relations: $g_b > g'_b > g_a > g'_a > m_b$, $g'_a - m_b > g_b - g'_b$, $g'_b - g_a > g_a - g'_a$.

But maybe the set of moves available to player 1 is incomplete. Perhaps we should also consider the possibility of uttering μ_a in situation b, and μ_b in situation a. Of course if player 1 uttered μ_a knowing that A is false, he would be lying, and, under the assumption that we are trying to analyse a case of patently honest communication this move would yield a bad outcome for both. But the other case cannot be dismissed so easily, remember that A is true in situation b. The payoff would actually be g'_a. The fact is that whatever the choice of 2, the gain would be higher if player 1 chose μ_b or ϕ. This means that, according to the model presented here, it is never rational for player 1 to choose to utter μ_a in situation b. In technical terms, any strategy where the speaker utters μ_a in situation b or μ_b in situation a is strongly dominated, and can be eliminated from the game. In this case the model simply predicts the existence of a scalar implicature, to the effect that if 1 utters μ_a, then 2 infers that it is not the case that 1 knows that B. Of course the ordering among payoffs that was depicted above presupposes that if 1 knew that B, then he would not conceal this information to the hearer. In situations not covered by this analysis, the speaker could utter μ_a knowing that B, if he did not want 2 to know.

Similarly, we could include a pair of 'don't say anything' moves for player 1. Of course, when he chooses one of these additional moves, she has no possibility to move, and the payoff should be equal to 0 for both players. I will assume that both g'_a and g'_b are strictly positive. If this is the case, then, again, any strategy involving one of these additional moves is strongly dominated, hence I will ignore this possible variant of the game. Yet, this shows that I have not mentioned a fact which is implicitly presupposed by our model, namely the fact that, for example, at node a, player 1 knows that A and also *wants* to convey this information. Maybe this is what is meant by Parikh when he says that the chance nodes 'represent [player 1's] intention to convey' A or B [7].[2] If this is the case, the objections I raised in section 1 miss the mark. But, even in this case, his original models should be modified. As I have already said, with cases like (1) or (2), the payoff obtained when player 2 rightly chooses the stronger interpretation, should be higher than that obtained when she rightly chooses the weaker one.

Now we have all the elements to build our game. I will first construct it as a game of imperfect information in extensive form, which will be called Γ^e. It has the structure of a tree, as is shown in Figure 1.

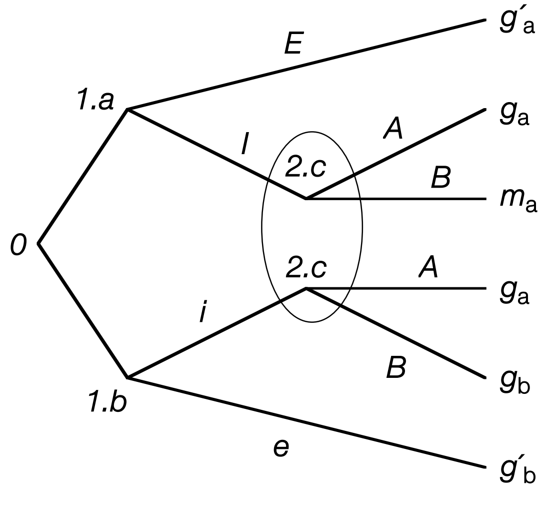


Figure 1. The game in extensive form

[2] P. 27.

This is a deviation from Parikh's path. In the model presented here, I imagine that the first event in the game is a chance move made by 'Nature', which determines whether 1 knows that A and does not know that B, or knows that A and B. At this point 1 can make his move. As usual, I also imagine that the whole structure of the game is common knowledge. Parikh's game in extensive form is not a tree but a pair of trees, since he argues that player 2 cannot construct anything before 1's utterance [7],[3] this is why he proposes the notion of a game of *partial information*. I have the impression that this is an unnecessary – but harmless – deviation from more traditional notions, unless we want our model to mirror the actual mental processes of speaker and hearer. I observe that if Parikh is right in his claim that these disambiguation games should not be treated as ordinary games of imperfect information, the same would hold, for example, for Spence's 'model of education' [12].

Then, a and b are chance events with prior probability p and $1-p$, respectively, where $1 > p > 0$, as usual. If player 1 is in situation a, he can utter either ϕ or μ_a, and we can label these two moves 'I' and 'E', respectively – where 'I' stands for 'implicit' and 'E' for 'explicit'. If he is in situation b, he can choose between ϕ and μ_b, and we can call these alternative moves 'i' and 'e'. Player 2 has a chance to move only if the game is in one of the states labelled '$2.c$', where the identucal labels and the ellipsis manifest the fact that she is not able to distinguish them, technically speaking they belong to the same *information set*. Her options are the two moves A and B.

The fact that there are only two alternative states in 2's information set follows from the characteristic features of the examples considered, namely the fact that one of the two readings is entailed by the other. It is not even necessary that this be a logical entailment – like it is in our example – but the entailment has to be common knowledge. If the two alternative interpretations were mutually exclusive, we would build another game with two epistemic possibilities, but there would be a difference in the ordering of the outcomes. The choice of A when 1 means B, for example, would lead to a bad result. If the two alternatives readings were logically and conceptually unrelated, player 2 would have an information set containing three elements. And of course we can conceive of cases where an ambiguous sentence admits of more than two readings.

3. Nash Equilibria in Strategic Form

The normal representation of our game is the set $\Gamma = \{N, C_1, C_2, u\}$, where $N = \{1, 2\}$ is the set of players, $C_1 = \{Ei, Ee, Ii, Ie\}$ and $C_2 = \{A, B\}$ are the sets of their pure strategies, and u is their payoff function, hence a function from $C_1 \times C_2$ to the real line $\mathbb{R}$. It satisfies the pattern shown in Table 1.

	A	B
Ei	$p \times g'_a + (1-p) \times g_a$	$p \times g'_a + (1-p) \times g_b$
Ee	$p \times g'_a + (1-p) \times g'_b$	$p \times g'_a + (1-p) \times g'_b$
Ii	g_a	$p \times m_b + (1-p) \times g_b$
Ie	$p \times g_a + (1-p) \times g'_b$	$p \times m_b + (1-p) \times g'_b$

Table 1. The game in strategic form

I will now establish a few properties of the model.

Theorem 3.1 *Ii is strongly dominated in Γ*

Proof. Ii is strongly dominated if and only if $\exists \sigma_1 \in \Delta(C_1)$ such that

$$\begin{aligned} &u(Ii, A) < \sigma_1(Ei)u(Ei, A)+ \\ &\sigma_1(Ee)u(Ee, A) + \sigma_1(Ie)u(Ie, A)+ \\ &(1 - \sigma_1(Ei) - \sigma_1(Ee) - \sigma_1(Ie))u(Ii, A) \end{aligned} \quad (3)$$

and

$$\begin{aligned} &u(Ii, B) < \sigma_1(Ei)u(Ei, B)+ \\ &\sigma_1(Ee)u(Ee, B) + \sigma_1(Ie)u(Ie, B)+ \\ &(1 - \sigma_1(Ei) - \sigma_1(Ee) - \sigma_1(Ie))u(Ii, B) \end{aligned} \quad (4)$$

Inequalities (3) and (4) are equivalent to

$$\frac{\sigma_1(Ei) + \sigma_1(Ee)}{\sigma_1(Ie) + \sigma_1(Ee)} < \frac{(1-p)(g'_b - g'_a)}{p(g_a - g'_a)} \quad (5)$$

and

$$\frac{\sigma_1(Ei) + \sigma_1(Ee)}{\sigma_1(Ie) + \sigma_1(Ee)} > \frac{(1-p)(g_b - g'_b)}{p(g'_a - m_b)} \quad (6)$$

respectively. Since $g'_b > g_a$, we have that $(g'_b - g'_a)/(g_a - g'_a) > 1$. Moreover, we stated above that $g'_a - m_b > g_b - g'_b$, therefore $1 > (g_b - g'_b)/(g'_a - m_b)$. This entails

$$\frac{g'_b - g'_a}{g_a - g'_a} > \frac{g_b - g'_b}{g'_a - m_b}$$

and hence

$$\frac{(1-p)(g'_b - g'_a)}{p(g_a - g'_a)} > \frac{(1-p)(g_b - g'_b)}{p(g'_a - m_b)}$$

At this point it is an easy task to find values for $\sigma_1(Ei)$, $\sigma_1(Ie)$, and $\sigma_1(Ee)$ that satisfy inequalities (5) and (6). QED

Observe that Theorem 3.1 entails that no strategy profile τ where $\tau_1(Ii) > 0$ is a Nash equilibrium.

[3] P. 83.

Theorem 3.2 *There is no equilibrium in Γ where both Ei and Ie have strictly positive probability.*

Proof. Assume that σ is such an equilibrium. Then the following inequalities have to be true:

$$\sum_{c_2 \in C_2} \sigma_2(c_2) u(Ei, c_2) \geq \sum_{c_2 \in C_2} \sigma_2(c_2) u(Ee, c_2)$$
$$\sum_{c_2 \in C_2} \sigma_2(c_2) u(Ie, c_2) \geq \sum_{c_2 \in C_2} \sigma_2(c_2) u(Ee, c_2)$$

They are equivalent to

$$\sigma_2(A) \leq \frac{g_b - g'_b}{g_b - g_a}$$
$$\sigma_2(A) \geq \frac{g'_a - m_b}{g_a - m_b}$$

respectively. But this cannot be. In fact, given the ordering among payoffs, $(g'_a - m_b)(g'_b - g_a) > (g_b - g'_b)(g_a - g'_a)$, $(g'_a - m_b)(g_b - g'_b) + (g'_a - m_b)(g'_b - g_a) > (g'_a - m_b)(g_b - g'_b) + (g_b - g'_b)(g_a - g'_a)$, $(g'_a - m_b)(g_b - g_a) > (g_b - g'_b)(g_a - m_b)$, and hence

$$\frac{g'_a - m_b}{g_a - m_b} > \frac{g_b - g'_b}{g_b - g_a} \tag{7}$$

QED

Theorem 3.3 *There is no equilibrium Γ where both Ie and Ee have strictly positive probability.*

Proof. Assume that σ is such an equilibrium. Then the following equation has to be true

$$\sum_{c_2 \in C_2} \sigma_2(c_2) u(Ie, c_2) = \sum_{c_2 \in C_2} \sigma_2(c_2) u(Ee, c_2)$$

which amounts to

$$\sigma_2(A) = \frac{g'_a - m_b}{g_a - m_b}$$

This means that $1 > \sigma_2(A) > 0$, hence in this equilibrium player 2 is indifferent between strategies A and B, and this means

$$\sum_{c_1 \in C_1} \sigma_1(c_1) u(c_1, A) = \sum_{c_1 \in C_1} \sigma_1(c_1) u(c_1, B) \tag{8}$$

Since $\sigma_1(Ei) = 0$ and $\sigma_1(Ii) = 0$, (8) becomes $g_a = m_b$, which is impossible. QED

Theorem 3.4 *There is no equilibrium where both Ei and Ee have strictly positive probability.*

Proof. Analogous to the preceding one. QED

How many equilibria are there? Of course there are two equilibria in pure strategies, namely $\eta = ([Ie], [A])$ and $\theta = ([Ei], [B])$, but there is also an infinite set of mixed equilibria.

Theorem 3.5 *If*

$$\pi_1(Ee) = 1 \tag{9}$$

and

$$\frac{g'_a - m_b}{g_a - m_b} \geq \pi_2(A) \geq \frac{g_b - g'_b}{g_b - g_a} \tag{10}$$

then π is a Nash equilibrium.

Proof. Consider a modified game $\Gamma^* = \{N, C^*_1, C_2, u^*\}$ where $C^*_1 = \{Ei, Ie, Ee\}$, and u^* is just u after its domain has been restricted accordingly. Since Ii is strongly dominated, 3.1, every equilibrium of Γ^* is an equilibrium of Γ, and vice versa. Suppose that π is a strategy profile that satisfies conditions (9) and (10). Define ω as $p(g'_a - g'_b) + g'_b$, which is the expected payoff of both players under π. Since player 2 is clearly indifferent between A and B when player 1's strategy is $[Ee]$, in order to show that π is an equilibrium, we only need to prove the following statements:

$$\omega \geq \sum_{c_2 \in C_2} \pi_2(c_2) u(Ei, c_2) \tag{11}$$
$$\omega \geq \sum_{c_2 \in C_2} \pi_2(c_2) u(Ie, c_2) \tag{12}$$

But the conjunction of conditions (11) and (12) is equivalent to (10). Hence π is a Nash equilibrium of Γ^* and therefore of Γ as well. QED

Theorems 3.1, 3.2, 3.3, and 3.4 entail that there are no other equilibria.

4. Efficiency

Summing up, there are two equilibria in pure strategies, namely η and θ, and many mixed equilibria π. All these mixed equilibria are somehow equivalent, since they yield the same expected payoff, and they all amount to the fact that player 1 goes for the costly but unambiguous option, and player 2 has no opportunity to move. These mixed equilibria are the least efficient ones. As for the equilibria in pure strategies, η is the unique Pareto efficient equilibrium iff

$$p > \frac{g_b - g'_b}{g_b - g'_b + g_a - g'_a}$$

and θ is the unique Pareto efficient equilibrium iff

$$p < \frac{g_b - g'_b}{g_b - g'_b + g_a - g'_a}$$

Parikh's account predicts that the players will tend to converge on the more efficient equilibrium. Robert Van Rooij rejects this solution concept, claiming that it is 'unusual' [10]. This claim is quite odd. On the one hand, there is some agreement among some scholars on the view that we should expect rational players to converge on efficient equilibria in cooperative games [3, 4]. And we should bear in mind that a rational justification of a solution concept is perhaps always desirable, but not strictly necessary, as long as any included profile is (at least) a (Nash) equilibrium and is empirically adequate [1].

Yet, even if I hold that Parikh's relying on Pareto efficiency is probably the most natural choice, I will propose a line of defence which is rejected by him. Imagine that the players were allowed some preplay communication [4], before the beginning of the game, hence before player 1 has access to his private information. Since they are given the opportunity to reach an agreement over the strategy to adopt during the game, they will presumably agree to converge on the equilibrium that is the most profitable one for both, namely on the uniquely Pareto efficient one. Of course an actual occurrence of this kind of communication is unrealistic, but the players do not have to be really engaged in it in order to know what would happen in such a counterfactual situation, because this can be inferred from the structure of the game, it is a feature of the game, which is common knowledge. According to Parikh this argument is untenable for two reasons. First, if you explain successful communication in terms of preplay communication you fall into an infinite regress. Second, 'even if such an infinite regress were avoidable, the solution would certainly require a great deal of effort suggesting that languages aren't quite so efficient as they in fact are' [7].[4] I argue that both of these tenets can be rejected. The model presented here is an account of disambiguation, which is a particular phenomenon occurring in communication. I claimed that our two players could converge on a unique equilibrium, if they considered what would have happened if they had had the opportunity to reach an agreement over a coordinated plan. If this imaginary preplay communication is conceived as involving only unambiguous sentences, there seems to be no danger of an infinite regress, yet the response is the same: they would have agreed to converge on the unique Pareto efficient equilibrium. The second point is less clear to me, since the kind of reasoning that we come to attribute to our players does not seem to involve a great deal of computational effort, compared to the construction of the model itself.

The main shortcoming the Pareto-Nash solution concept that I borrowed from Parikh is that it does not explain what should happen in the limit case where

$$p = \frac{g_b - g'_b}{g_b - g'_b + g_a - g'_a}$$

and therefore both η and θ are (weakly) Pareto efficient. Being uncertain over which course of action should be chosen, our players could end up converging on one of the mixed equilibria. The reasoning is as follows. Resume the argument from preplay communication of the preceding paragraph. The upshot of a counterfactual conversation like the one described would be indeterminate, in this case, they cannot tell what they would have agreed on, just considering the structure of the game. They know for sure that they would have agreed to converge on one of the two weakly efficient equilibria, but they do not know which one, they are equally probable. Therefore, the beliefs that player 1 will end up choosing $[Ei]$ and that he will end up choosing $[Ie]$ are equally probable for player 2

player 2 deems to be equally probable the belief that player 1 will end up choosing $[Ei]$ and the belief that he will end up choosing $[Ie]$. In other words, he comes to believe that he will choose the mixed strategy σ where $\sigma_1(Ei) = \sigma_1(Ie) = \frac{1}{2}$. But this expectation is self-refuting, since, by theorem 3.2, it is not part of a Nash equilibrium. But a similar reasoning would lead 1 to expect that 2 will choose the mixed strategy $\tau_2(A) = \tau_2(B) = \frac{1}{2}$. This is nothing more than the belief that 2 does not know what to do, and hence that she will choose at random. This belief is not self-refuting, since it belongs to one of the mixed equilibria by theorem 3.5, because

$$\frac{g'_a - m_b}{g_a - m_b} > \frac{1}{2} > \frac{g_b - g'_b}{g_b - g_a}$$

And the very same reasoning that can lead 1 to form this belief can lead 2 to believe that 1 has this belief, and so on. This argument is rather unorthodox, and I take it to bee an interim solution to the problem I raised. Another line of reasoning which was promising, at least *a priori*, proved to be a dead end. I will present it anyway in the following two sections, since it gives the occasion to analyze some features of the model which are interesting in themselves, and it will help to meet a possible objection.

5. Trembling Hand Perfect Equilibria

One might hope to select a unique equilibrium arguing that in our analysis player 2 does not exploit all the evidence she has at her disposal, since in order to make a rational choice she must consider not the prior probability of a and b, but the conditional probability of those events, given that player 1 decided to utter ϕ. This suggests that we consider this a sequential game.

[4] P. 39n.

The multiagent representation [4] – also called agent-normal form [11], and agent strategic [5] – is a way to represent games in extensive form as games in strategic form, alternative to the normal representation. In the multiagent representation, there is a player, called (temporary) agent, for every information set of every player. Hence, as far as our game is concerned, player 1 is represented by two agents in the multiagent representation, say a and b. While there is only one agent for player 2, say c.

A *behavioural strategy profile* of a game in extensive form is a mixed strategy profile of its multiagent representation. A generic behavioural strategy profile of our game is $(\sigma_a, \sigma_b, \sigma_c)$, and it specifies a probability distribution for every agent of every player. The behavioural strategy profile $([I], [e], [A])$ corresponds to our Nash equilibrium η in an intuitive way, so that it can be called its *behavioural representation* [4]. Since there should not be any danger of misunderstanding, until the end of this section, I will use the names of the strategy profiles of (the normal representation of) our original game to refer to their behavioural representations. Hence, I will set $\eta = (\eta_a, \eta_b, \eta_c) = ([I], [e], [A])$, and similarly for the other equilibria.

Definition 5.1 A trembling hand perfect equilibrium of a game in extensive form is a trembling hand perfect equilibrium of its multiagent representation [4, 5]. ◁

Searching for trembling hand perfect equilibria of games in extensive form, is precisely a way to search for strategy profiles that are coherent if we consider the sequential nature of a game.

Theorem 5.2 *η is a trembling hand perfect equilibrium of* Γ^e

Proof. Recall that η is a perfect equilibrium iff there exists a sequence $(\eta^k)_{k=1}^{\infty}$ such that each η^k is a perturbed behavioural strategy profile where every move gets positive probability, and, moreover

(i)

$$\lim_{k\to\infty} \eta_s^k(d_s) = \eta_s(d_s) \quad \forall s \in S \quad \forall d_s \in D_s$$

(ii)

$$\eta_s \in \text{argmax}_{\tau_s\in\Delta(D_s)} \sum_{d\in D}\left(\prod_{r\in N-s}\eta_r^k(d_r)\right)\tau_s(d_s)u(d) \quad \forall s \in S$$

where $S = (a, b, c)$ is the set of all information states of all players, and, for each $s \in S$, D_s is the set of moves available to the relevant player in state s, and $D = \times_{s\in S} D_s$. It is not difficult to find a sequence satisfying these criteria. Set

$$\xi = \frac{(1-p)(g_b - g_a)}{p(g_a - m_b)}$$

Then $\forall k \in (1, 2, 3, ...)$, if $\xi \geq 1$,

$$\eta_a^k(I) = \frac{2k-1}{2k} \quad \eta_b^k(i) = \frac{1}{2k\xi} \quad \eta_c^k(A) = 1 - \frac{g_a - g'_a}{k(g_a - m_b)}$$

If $\xi < 1$, instead, set

$$\eta_b^k(i) = \frac{1}{2k}$$

and the rest as before. You can see at a glance that these sequences satisfy condition (i). Consider now the expected payoff for player 1 when he is in state a and is planning to make move $\tau_a \in \Delta(D_a)$, and all moves at all other states are made according to scenario η^k. It is equal to

$$\sum_{d-a\in D-a}\left(\prod_{r\in N-a}\eta_r^k(d_r)\right)\times \left[\tau_a(I)u(d_{-a}, I) + (1-\tau_a(I))u(d_{-a}, E)\right] \tag{13}$$

We can consider (13) as a function of $\tau_a(I)$, and if we calculate the derivative of this function we get

$$p[\eta_c^k(A)(g_a - m_b) + m_b - g'_a]$$

As you can easily verify, this value is either null or positive for all k, and this means that, since $\eta_a(I) = 1$

$$\eta_a \in \text{argmax}_{\tau_a\in\Delta(D_a)} \sum_{d\in D}\left(\prod_{r\in N-a}\eta_r^k(d_r)\right)\tau_a(d_a)u(d)$$

Similarly, if you consider the corresponding expected payoff for player 1 when he is in state b, i.e.

$$\sum_{d-b\in D-b}\left(\prod_{r\in N-b}\eta_r^k(d_r)\right)\times \left[\tau_b(i)u(d_{-b}, i) + (1-\tau_b(i))u(d_{-b}, e)\right]$$

regard it as a function of $\tau_b(i)$, and calculate its derivative, you get

$$(1-p)[\eta_c^k(A)(g_a - g_b) + g_b - g'_b]$$

which is either null or negative for all k, because of inequality (7), and this means that, since $\eta_b(i) = 0$,

$$\eta_b \in \text{argmax}_{\tau_b\in\Delta(D_b)} \sum_{d\in D}\left(\prod_{r\in N-b}\eta_r^k(d_r)\right)\tau_b(d_b)u(d)$$

Finally, if you calculate the expected payoff for player 2, you have

$$\sum_{d-c \in D-c} \left(\prod_{r \in N-c} \eta_r^k(d_r) \right) \times$$
$$[\tau_c(A)u(d_{-c}, A) + (1 - \tau_c(A))u(d_{-c}, B)]$$

whose derivative is

$$\eta_a^k(I)p(g_a - m_b) + \eta_b^k(i)(1-p)(g_a - g_b)$$

which is either null or positive for all k, and this entails

$$\eta_c \in \text{argmax}_{\tau_c \in \Delta(D_c)}$$
$$\sum_{d \in D} \left(\prod_{r \in N-c} \eta_r^k(d_r) \right) \tau_c(d_c)u(d)$$

QED

The case of θ is completely analogous.

Theorem 5.3 *θ is a trembling hand perfect equilibrium of Γ^e*

Proof. A suitable sequence is

$$\theta_a^k(I) = \frac{1}{2k} \quad \theta_b^k(i) = \frac{2k-1}{2k} \quad \theta_c^k(A) = \frac{g_b - g_b'}{k(g_b - g_a)}$$

if $\xi \geq 1$, and

$$\theta_a^k(I) = \frac{\xi}{2k} \quad \theta_b^k(i) = \frac{2k-1}{2k} \quad \theta_c^k(A) = \frac{g_b - g_b'}{k(g_b - g_a)}$$

if $\xi < 1$. QED

As for the mixed equilibria the case is simpler.

Theorem 5.4 *The mixed equilibria π are trembling hand perfect in the extensive form of the game*

Proof. Since $1 > \pi > 0$, we can set $\pi_c^k(A) = \pi_c(A)$, and

$$\eta_a^k(I) = \frac{1}{2k} \quad \eta_b^k(i) = \frac{1}{2k\xi}$$

whenever $\xi \geq 1$, and

$$\eta_a^k(I) = \frac{\xi}{2k} \quad \eta_b^k(i) = \frac{1}{2k}$$

otherwise. QED

Theorems 5.2, 5.3, and 5.4 show that all of the Nash equilibria can be legitimately regarded as tenable, even if take into account the fact that the players do not act simultaneously, and hence that player 2 must update her beliefs upon the evidence that player 1 chose the ambiguous utterance ϕ. This is indeed trivial for the pure equilibria, but not for the mixed ones. The question is, what does 2 believe when she sees that 1 has uttered ϕ? But an answer to this question is relative to a strategy profile. For example, under η, if 1 utters ϕ, she knows for sure that she is in her upper node, since the conditional probability of this event, upon the evidence that he has chosen either I or i, is equal to 1. But what does she believe in the same situation under one of the mixed equilibria, where all the nodes in her information set have null prior probability, and hence the traditional Bayesian theory leaves the corresponding conditional probability undefined? Many of the refinements of the Nash equilibrium concept are an attempt to give an answer to this question. The reasoning behind the notion of trembling hand perfect equilibrium in extensive games is this: when a player comes to know that an event with null prior probability has actually occurred, she believes that this was due to a mistake made by one of the other players, when performing his strategy. Then she updates her beliefs, assuming that all the possible mistakes had an infinitesimal prior probability.

6. Proper Equilibria

Summing up, theorems 5.2, 5.3, and 5.4 strengthen the overall strategy of this paper, which amounts to the claim that the players will converge on the unique Pareto efficient equilibrium, whenever there is one, and on one of the mixed equilibria otherwise. For completeness, I will mention another feature of the model which is less reassuring. The concept of proper equilibrium builds on the idea behind trembling hand perfect equilibrium, and adds the restriction that less costly mistakes have a higher probability than more dangerous ones [4].

Definition 6.1 A mixed strategy profile σ is an **ε-proper equilibrium** iff all pure strategies get strictly positive probability and, for every player i and any pair of pure strategies c_i and e_i in C_i,

$$\text{if} \quad u_i(\sigma_{-i}, [c_i]) < u_i(\sigma_{-i}, [e_i]), \quad \text{then} \quad \sigma_i(c_i) \leq \varepsilon \sigma_i(e_i)$$

◁

Definition 6.2 A mixed strategy profile σ is a **proper equilibrium** iff there is a sequence $(\varepsilon^k, \sigma^k)_{k=1}^{\infty}$ such that

$$\lim_{k \to \infty} \varepsilon^k = 0, \quad \lim_{k \to \infty} \sigma_i^k(c_i) = \sigma_i(c_i), \quad \forall i \in N, \forall c_i \in C_i,$$

and, for every k, σ^k is an ε^k-proper equilibrium. ◁

Proper equilibria are usually applied to the normal representation of games in extensive form. It is probably evident that η and θ are proper equilibria, the interesting case is the following one.

Theorem 6.3 *A mixed strategy profile* π *such that* $\pi_1(Ee) = 1$ *is a proper equilibrium of* Γ *iff*

$$\pi_2(A) = \frac{p(g'_a - m_b + g'_b - g_b) + g_b - g'_b}{p(g_a - m_b + g_a - g_b) + g_b - g_a}$$

I will not provide the proof here, I only remark that for every game, there are three and only three equilibria, and that the suitable value for $\pi_2(A)$ is always included in the open interval

$$\left(\frac{g_b - g'_b}{g_b - g_a}, \frac{g'_a - m_b}{g_a - m_b}\right)$$

I admit that this fact is not welcome, since it weakens the strategy adopted so far. I just take it as a reason for not adopting proper equilibrium as a solution concept in disambiguation games.

7. Conclusion

Summing up, the substance of this work is a new game-theoretic analysis of the capacity humans have to communicate using ambiguous expressions. The background hypothesis is that these tasks are accomplished because humans are rational creatures, and, when two people are involved in a conversation, they crucially capitalize on this fact, assuming that it is common knowledge. I built on ideas first developed by Prashant Parikh, raising some objections that led me to modify his models.

I built a game of imperfect information in extensive form, where a hearer and a speaker are the two players, the speaker has some private information, and his task is to convey this piece of information to the hearer. Here lies the main difference between my analysis and Parikh's, since, in his model, the relevant private information of the speaker is the intended meaning of his speech act. I argued that my reform renders the theory more natural and conceptually simpler.

The examples I chose as sample cases were simpler to analyze than more general cases, because of the structural features of the resulting model. In the end I retain Parikh's conclusion that speakers tend to focus on efficient equilibria, but I also proposed a solution to a problem that had been left open, namely, the strategy adopted by the speakers when there is not a unique efficient equilibrium. I argued that, in this case, the speaker goes for the ambiguous expression, which is costly, but safe. The argument I used to back both of these tenets hinges on the idea that the players are able to guess the joint strategy they would agree on, were they allowed some preplay communication before the beginning of the game. This kind of argument is not new. It is crucial that the players do not really need to entertain this kind of communication in order to know what would ensue from it. Yet, I acknowledged that my argument is partially unorthodox, from the point of view of the existing literature.

I also showed that the relevant equilibria are plausible even if we consider that a conversation is sequential in nature, proving that they are trembling hand perfect. And I ended stating, omitting the proof, that not all the equilibria are proper, which I take to be an unwelcome result.

Now the task is to extend this analysis to other, more general and more complex cases, and check whether the claims that have been put forward here have a wider application.

References

[1] N. Allott. Game theory and communication. In A. Benz, editor, *Game Theory and Pragmatics*. Macmillan, Palgrave, 2006.

[2] J. Harsanyi and R. Selten. *A General Theory of Equilibrium Selection in Games*. MIT Press, Cambridge Massachusetts, 1988.

[3] R. Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, Cambridge Massachusetts, 1991.

[4] M. Osborne and A. Rubinstein. *A Course in Game Theory*. MIT Press, Cambridge Massachusetts, 1994.

[5] P. Parikh. A game-theoretic account of implicature. In Y. Moses, editor, *Theoretical Aspects of Reasoning about Knowledge*, 1992.

[6] P. Parikh. *The Use of Language*. CSLI, Stanford CA, 2001.

[7] P. Parikh. Radical semantics: A new theory of meaning. *Journal of Philosophical Logic*, 35:349–391, 2006.

[8] W. Quine. *The Ways of Paradox and other Essays*. Harvard University Press, Cambridge Massachusetts, 2 edition, 1976.

[9] R. V. Rooij. Signalling games select horn strategies. *Linguistics and Philosophy*, 27:493–527, 2004.

[10] R. Selten. Reexamination of the perfectness concepts for equilibrium points in extensive games. *International Journal of Game Theory*, 4:25–55, 1975.

[11] M. Spence. Job market signaling. *The Quarterly Journal of Economics*, 87:355–374, 1975.

Intentions and transformations of strategic games[*]

Olivier Roy
University of Groningen
O.Roy@rug.nl

Abstract

In this paper we take a game-theoretic perspective to study the effects of previously adopted intentions in rational decision making. We investigate the question of how agents transform the decision problems they face in the light of what they intend, and provide conditions under which such transformations, when iterated, leave room for deliberation, i.e. do not exclude all the options of the decision maker.

1 Introduction

In this paper we take a game-theoretic perspective to study the effects of previously adopted intentions in rational decision making. We investigate the question of how agents transform the decision problems they face in the light of what they intend, and provide conditions under which such transformations, when iterated, leave room for deliberation, i.e. do not exclude all the options of the decision maker.

There is a broad consensus among philosophers of action, e.g. [Bratman, 1987] and Velleman [2003], that previously adopted intentions, alongside beliefs and desires, shape decision problems. This role, however, has until now attracted little if no attention in game theory. The present work attempts at (partially) filling this gap. This not only results in a richer game-theoretic framework, but also sheds new lights on the philosophical theory of intentions, especially concerning the interactive character of intention-based transformations of decision problems.

The approach in this paper differs in important respects from the one in "BDI" architectures, e.g. Georgeff et al. [1998] and van der Hoek et al. [2007]. Studies in that paradigm have mainly focused on the relation that intentions can or should have with beliefs and desires, and on different policies of intention revision. Furthermore, these approaches do not directly use game-theoretic formalisms, but rather frameworks tailored for the analysis of multi-agent systems. Here we use strategic form games, and focus on how intentions transform them.

We consider two ways of transforming decision problems on the basis of the agents' intentions. For each of them we characterize the conditions under which they do not remove all possible choices for the agents. Proofs of the technical results can be found in the Appendix.

2 Strategic games with intentions

We use standard strategic form games, as in e.g. [Osborne and Rubinstein, 1994], except that preferences are represented qualitatively. A *decision problem* or *strategic game* $\mathbb{G}$ is a tuple $\langle I, S_i, X, \pi, \leq_i \rangle$ such that :

- I is a finite set of agents.
- S_i is a finite set of *actions* or *strategies* for i. A *strategy profile* $\sigma \in \Pi_{i \in I} S_i$ is a vector of strategies, one for each agent in I. The strategy s_i which i plays in the profile σ is noted $\sigma(i)$.
- X is a finite set of *outcomes*.
- $\pi : \Pi_{i \in I} S_i \to X$ is an *outcome function* that assigns to every strategy profile $\sigma \in \Pi_{i \in I} S_i$ an outcome $x \in X$. We use $\pi(s_i)$ to denote the set of outcomes that can result from the choice of s_i. Formally: $\pi(s_i) = \{x : x = \pi(s_i, \sigma_{j \neq i})$ for some $\sigma_{j \neq i} \in \Pi_{j \neq i} S_j\}$.
- $\leq_i$ is a reflexive, transitive and total preference relation on X.

We study the effect of *previously adopted* intentions on such decision problems, rather than the process by which the agents form these intentions. Furthermore, we restrict our attention to intentions to *realize certain outcomes* in the game, in contrast with intentions to play certain strategy—although there is an obvious connection between the two. We thus assign to each agent $i \in I$ an *intention set* $\iota_i \subseteq X$. The intention set ι_i of agent i should be thought as the intentions that i has formed some time before entering the

[*] This is an extended abstract of [Roy, 2008, chap.4]. Some results presented here have been obtained in collaboration with Martin van Hees (Rijksuniversiteit Groningen).

game, and on the basis of which he now has to make his decision. Following common assumption by philosophers of action we suppose that $\iota_i \neq \emptyset$, which amounts to agents not having inconsistent intentions. An *intention profile* ι is a vector of intention sets, one for each agent.

Many philosophers of action have stressed that previously adopted intentions *transform* decision problems, a phenomenon which is called the *reasoning-centered commitment* of intentions. They imposes a "standard for *relevance* for options considered in deliberation. And they constrain solutions to these problems, providing a *filter of admissibility* for options." [Bratman, 1987, p.33, emphasis in the original]. These are the two effects of intentions on deliberation that we study in the next sections.

3 Filter of admissibility

We take "providing a filter of admissibility" to mean ruling out options that are incompatible with the agents achieving their intentions. Agents, in that sense, discard some of their strategies because they are incompatible with what they intend. We will study two different admissibility/compatibility criteria, depending on whether the agents take each others' intentions into account.

We start with a generic definition of the discarding process, which we call *cleaning*, and in which the two notions of admissibility will be plugged in. The *cleaned version* $cl(S_i)$ of a strategy set S_i is defined as:

$$cl(S_i) = \{s_i \mid s_i \text{ is admissible for deliberation for } i\}$$

The *cleaned* version of a game $\mathbb{G}$ with intention profile ι is the tuple $cl(\mathbb{G}) = \langle I, X^{cl}, \{cl(S_i), \leq_i^{cl}\}_{i \in I}, \pi^{cl} \rangle$ such that:

- $X^{cl} = \pi(\Pi_{i \in I} cl(S_i)) = \{x \mid x = \pi(\sigma) \text{ for some } \sigma \in \Pi_{i \in I} cl(S_i)\}$.
- $\leq_i^{cl}$ is the restriction of $\leq_i$ to X^{cl}.
- π^{cl} is π with the restricted domain $\Pi_{i \in I} cl(S_i)$.

We do not study intention revision, and so we assume that the agents adapt their intentions to the decision problem they face after cleaning by giving up on achieving the outcomes that are no longer achievable. We thus take the cleaned version ι_i^{cl} of the intention set ι_i to be $\iota_i \cap X^{cl}$, reminding plain belief expansion in e.g. Rott [2001] and Gärdenfors [2003].

The first criterion for admissibility we consider is individualistic: we say that a strategy s_i of agent i is *individualistically admissible* for him when choosing it can yield an outcome he intends. Formally, a strategy s_i of agent i is individualistically admissible with respect to his intention set ι_i when $\pi(s_i) \cap \iota_i \neq \emptyset$. Conversely, a strategy is not admissible for i when choosing it would not realize any of his intentions.

It can be that no strategy survive cleaning with individualistic admissibility, simply because some outcome x can be unrealizable, i.e. it can happen that there is no profile σ such that $\pi(\sigma) = x$. In such case we say that cleaning *empties* a decision problem for the agent. Intuitively agents should avoid intentions which, once used for cleaning, empty the decision problem. This this leaves them no strategy to choose. It is thus important to characterize the intention sets that do not lead to empty cleaned games.

When there is only one agent, cleaning empties a decision problem if and only if ι_i contains no realizable outcomes. In interactive situations, however, agents who clean individualistically can make intentions of others unrealizable. Table 1 is an example, with the numbers in the cells representing which outcomes are in ι_i for the corresponding agent, 1 being the row and 2 being the column player. When more than one agent is involved, to have realizable intentions is thus not enough to avoid ending up with empty strategy sets after cleaning. To pinpoint the conditions which ensure such non-emptiness in the general case, we look at iteration of cleaning, in a way that draws from van Benthem [2003] and Apt [2007].

$\mathbb{G}$	t_1	t_2
s_1	1	
s_2	2	

$cl(\mathbb{G})$	t_1
s_1	1

Table 1. A game which an empty cleaning.

Given a strategic game $\mathbb{G}$, let $cl^k(\mathbb{G}) = \langle I, X^{cl^k}, \{cl^k(S_i), \leq_i^{cl^k}\}_{i \in I}, \pi^{cl^k} \rangle$ be the strategic game that results after k iterations of the cleaning of $\mathbb{G}$. That is, $cl^1(\mathbb{G}) = cl(\mathbb{G})$ and $cl^{k+1}(\mathbb{G}) = cl(cl^k(\mathbb{G}))$. The smallest cleaning *fixed-point* $cl^{\#}(\mathbb{G})$ of $\mathbb{G}$ is defined as $cl^k(\mathbb{G})$ for the smallest k such that $cl^k(\mathbb{G}) = cl^{k+1}(\mathbb{G})$. In what follows we ignore the "smallest" and only write about the fixed point.

Every game has a unique cleaning fixed point with individualistic cleaning but, as just noted, it may be an empty one. This is avoided only if the intentions of the agents are sufficiently entangled with one another.

Let us call the *cleaning core* of a strategic game $\mathbb{G}$ is the set of strategy profile S^* inductively defined as follows, with $\pi^{S^n}(s_i) = \pi(s_i) \cap \{\pi(\sigma') : \sigma' \in S^n\}$.

- $S^0 = \Pi_{i \in I} S_i$.
- $S^{n+1} = S^n - \{\sigma : \text{ there is an } i \text{ such that } \pi^{S^n}(\sigma(i)) \cap \iota_i = \emptyset\}$.

- $S^* = \bigcap_{n<\omega} S^n$.

For each strategy s_i and profile σ in the cleaning core such that $\sigma(i) = s_i$, there is at least one agent j for whom strategy $\sigma(j)$ is admissible, by looking only at what can result from the profiles in the core.

Fact 3.1 *For any strategic game $\mathbb{G}$ and intention profile ι, $S^* \neq \emptyset$ iff $cl^{\#}(\mathbb{G})$ is not empty.*

From this we learn that the individualistic character of admissibility must be compensated by an interlocking web of intentions and strategies if cleaning is not to make the game empty. Intentions which yield a non-empty cleaning core closely fit the admissible strategies of *all* agents. By intending outcomes that are realizable in the cleaning core, an agent somehow acknowledges that he interacts with other agents who, like him, clean inadmissible options from their strategy set.

The following alternative form of admissibility emphasizes this interactive character. A strategy s_i of agent i is *altruistically admissible* with respect to his intention set ι_i when there is a $j \in I$ such that $\pi(s_i) \cap \iota_j \neq \emptyset$. Following this second criterion, a strategy of agent i is admissible whenever it can yield an outcome that some agent, *not necessarily* i, intends. When agents discard option on the basis of this criterion, there is no risk of emptying the game, and the process does not need to be iterated.

Fact 3.2 *For $\mathbb{G}$ an arbitrary strategic game, $cl^{\#}(\mathbb{G}) = cl(\mathbb{G})$ for cleaning with altruistic admissibility.*

Fact 3.3 *For any strategic game $\mathbb{G}$, intention profile ι and cleaning with altruistic admissibility, there is, for all i, a realizable $x \in \iota_i$ iff $cl^{\#}(\mathbb{G})$ is not empty.*

It is thus crucial for agents to take the others' intentions into account when ruling out options in strategic games. If, on the one hand, agents rule out options without taking care of what the others intend, they run the risk of ending up with no strategy at all, unless their intentions are already attuned to those of their co-players. If, on the other hand, their intentions do not fit so well with those of others, then they should at least take heed of what the others intend when ruling out options. This aspect of the reason-centered commitment of intentions has, up to now, been overlooked in philosophical theories of intentions.

4 Standard of relevance

We now turn to the second aspect of the reason-centered commitment of intentions: transformations of decision problem based on the "standard of relevance".

Here we take this idea to mean discarding options which differences are not relevant in terms of what one intends. We say that such options are *redundant*. Formally, two strategies s_1 and s_2 in S_i are *redundant*, noted $s_1 \approx s_2$, whenever $\pi(s_1, \sigma_{j\neq i}) \in \iota_i$ iff $\pi(s_2, \sigma_{j\neq i}) \in \iota_i$ for all combinations of actions of other agents $\sigma_{j\neq i} \in \Pi_{j\neq i} S_j$. Strategies s_1 and s_2 in Table 2 are redundant for the row player in that sense.

	t_1	t_2	t_3
s_1	$1, 2$	2	1
s_2	1	2	1
s_3		1	2

Table 2. A game with redundant strategies for the row player.

The relation $\approx$ clearly induces a partition of the set of strategies S_i into subsets $[s_i]^{\mathbb{G}}_{\approx} = \{s'_i \in S_i | s'_i \approx s_i\}$, each of which represents a distinct "means" for agent i to achieve what he intends. We take the standard of relevance imposed by intentions to induce such a means-oriented perspective on decision problems.

To make a decision from that perspective agents have to sort out these means according to some preference ordering. Here we assume that they "pick" a representative strategy for each means, and collect them to form their new strategy set. This allows to define preferences in the game that result from this transformation from those in the original game. Regarding the picking process itself, we take an abstract point of view and leave implicit the criterion which underlies it.

Given a strategic game $\mathbb{G}$, a function $\theta_i : \mathcal{P}(S_i) \to S_i$ such that $\theta_i(S) \in S$ for all $S \subseteq S_i$ is called i's *picking function*. A *profile* of picking functions Θ is a combination of such θ_i, one for each agent $i \in I$. These functions return, for each set of strategies—and in particular each equivalence class $[s_i]_{\approx}$—the strategy that the agents picks in that set. We define them over the whole power set of strategies to facilitate the technical analysis.

The *pruned version* $pr(S_i)$ of a strategy set S_i, with respect to an intention set ι_i and a picking function θ_i is defined as:

$$pr(S_i) = \{\theta([s_i]^{\mathbb{G}}_{\approx}) : s_i \in S_i\}$$

Pruned version of a strategic game $\mathbb{G}$ are defined similarly as cleaned ones: given an intention profile ι and a profile of picking function Θ, the pruned version of $\mathbb{G}$ is the tuple $pr(\mathbb{G}) = \langle I, X^{pr}, \{pr(S_i), \leq_i^{pr}\}_{i\in I}, \pi^{pr}\rangle$ such that:

- $X^{pr} = \pi(\Pi_{i\in I} pr(S_i))$.

- $\leq_i^{pr}$ is the restriction of $\leq_i$ to X^{pr}.

- π^{pr} is π with the restricted domain $\Pi_{i \in I} pr(S_i)$.

The pruned version ι_i^{pr} of an intention set ι_i is $\iota_i \cap X^{pr}$. Agents, again, adapt their intentions in the process of pruning.

We once again take a general point of view and analyze iterations of pruning. Given a strategic game $\mathbb{G}$, let $pr^k(\mathbb{G})$ be the strategic game that results after k iterations of the pruning of $\mathbb{G}$. That is, $pr^0(\mathbb{G}) = \mathbb{G}$ and $pr^{k+1}(\mathbb{G}) = pr(pr^k(\mathbb{G}))$. The *pruning fixed point* $pr^{\#}(\mathbb{G})$ of $\mathbb{G}$ is defined as $pr^k(\mathbb{G})$ for the smallest k such that $pr^k(\mathbb{G}) = pr^{k+1}(\mathbb{G})$.

As for cleaning, it can happen that agents end up with empty intentions after a few rounds of pruning, but no pruning makes a game empty.

Fact 4.1 *For all strategic game $\mathbb{G}$ and agent $i \in I$, $pr^{\#}(S_i) \neq \emptyset$.*

Furthermore, the existence of pruning fixed points where all agents have non-empty intentions depends on whether they intend "safe" outcomes. Given a strategic game $\mathbb{G}$, an intention profile ι and a profile of picking functions Θ, the outcome $x = \pi(\sigma)$ is:

- *Safe for pruning at stage 1* iff for all agents i, $\theta_i([\sigma(i)]) = \sigma(i)$.

- *Safe for pruning at stage* $n + 1$ whenever it is safe for pruning at stage n and for all agents i, $\theta_i([\sigma(i)]^{pk^n(\mathbb{G})}) = \sigma(i)$.

- *Safe for pruning* when it is safe for pruning at all stages n.

Safe outcomes are those which the picking functions retain, whatever happens in the process of pruning. Intending safe outcomes is necessary and sufficient for an agent to keep his intention set non-empty in the process of pruning.

Fact 4.2 *For any strategic game $\mathbb{G}$, intention profile ι, profile of picking function Θ and for all $i \in I$, $\iota_i^{pr^{\#}} \neq \emptyset$ iff there is a $\pi(\sigma) \in \iota_i$ safe for pruning in $\mathbb{G}$.*

Agents are thus required to take the others' intentions *and* picking criteria into account if they wish to avoid ending up with empty intentions after pruning. In single-agent cases pruning never makes the intention set of the agent empty, as long as the agent has realizable intentions. This shows, once again, that reasoning-centered commitment really gains an interactive character in situations of strategic interaction.

5 Putting the two transformations together

We now look at how the pruning and cleaning interact with one another, in order to get a more general picture of the reasoning-centered commitment of intentions. We investigate sequential applications of these operations, and consider individualistic admissibility only.

Given a strategic game $\mathbb{G}$, let $t(\mathbb{G})$ be either $pr(\mathbb{G})$ or $cl(\mathbb{G})$. A *sequence of transformation of length* k is any $t^k(\mathbb{G})$ for $k \geq 0$, where $t^1(\mathbb{G}) = t(\mathbb{G})$ and $t^{k+1}(\mathbb{G}) = t(t^k(\mathbb{G}))$. A sequence of transformation $t^k(\mathbb{G})$ is a *transformation fixed point* whenever both $cl(t^k(\mathbb{G})) = t^k(\mathbb{G})$ and $pr(t^k(\mathbb{G})) = t^k(\mathbb{G})$.

The first notable fact about cleaning and pruning sequences is that these operations do not in general commute. Table 3 is a counterexample, with $\theta_2([t_1]) = t_1$. They do commute, however, in the single-agent case.

$\mathbb{G}$	t_1	t_2
s_1		1
s_2	1, 2	1, 2

$pr(\mathbb{G})$	t_1
s_1	
s_2	1, 2

$cl(pr(\mathbb{G}))$	t_1
s_2	1, 2

Table 3. Counter-example to commutativity.

Fact 5.1 $pr(cl(\mathbb{G})) = cl(pr(\mathbb{G}))$ *for any strategic game $\mathbb{G}$ with only one agent, intention set ι_i and picking function θ_i.*

Sequential cleaning and pruning creates new possibilities for empty fixed points. Neither the existence of a cleaning core nor of safe outcomes, and not even a combination of the two criteria are sufficient to ensure non-emptiness. Furthermore, there might not be a unique fixed point, as revealed in Tables 4, 5 and 6, with $\theta_1(\{s_1, s_2\}) = s_2$, $\theta_1(\{s_1, s_2, s_3\}) = s_1$ and $\theta_1(\{s_2, s_3\}) = s_2$.

$\mathbb{G}$	t_1	t_2	t_3
s_1	1	2	
s_2	1, 2		
s_3	1		1

Table 4. A game with two different fixed-points.

Ignoring redundant transformations, all sequences of cleaning and pruning reach a fixed point in a finite number of steps, for every finite strategic games. Non-emptiness of this fixed point is ensured by the following strengthening of safety for pruning and cleaning core. The outcome x of profile $\sigma \in \Pi_{i \in I} S_i$ is:

$cl(\mathbb{G})$	t_1	t_2
s_1	1	2
s_2	1, 2	
s_3	1	

$pr(cl(\mathbb{G}))$	t_1	t_2
s_1	1	2

$cl(pr(cl(\mathbb{G})))$	t_2
s_1	2

Table 5. The route to the first (empty) fixed point of the game in Table 4.

$pr(\mathbb{G})$	t_1	t_2	t_3
s_2	1, 2		
s_3	1		1

$cl(pr(\mathbb{G}))$	t_1
s_2	1, 2
s_3	1

$pr(cl(pr(\mathbb{G})))$	t_2
s_2	1, 2

Table 6. The second fixed point of the game in Table 4.

- *Safe for iterated transformations at stage 1* whenever, for all $i \in I$:
  1. $\pi(\sigma(i)) \cap \iota_i \neq \emptyset$.
  2. $\theta_i[\sigma(i)]^{\mathbb{G}}_{\approx} = \sigma(i)$.
- *Safe for iterated transformations at stage n + 1* whenever it is safe for iterated transformation at stage n and for all $i \in I$:
  1. $\pi^{t^n(\mathbb{G})}(\sigma(i)) \cap \iota_i^{t^n(\mathbb{G})} \neq \emptyset$.
  2. $\theta_i[\sigma(i)]^{t^n(\mathbb{G})}_{\approx} = \sigma(i)$.
- *Safe for iterated transformations* whenever it is safe for transformation at all n.

Fact 5.2 *For any strategic game $\mathbb{G}$, intention profile ι and profile of consistent picking function Θ, if $\pi(\sigma)$ is safe for transformation in $\mathbb{G}$ then for all fixed points $t^{\#}(\mathbb{G})$, $\sigma \in \Pi_i t^{\#}(S_i)$.*

The presence of safe outcomes is thus sufficient not only to ensure that a game has no empty fixed point, but also that all fixed points have a non-empty intersection. Precisely because of that, this does not entail that any game which has no empty fixed point contains safe outcomes. If it can be shown that whenever a game has a non-empty fixed-point then this fixed-point is unique, we would know that safety for transformation exactly captures non-emptiness. Whether this is the case is still open to us at the moment. We do know, however, that the converse of Fact 5.2 holds if we constraint the picking functions.

In the spirit of Sen's [1970] "property α" , let a picking function θ_i be called *consistent* if $\theta_i(X) = s_i$ whenever $\theta_i(Y) = s_i$, $X \subseteq Y$ and $s_i \in X$.

Fact 5.3 *For any strategic game $\mathbb{G}$, intention profile ι and profile of consistent picking function Θ, if $\sigma \in \Pi_i t^{\#}(S_i)$ for all fixed points $t^{\#}(\mathbb{G})$, then $\pi(\sigma)$ is safe for transformation in $\mathbb{G}$.*

If all players intend safe outcomes we thus know that all fixed-point are non-empty, and we can "track" safe outcomes in the agents' original intentions by looking at those they keep intending in all fixed-points.

The existence of empty transformation fixed points and the definition of safety for transformation once again highlight the importance of taking each others' intention into account while simplifying decision problems. The fact that the pruning and cleaning do commute when there is only one agent is in that respect illuminating.

6 Conclusion

We have studied two aspects of the reason-centered commitment of intentions, by extending game-theoretic formalisms with two new operations on strategic form games. We characterized conditions under which these operations keep the games or the intentions of the agents non-empty. This has revealed an important interactive character to the reason-centering commitment, one which went up to now unnoticed in philosophical theories of intentions. This work thus extends game-theoretic models and shew new lights on the theory of intentions.

Taking a epistemic perspective, in the line of Aumann [1999], van Benthem [2003], Brandenburger [2007] and Bonanno [2007], would surely enhance the present work. Mutual knowledge of each others' intentions seems crucial in the process of cleaning and pruning. It would also be interesting to relate the current proposal with game-theoretic work on intention formation and reconsideration, e.g. McClennen [1990] and Gul and Pesendorfer [2001], and with the BDI architectures cited in the introduction.

References

K.R. Apt. The many faces of rationalizability. *The B.E. Journal of Theoretical Economics*, 7(1), 2007. Article 18.

R.J. Aumann. Interactive epistemology I: Knowledge. *International Journal of Game Theory*, 28:263–300, 1999.

G. Bonanno. Two lectures on the epistemic foundations of game theory. URL `http://www.econ.ucdavis.`

`edu/faculty/bonanno/wpapers.htm`. Delivered at the Royal Netherlands Academy of Arts and Sciences (KNAW), February 8, 2007.

A. Brandenburger. The power of paradox: some recent developments in interactive epistemology. *International Journal of Game Theory*, 35:465–492, 2007.

M. Bratman. *Intention, Plans and Practical Reason*. Harvard University Press, London, 1987.

P. Gärdenfors, editor. *Belief Revision*. Cambridge UP, 2003.

M. Georgeff, B. Pell, M.E. Pollack, M. Tambe, and M. Wooldridge. The belief-desire-intention model of agency. In J. Muller, M. Singh, and A. Rao, editors, *Intelligent Agents V*. Springer, 1998.

F. Gul and W. Pesendorfer. Temptation and self-control. *Econometrica*, 69(6):1403–1435, November 2001.

E.F. McClennen. *Rationality and Dynamic Choice : Foundational Explorations*. Cambridge University Press, 1990.

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

H. Rott. *Change, Choice and Inference: A Study of Belief Revision and Nonmonotonic Reasoning*. Oxford Logic Guides. ford University Press, Oxford, 2001.

O. Roy. *Thinking before Acting: Intentions, logic, rational choice*. PhD thesis, Universiteit van Amsterdam, February 2008.

A. Sen. *Collective Choice and Social Welfare*. Holden-Day, 1970.

J. van Benthem. Rational dynamic and epistemic logic in games. In S. Vannucci, editor, *Logic, Game Theory and Social Choice III*, pages 19–23. University of Siena, department of political economy, 2003.

W. van der Hoek, W. Jamroga, and M. Wooldrige. Towards a theory of intention revision. *Synthese*, 155, March 2007. Knowledge, Rationality & Action 103-128.

J.D. Velleman. What good is a will? Downloaded from the author's website on April 5th 2006, April 2003.

7 Appendix

7.1 Proof of Fact 3.1

For any strategic game $\mathbb{G}$ and intention profile ι, $S^* \neq \emptyset$ iff $cl^{\#}(\mathbb{G})$ is not empty.

By Definition, $S^* \neq \emptyset$ is the same as saying that we can find a $\sigma \in S^*$ such that for all i, $\pi^{S^*}(\sigma(i)) \cap \iota_i \neq \emptyset$. We show by induction that $\pi(S^k) = X^{cl^k}$, for all k. This is enough to show the equivalence, for then we know that $X^{cl^{\#}} \cap \iota_i \neq \emptyset$, which we know is the same as $cl^{\#}(\mathbb{G})$ being non-empty. The basic case of the induction, $k = 0$, is trivial. For the induction step, assume the claim is proved for k. We have that $x \in \pi(S^{k+1})$ iff there is a $\sigma \in S^{k+1}$ such that $\pi(\sigma) = x$. This in turns happens iff $\pi^{S^k}(\sigma(i)) \cap \iota_i \neq \emptyset$, for all i. But by the inductive hypothesis this is just to say that $\pi(\sigma(i)) \cap X^{cl^k} \cap \iota_i \neq \emptyset$, which is just the definition of x being in X^{x+1}.

7.2 Proof of Fact 3.2

For $\mathbb{G}$ an arbitrary strategic game, $cl^{\#}(\mathbb{G}) = cl(\mathbb{G})$ for cleaning with altruistic admissibility.

We show that $cl(cl(\mathbb{G})) = cl(\mathbb{G})$. Given the definition of the cleaning operation, it is enough to show that $cl(cl(S_i)) = cl(S_i)$ for all i. It should be clear that $cl(cl(S_i)) \subseteq cl(S_i)$. It remains to show the converse. So assume that $s_i \in cl(S_i)$. Since cleaning is done with altruistic admissibility, this means that there is a σ such that $\sigma(i) = s_i$ and a $j \in I$ such that $\pi(\sigma) \in \iota_j$. But then $\sigma(i') \in cl(S_{i'})$ for all $i' \in I$, and so $\sigma \in \Pi_{i \in I} cl(S_i)$. This means that $\pi(\sigma) \in X^{cl}$, which in turns implies that $\pi^{cl}(\sigma) \in \iota_j^{cl}$. We thus know that there is a $\sigma \in \Pi_{i \in I} cl(S_i)$ such that $\sigma(i) = s_i$ and a j such that $\pi^{cl}(\sigma) \in \iota_j^{cl}$, which means that $s_i \in cl(cl(S_i))$.

7.3 Proof of Fact 3.3

For any strategic game $\mathbb{G}$, intention profile ι and cleaning with altruistic admissibility, there is, for all i, a realizable $x \in \iota_i$ iff $cl^{\#}(\mathbb{G})$ is not empty.

There is a realizable $x \in \iota_i$ for all i iff for all i there is a σ such that $\pi(\sigma) \in \iota_i$. But this is this same as to say that for all j there is a strategy s_j such that $\sigma(j) = s_j$ and an i such that $\pi(\sigma) \in \iota_i$ which, by Facts 3.1 and 3.2, means that $cl^{\#}(\mathbb{G})$ is not empty.

7.4 Proof of Fact 4.1

For all strategic game $\mathbb{G}$ and agent $i \in I$, $pr^{\#}(S_i) \neq \emptyset$.

This is shown by induction on $pr^k(\mathbb{G})$. The basic case is trivial. For the induction step, observe that the picking function θ_i is defined for the whole power set of S_i. This means, given the inductive hypothesis, that $\theta_i([s_i]_{\approx}^{pr^k(\mathbb{G})})$ is well-defined and in $[s_i]^{pr^k(\mathbb{G})}$ for any $s_i \in pr^k(S_i)$, which is enough to show that $pr^{k+1}(S_i)$ is also not empty.

7.5 Proof of Fact 4.2

For any strategic game $\mathbb{G}$, intention profile ι, profile of picking function Θ and for all $i \in I$, $\iota_i^{pr^\#} \neq \emptyset$ iff there is a $\pi(\sigma) \in \iota_i$ safe for pruning in $\mathbb{G}$.

From right to left. Take any $x \in \iota_i^{pr^\#}$. By definition we know that there is a $\sigma \in \Pi_{i \in I} pr^\#(S_i)$ such that $\pi(\sigma) = x$. But this happens iff $\sigma \in \Pi_{i \in I} pr^k(S_i)$ for all k, and so that $\theta_i([\sigma(i)]_{\approx}^{pr^k(\mathbb{G})}) = \sigma(i)$ also for all k, which in turns means that x is safe for pruning in $\mathbb{G}$. Left to right, take any such $\pi(\sigma) \in \iota_i$. We show that $\pi(\sigma) \in X^{pr^k}$ for all k. The basic case is trivial, so assume that $\pi(\sigma) \in X^{pr^k}$. We know by definition that $\pi(\sigma)$ is safe for pruning at k, which gives automatically that $\pi(\sigma) \in X^{pr^{k+1}}$.

7.6 Proof of Fact 5.2

For any strategic game $\mathbb{G}$, intention profile ι and profile of consistent picking function Θ, if $\pi(\sigma)$ is safe for transformation in $\mathbb{G}$ then for all fixed points $t^\#(\mathbb{G})$, $\sigma \in \Pi_i t^\#(S_i)$.

This is shown by induction on k for an arbitrary fixed point $t^k(S_i)$. The proof is a direct application of the definition of safety for transformation.

7.7 Proof of Fact 5.3

For any strategic game $\mathbb{G}$, intention profile ι and profile of consistent picking function Θ, if $\sigma \in \Pi_i t^\#(S_i)$ for all fixed points $t^\#(\mathbb{G})$, then $\pi(\sigma)$ is safe for transformation in $\mathbb{G}$.

We show by “backward” induction that $\pi(\sigma)$ is safe for transformation at any k for all sequences $t^k(\mathbb{G})$. For the basic case, take k to be the length of the longest, non-redundant fixed point of $\mathbb{G}$. I show that $\pi(\sigma)$ is safe for transformation at stage k for all sequences of that length. Observe that by the choice of k all $t^k(\mathbb{G})$ are fixed points. We thus know by assumption that $\sigma \in \Pi_{i \in I} t^k(S_i)$. But then it must be safe for transformation at stage k. If clause (1) was violated at one of these, say $t'^k(\mathbb{G})$, then we would have $cl(t'^k(\mathbb{G})) \neq t'^k(\mathbb{G})$, against the fact that $t'^k(\mathbb{G})$ is a fixed point. By the same reasoning we know that clause (2) cannot be violated either. Furthermore, by the fact that $t'^{k+1}(\mathbb{G}) = t'^k(\mathbb{G})$, we know that it is safe for transformation at all stages $l > k$.

For the induction step, take any $0 \leq n < k$ and assume that for all sequences $t^{n+1}(\mathbb{G})$ of length $n+1$, $\pi(\sigma)$ is safe for transformation at stage $n+1$. Take any $t^n(\mathbb{G})$. By our induction hypothesis, that $\pi(\sigma)$ is safe for transformation at both $cl(t^n(\mathbb{G}))$ and $pr(t^n(\mathbb{G}))$. This secures clause (2) of the definition of safety for transformation, and also gives us that $\sigma \in \Pi_{i \in I} t^n(S_i)$. Now, because it is safe for transformation in $cl(t^n(\mathbb{G}))$, we know that $\pi^{cl(t^n(\mathbb{G}))}(\sigma(i)) \cap \iota_i^{cl(t^n(\mathbb{G}))} \neq \emptyset$ for all i. But since $\pi^{cl(t^n(\mathbb{G}))}(\sigma(i)) \subseteq \pi^{t^n(\mathbb{G})}(\sigma(i))$, and the same for the intention set, we know that $\pi^{t^n(\mathbb{G})}(\sigma(i)) \cap \iota_i^{t^n(\mathbb{G})} \neq \emptyset$ for all i. For condition (2), we also know that $\theta_i[\sigma(i)]_{\approx}^{cl(t^n(\mathbb{G}))} = \sigma(i)$ for all i from the fact that $\pi(\sigma)$ is safe for transformation at stage $n+1$. By Lemma 7.1 (below) and the assumption that θ_i is consistent for all i, we can conclude that $\theta_i[\sigma(i)]_{\approx}^{t^n(\mathbb{G})} = \sigma(i)$, which completes the proof because we took an arbitrary $t^n(\mathbb{G})$.

Lemma 7.1 *For any game strategic game $\mathbb{G}$ and intention set ι_i and strategy $s_i \in cl(S_i)$, $[s_i]_{\approx}^{\mathbb{G}} \subseteq [s_i]_{\approx}^{cl(\mathbb{G})}$.*

Proof. Take any $s'_i \in [s_i]_{\approx}^{\mathbb{G}}$. Since $s_i \in cl(S_i)$, we know that there is a $\sigma_{j \neq i}$ such that $\pi(s_i, \sigma_{j \neq i}) \in \iota_i$. But because $s'_i \approx s_i$, it must also be that $\pi(s'_i, \sigma_{j \neq i}) \in \iota_i$, and so that $s'_i \in cl(S_i)$. Now, observe that $\{\sigma \in \Pi_{i \in I} cl(S_i) : \sigma(i) = s_i\} \subseteq \{ \sigma \in S_i : \sigma(i) = s_i\}$, and the same for s'_i. But then, because $s'_i \approx s_i$, it must also be that $s'_i \in [s_i]_{\approx}^{cl(\mathbb{G})}$. QED

A Logic for Cooperation, Actions and Preferences

Lena Kurzen
Universiteit van Amsterdam
L.M.Kurzen@uva.nl

Abstract

In this paper, a logic for reasoning about cooperation, actions and preferences of agents is developed. It is shown to be sound and complete and the satisfiability problem of its fragment that does not contain strict preferences is shown to be NExpTime-complete.

1. Introduction

When analyzing interactive situations involving multiple agents, we are interested in what results agents can achieve – individually or together as groups. In many cases, agents can have various plans for achieving some result. These plans can differ significantly, e.g. with respect to their feasibility, costs or side-effects. Hence, it is not only relevant which results groups of agents can achieve but also *how* exactly they can do so. In other words, plans and actions also play a central role in interactive processes. Cooperative ability of agents expressed only in terms of results and actions that lead to these results does not tell us *why* agents would actually decide to achieve a certain result. We also need to take into account the preferences based on which the agents decide what to do.

Summarizing, we can say that in interactive situations, the following three questions are of interest and moreover tightly connected:

- ***What** results can groups of agents achieve?*
- ***How** can they achieve something?*
- ***Why** would they want to achieve a certain result?*

The above considerations show that coalitional power, actions and preferences play a major role in interactive situations and are moreover tightly connected. Thus, a formal theory for reasoning about agents' cooperative abilities in an explicit way should also take into account actions/plans of agents and their preferences.

In logic, these aspects have mostly been addressed separately. Coalitional power has mainly been investigated within the frameworks of ATL [4], Coalition Logic [12] and their extensions. These logics focus on what results groups can achieve and do not represent explicitly how exactly results can be achieved. Recently, there have been some attempts to develop logics for reasoning about coalitional power that also take into account either agents' preferences or actions. One group of such logics looks at cooperation from the perspective of cooperative games [1, 2]. Another path that has been taken in order to make coalitional power more explicit is to combine cooperation logics with action logics [14, 6, 7].

In this paper, a logic for reasoning about cooperation, actions and preferences (CLA+P) is developed, which is obtained by combining the cooperation logic with actions CLA [14] with a preference logic [15, 16]. Soundness and completeness are shown and the logics expressivity and computational complexity are investigated.

The remainder of this paper is structured as follows. Section 2 gives a brief overview of the cooperation logic with actions CLA [14]. In Section 3, a cooperation logic with actions and preferences (CLA+P) is developed and soundness and completeness are shown. Also its expressivity is discussed. Section 4 investigates the computational complexity of CLA+P.

2. Cooperation Logic with Actions (CLA)

In this section, we briefly present the cooperation logic with actions (CLA) developed by Sauro et al. [14], which will be extended in the next section by combining it with a preference logic. The idea of CLA is to make coalitional power explicit by expressing it in terms of the ability to perform actions instead of expressing it directly in terms of the ability to achieve certain results. CLA is a modular modal logic, consisting of an environment module for reasoning about actions and their effects, and an agents module for reasoning about agents' abilities to perform actions. By combining both modules, a framework is obtained in which cooperative ability can be made more explicit.

The environment, which at a later step will be populated by agents, is modelled as a labelled transition system whose edges are labeled with sets of atomic actions.

Definition 2.1 [Environment Model [14]] An environment model is a set-labelled transition system

$$E = \langle S, Ac, (\rightarrow)_{A \subseteq Ac}, V \rangle.$$

S is a set of states, Ac is a finite set of atomic actions, $\rightarrow_A \subseteq S \times S$ for each $A \subseteq Ac$ and V is a propositional valuation. $\rightarrow_A$ is required to be serial for each $A \subseteq Ac$. ◁

Then a modal language is defined with modalities of the form $[\alpha]$, for α being a propositional formula built from atomic actions. The intended meaning of $[\alpha]\varphi$ is that every transition $\rightarrow_A$ such that $A \vDash \alpha$ (using the satisfaction relation of propositional logic) leads to a state that satisfies φ. Formally,

$$E, s \vDash [\alpha]\varphi \quad \text{iff} \quad \forall A \subseteq Ac, s' \in S : \text{if } A \vDash \alpha \text{ and } s \rightarrow_A s' \text{ then } E, s' \vDash \varphi.$$

An environment logic Λ^E is developed, which is sound and complete with respect to the class of environment models [14]. It contains seriality axioms and the **K** axiom for each modality $[\alpha]$, for α being consistent. For the details, the reader is referred to Sauro et al. [14]. The environment logic, can then be used for reasoning about the effects of concurrent actions.

Before agents are introduced into the environment, a separate agents module is developed for reasoning about the ability of (groups of) agents to perform actions. Each agent is assigned a set of atomic actions that he can perform and a group is assigned the set of actions its members can perform.

Definition 2.2 [Agents Model [14]] An agents model is a triple $\langle Ag, Ac, \mathsf{act} \rangle$, where Ag is a set of agents, Ac is a set of atomic actions and act is a function $\mathsf{act} : Ag \rightarrow \mathcal{P}(Ac)$ such that $\bigcup_{i \in Ag} \mathsf{act}(i) = Ac$. For $G \subseteq Ag$, $\mathsf{act}(G) := \bigcup_{i \in G} \mathsf{act}(i)$. ◁

We are not only interested in what atomic actions agents can perform but also in their abilities to enforce more complex actions. An agent laguage is developed with expressions $\langle\langle G \rangle\rangle \alpha$, meaning that the group G can force that a concurrent action is performed that satisfies α. This means that G can perform some set of atomic actions such that no matter what the other agents do, the resulting set of actions satisfies α.

$$\langle Ag, Ac, \mathsf{act} \rangle \vDash \langle\langle G \rangle\rangle \alpha \quad \text{iff} \quad \exists A \subseteq \mathsf{act}(G) : \forall B \subseteq \mathsf{act}(Ag \setminus G) : A \cup B \vDash \alpha.$$

Then a cooperation logic for actions is developed, which is very much in the style of Coalition Logic [12] – the main difference being that it is concerned with the cooperative ability to force *actions*.

Definition 2.3 [Coalition Logic for Actions [14]] The coalition logic for actions Λ^A is defined to be the logic derived from the following set of axioms.

1. $\langle\langle G \rangle\rangle \top$, for all $G \subseteq Ag$,
2. $\langle\langle G \rangle\rangle \alpha \rightarrow \neg \langle\langle Ag \setminus G \rangle\rangle \neg \alpha$,
3. $\langle\langle G \rangle\rangle \alpha \rightarrow \langle\langle G \rangle\rangle \beta$ if $\vdash \alpha \rightarrow \beta$ in propositional logic,
4. $\langle\langle G \rangle\rangle a \rightarrow \bigvee_{i \in G} \langle\langle \{i\} \rangle\rangle a$ for all $G \subseteq Ag$ and atomic $a \in Ac$,
5. $(\langle\langle G_1 \rangle\rangle \alpha \wedge \langle\langle G_2 \rangle\rangle \beta) \rightarrow \langle\langle G_1 \cup G_2 \rangle\rangle (\alpha \wedge \beta)$, for $G_1 \cap G_2 = \emptyset$,
6. $(\langle\langle G \rangle\rangle \alpha \wedge \langle\langle G \rangle\rangle \beta) \rightarrow \langle\langle G \rangle\rangle (\alpha \wedge \beta)$ if α and β have no common atomic actions,
7. $\langle\langle G \rangle\rangle \neg a \rightarrow \langle\langle G \rangle\rangle a$ for atomic $a \in Ac$,
8. $\langle\langle G \rangle\rangle \alpha \rightarrow \bigvee \{ \langle\langle G \rangle\rangle \bigwedge \Psi | \Psi$ is a set of literals such that $\bigwedge \Psi \rightarrow \alpha \}$.

The rule of inference is modus ponens. ◁

Axiom 5 says how groups can join forces. The coalition logic for actions is sound and complete with respect to the class of agents models [14].

Next, agents are introduced as actors into the environment. This is done by combining the environment models with the agents models. In the resulting models, the agents can perform actions which then have the effect of changing the current state of the environment.

Definition 2.4 [Multi-agent System [14]] A multi-agent system (MaS) is a tuple

$$M = \langle S, Ac, (\rightarrow)_{A \subseteq Ac}, V, Ag, \mathsf{act} \rangle,$$

where $\langle S, Ac, (\rightarrow)_{A \subseteq Ac}, V \rangle$ is an environment model and $\langle Ac, Ag, \mathsf{act} \rangle$ an agents model. ◁

Now, we can reason about what states of affairs groups of agents can achieve by performing certain actions. The corresponding language contains all expressions of the environment logic and the cooperation logic for actions and additionally expressions for saying that a group has the power to achieve φ.

Definition 2.5 [Language for MaS [14]] The language for multi-agent systems $\mathcal{L}_{cla}$ is generated by the following grammar:

$$\varphi ::= \quad p \mid \varphi \wedge \varphi \mid \neg\varphi \mid [\alpha]\varphi \mid \langle\langle G\rangle\rangle\alpha \mid \langle\langle G\rangle\rangle\varphi$$

for $G \subseteq Ac$ and α being an action expression. ◁

$\langle\langle G\rangle\rangle\varphi$ means that G can force that the system moves into a φ-state, i.e. G can perform some set of actions such that no matter what the other agents do, the system will move into a φ-state.

$$M, s \vDash \langle\langle G\rangle\rangle\varphi \quad \text{iff} \quad \exists A \subseteq \mathsf{act}(G) \text{ such that } \forall B \subseteq \mathsf{act}(Ag \setminus G), t \in S: \text{ if } s \rightarrow_{A\cup B} t, \text{ then } M, s \vDash \varphi.$$

The environment logic and the coalition logic for agents are combined by adding two interaction axioms.

Definition 2.6 [Cooperation Logic with Actions [14]] The cooperation logic with actions Λ^{CLA} combines the environment logic Λ^{E} and the coalition logic for actions Λ^{A} by adding

1. $(\langle\langle G\rangle\rangle\alpha \wedge [\alpha]\varphi) \rightarrow \langle\langle G\rangle\rangle\varphi$,
2. $\langle\langle G\rangle\rangle\varphi \rightarrow \bigvee\{\langle\langle G\rangle\rangle\alpha \wedge [\alpha]\varphi | \alpha$ is the conjunction of a set of atomic actions or their negations$\}$.

◁

CLA provides us with a formal framework for reasoning about what states of affairs groups of agents can achieve and how they can do so. For a detailed discussion of CLA, the reader is referred to Sauro et al. [14]. Now, we procede by adding an explicit representation of the agents' preferences to CLA.

3. Cooperation Logic with Actions and Preferences (CLA+P)

In this section, a logic for reasoning about cooperation, actions and preferences is developed. This is done by adding a preference logic to CLA. For a more detailed discussion of what is covered in this section and detailed proofs, see [8].

3.1. Preference Logic

There are various ways how preferences can be added to a logic for cooperation and actions. We could for instance let the preferences of each agent range over the actions that he can perform. Alternatively, we can think of each agent having preferences over the set of successor states of the current state.

In the current work, we consider preferences of single agents over the states of the environment. This is reasonable since by performing actions the agents can change the current state of the environment, and the preferences over those states can be seen as the base of how the agents decide how to act. Such a preference relation can also be lifted to one over formulas [15, 16].

Definition 3.1 [Preference Model [16]] A preference model is a tuple

$$M^P = \langle S, Ag, \{\preceq_i\}_{i\in Ag}, V\rangle,$$

where S is a set of states, Ag is a set of agents, for each $i \in Ag$, $\preceq_i \subseteq S \times S$ is a reflexive and transitive relation and V is a propositional valuation. ◁

As a language, we use a fragment of the basic preference language developed by van Benthem et al. [15]. It has strict and non-strict unary preference modalities for each agent.

Definition 3.2 [Preference Language] Given a set of propositional variables and a finite set of agents Ag, define the preference language $\mathcal{L}_p$ to be the language generated by the following syntax:

$$\varphi := \quad p \mid \neg\varphi \mid \varphi \vee \varphi \mid \Diamond^{\preceq_i}\varphi \mid \Diamond^{\prec_i}\varphi.$$

◁

$\Diamond^{\preceq_i}\varphi$ says that there is a state satisfying φ, and agent i prefers this state over the current one, i.e.

$$M^P, s \vDash \Diamond^{\preceq_i}\varphi \text{ iff } \exists t: s \preceq_i t \text{ and } M^P, t \vDash \varphi.$$

$\Diamond^{\prec_i}\varphi$ is interpreted analogously. The preference relation $\preceq$ should be reflexive and transitive and $\prec$ should be its largest irreflexive subrelation. Thus, the following axiomatization is chosen.

Definition 3.3 [Preference Logic Λ^P] For a given set of agents Ag, let Λ^P be the logic generated by the following axioms for each agent $i \in Ag$:

For $\Diamond^{\preceq_i}$ and $\Diamond^{\prec_i}$, we have duality axioms and **K**. For $\Diamond^{\preceq_i}$, we also have reflexivity and transitivity axioms. Additionally, there are four axioms for the interaction between the strict and non-strict modalities:

1. $\Diamond^{\prec_i}\varphi \rightarrow \Diamond^{\preceq_i}\varphi$,
2. $\Diamond^{\preceq_i}\Diamond^{\prec}\varphi \rightarrow \Diamond^{\prec_i}\varphi$,
3. $\Diamond^{\prec_i}\Diamond^{\preceq_i}\varphi \rightarrow \Diamond^{\prec_i}\varphi$,
4. $\varphi \wedge \Diamond^{\preceq_i}\psi \rightarrow (\Diamond^{\prec_i}\psi \vee \Diamond^{\preceq_i}(\psi \wedge \Diamond^{\preceq_i}\varphi))$.

The inference rules are modus ponens, necessitation and substitution of logical equivalents. ◁

Note that transitivity for $\Diamond^{\prec_i}$ follows from the above axioms. We can show soundness and completeness of the preference logic. The fact that $\prec$ is supposed to be the greatest irreflexive subrelation of $\preceq$ can be dealt with by using the bulldozing technique. For the details, the reader is referred to van Benthem et al. [15].

Theorem 3.4 Λ^P *is sound and complete with respect to the class of preference models.*

Proof. Follows from Theorem 3.9 in [15]. QED

3.2. Environment Logic with Preferences

As an intermediate step towards a logic for reasoning about cooperation, actions and preferences, we first combine the preference logic and the environment logic. The two models are combined by identifying their sets of states. Then the preferences of the agents range over the states of the environment. In such a system, the agents cannot act in the environment, but they can rather be seen as observers that observe the environment from the outside and have preferences over the states.

Definition 3.5 [Environment with Preferences] An environment model with preferences is a tuple

$$E^{\preceq} = \langle S, Ac, (\rightarrow)_{A \subseteq Ac}, \{\preceq_i\}_{i \in Ag}, V \rangle,$$

where $\langle S, Ac, (\rightarrow)_{A \subseteq Ac}, \{\preceq_i\}_{i \in Ag}, V \rangle$ is an environment model and $\langle S, Ag, \{\preceq_i\}_{i \in Ag}, V \rangle$ is a preference model. $\lhd$

We combine the languages for the environment and the preferences and add expressions for saying that "every state accessible by an α transition is (strictly) preferred by agent i over the current state".

Convention 3.6 *In what follows, we write the symbol $\lhd$ in statements that hold for both $\preceq$ and $\prec$, each uniformly substituted for $\lhd$.*

Definition 3.7 The language $\mathcal{L}_{ep}$ contains all expressions of the environment language and the preference language and additionally formulas of the forms $[\alpha]^{\preceq_i}\top$ and $[\alpha]^{\prec_i}\top$, for α being an action expression.

Boolean combinations and expressions of the previously defined languages are interpreted in the standard way. For the newly introduced expressions, we have:

$$E^{\preceq}, s \vDash [\alpha]^{\lhd_i}\top \quad \text{iff} \quad \forall A \subseteq Ac, t \in S: \text{ if } s \rightarrow_A t \text{ and } A \vDash \alpha \text{ then } s \lhd_i t.$$

$\lhd$

Expressions of the form $[\alpha]^{\lhd_i}\top$ cannot be defined using just the preference language and the environment language. To see this, note that $[\alpha]^{\preceq_i}\top$ says that for every state accessible by an α-transition it holds that this same state is accessible by $\preceq$. Thus, we would have to be able to refer to particular states. Therefore, we add two inference rules for deriving the newly introduced expressions.

$$\text{(PREF-ACT)} \qquad \frac{\Box^{\preceq_i}\varphi \rightarrow [\alpha]\varphi}{[\alpha]^{\preceq_i}\top}$$

$$\text{(STRICT PREF-ACT)} \qquad \frac{\Box^{\prec_i}\varphi \rightarrow [\alpha]\varphi}{[\alpha]^{\prec_i}\top}$$

In order to obtain a complete axiomatization, two axioms are added which correspond to the converse of the inference rules.

Theorem 3.8 *Let Λ^{EP} be the logic generated by all axioms of the environment logic Λ^E, all axioms of the preference logic Λ^P, and*

1. $[\alpha]^{\preceq_i}\top \rightarrow (\Box^{\preceq_i}\varphi \rightarrow [\alpha]\varphi)$,

2. $[\alpha]^{\prec_i}\top \rightarrow (\Box^{\prec_i}\varphi \rightarrow [\alpha]\varphi)$.

The inference rules are modus ponens, substitution of logical equivalents, PREF-ACT and STRICT PREF-ACT. Then Λ^{EP} is sound and complete with respect to the class of environment models with preferences.

Proof. Completeness follows from completeness of the sublogics and the closure under the new rules. QED

In the environment logic with preferences, the performance of concurrent actions changes the current state of the system also with respect to the 'happiness' of the agents: A transition from one state to another can also be a transition up or down in the preference orderings of the agents.

3.3. Cooperation Logic with Actions and Preferences

Now, agents are introduced as actors by combining the environment models with preferences with agents models. The resulting model is then called a multi-agent system with preferences (henceforth MaSP).

Definition 3.9 [Multi-agent System with Preferences] A multi-agent system with preferences (MaSP) $M^{\preceq}$ is a tuple

$$M^{\preceq} = \langle S, Ac, (\rightarrow)_{A \subseteq Ac}, Ag, \mathsf{act}, \{\preceq_i\}_{i \in Ag}, V \rangle,$$

where $\langle S, Ac, (\rightarrow)_{A \subseteq Ac}, V, Ag, \mathsf{act} \rangle$ is a MaS, $\langle S, Ag, \{\preceq_i\}_{i \in Ag}, V \rangle$ is a preference model and $\langle S, Ac, (\rightarrow)_{A \subseteq Ac}, \{\preceq_i\}_{i \in Ag}, V \rangle$ is an environment with preferences. $\lhd$

In order to get some intuitions about how MaSP's are related to other models of interaction, note that given a deterministic MaSP in which each preference relation $\preceq_i$ is total, we can consider each state s as having a strategic game

$$\mathcal{G}_s = \langle Ag, (\mathcal{P}(\mathsf{act}(i)))_{i \in Ag}, (\lesssim_i)_{i \in Ag} \rangle$$

attached to it, where $\times_{i=1}^{n} A_i \quad \lesssim_i \quad \times_{i=1}^{n} A'_i$ iff $t \quad \preceq_i t'$ for $s \rightarrow_{\bigcup_{i \in Ag} A_i} t$ and $s \rightarrow_{\bigcup_{i \in Ag} A'_i} t'$.

For talking about the cooperative ability of agents with respect to preferences, we introduce two expressions saying that a group can force the system to move to a φ-state that some agent (strictly) prefers over the current one.

Definition 3.10 [Language $\mathcal{L}_{cla+p}$] The language $\mathcal{L}_{cla+p}$ extends $\mathcal{L}_{cla}$ by formulas of the form

$$\Diamond^{\preceq_i} \varphi \mid \Diamond^{\prec_i} \varphi \mid [\alpha]^{\preceq_i} \top \mid [\alpha]^{\prec_i} \top \mid \langle\langle G^{\preceq_i} \rangle\rangle \varphi \mid \langle\langle G^{\prec_i} \rangle\rangle \varphi.$$

The first four expressions are interpreted as in the environment logic with preferences and for the last two we have the following.

$M^{\preceq}, s \vDash \langle\langle G^{\lhd_i} \rangle\rangle \varphi$ iff $\exists A \subseteq \mathsf{act}(G)$ such that $\forall B \subseteq \mathsf{act}(Ag \setminus G), t \in S$: if $s \rightarrow_{A \cup B} t$, then $M^{\preceq}, t \vDash \varphi$ and $s \lhd_i t$.

◁

Let us now look at how coalitional power to achieve an improvement for an agent is made explicit in CLA+P. We can show that $\langle\langle G^{\lhd_i} \rangle\rangle \varphi$ is equivalent to the existence of an action expression α that G can force and that has the property that all transitions of type α are guaranteed to lead to a φ-state preferred by agent i.

Observation 3.11 *Given a MaSP $M^{\preceq}$ and a state s of its environment,*

$M^{\preceq}, s \vDash \langle\langle G^{\lhd_i} \rangle\rangle \varphi$ *iff there exists an action expression α such that $M^{\preceq}, s \vDash \langle\langle G \rangle\rangle \alpha \wedge [\alpha] \varphi \wedge [\alpha]^{\lhd_i} \top$.*

Proof. Analogous to that of Observation 14 in [14]. For the left-to-right direction, use the action expression $\bigwedge \Phi(A, G) := \bigwedge (A \cup \{\neg a | a \in (\mathsf{act}(G) \setminus A), a \notin \mathsf{act}(Ag \setminus G)\})$. QED

Thus, formulas of the form $\langle\langle G^{\lhd_i} \rangle\rangle \varphi$ can be reduced to expressions of the sublogics. We also need new axioms establishing a relationship between the newly added formulas and the expressions of the sublogics.

Definition 3.12 [Cooperation Logic with Actions and Preferences Λ^{CLA+P}] Define Λ^{CLA+P} to be the smallest logic generated by the axioms of the cooperation logic with actions, the environment logic with preferences and

1. $(\langle\langle G \rangle\rangle \alpha \wedge [\alpha] \varphi \wedge [\alpha]^{\lhd_i} \top) \rightarrow \langle\langle G^{\lhd_i} \rangle\rangle \varphi$,

2. $\langle\langle G^{\preceq_i} \rangle\rangle \varphi \rightarrow \bigvee \{\langle\langle G \rangle\rangle \alpha \wedge [\alpha] \varphi \wedge [\alpha]^{\preceq_i} \top | \alpha$ is a conjunction of action literals$\}$,

3. $\langle\langle G^{\prec_i} \rangle\rangle \varphi \rightarrow \bigvee \{\langle\langle G \rangle\rangle \alpha \wedge [\alpha] \varphi \wedge [\alpha]^{\prec_i} \top | \alpha$ is a conjunction of action literals$\}$.

The inference rules are modus ponens, necessitation for action modalities and preference modalities $(\Box^{\preceq_i}, \Box^{\prec_i})$, substitution of logical equivalents, $\mathrm{PREF-ACT}$ and $\mathrm{STRICT\ PREF-ACT}$. ◁

Soundness of the axioms is straightforward and completeness follows from completeness of the sublogics.

Theorem 3.13 *The logic Λ^{CLA+P} is sound and complete with respect to the class of MaSP's.*

3.4. Expressivity of CLA+P

The framework of CLA+P allows us to reason about coalitional power in an explicit way since we can express how groups of agents can achieve the truth of some formula, and moreover we can also express how coalitional power and actions relate to the agents' preferences.

In game theory, coalitional power has mostly been studied within coalitional games [11]. One of the most important solution concepts of coalitional games is the core, which is the set of outcomes the grand coalition can achieve that have the property that no coalition can achieve some other outcome that is strictly better for all its members.

In the framework of CLA+P, we can characterize the states in a model that have a very similar property: The formula $\hat{\psi}$ characterizes the set of states in which no group has the power of making the system move into a state that is strictly better for all its members:

$$\hat{\psi} := \bigwedge_{G \subseteq Ag} \bigwedge_{A \subseteq \mathsf{act}(G)} \left(\bigvee_{i \in G} \neg [\bigwedge \Phi(A, G)]^{\prec_i} \top \right).$$

For the definition of $\bigwedge \Phi(A, G)$, see the proof of Observation 3.11.

In interactive situations, there can be different ways of how agents can achieve some result. These ways can consist of different plans that the agents can execute, and whereas all the actions or plans might lead to the same result φ, executing one plan might be better for the agents than executing another one. Being 'better' could be in the sense that one plan leads to an improvement of the situation for more agents than executing another plan does. In CLA+P, we can express that a group G can achieve φ in such a way that the situation improves for all its members:

$$\bigvee_{A\subseteq \mathsf{act}(G)} \Big(([\bigwedge \Phi(A,G)]\,\varphi) \wedge \bigwedge_{i\in G} [\bigwedge \Phi(A,G)]^{\preceq_i} \top \Big).$$

Thus, the explicit representation of actions and preferences allows us to reason about how exactly a group would choose to achieve some result, assuming that the members make their decisions according to a certain solution concept.

Alternatively, executing one plan might be better than another one in the sense that it is cheaper. By having both actions and preferences in our framework, we can also express how actions and preferences interact and thereby our framework can also give rise to a formal model for cost-benefit analysis. In cost-benefit-analysis, decisions are made by comparing the expected cost of executing actions and the expected benefit.

4. Complexity of CLA+P

In this section, we investigate the complexity of the satisfiability problem of CLA+P. Let us start by trying to determine an upper bound.

4.1. Upper Bound for CLA+P

In order to establish an upper bound, it has to be shown that computing whether some formula is satisfiable can be done using a certain amount of time or space. The first step is to show that only a restricted class of models of CLA+P needs to be checked.

For a formula φ, let $Ag(\varphi)$ denote the set of agents occurring in φ. Now, we ask: Is any satisfiable formula φ also satisfiable in a MaSP whose set of agents is $Ag(\varphi)$? In Coalition Logic, the answer is negative due to formulas such as

$$\varphi' = \neg\langle\langle\{1\}\rangle\rangle p \wedge \neg\langle\langle\{1\}\rangle\rangle q \wedge \langle\langle\{1\}\rangle\rangle(p \vee q),$$

which can only be satisfied in coalition models with at least two agents [13]. However, as in CLA+P the underlying environment models can be nondeterministic, here φ' can indeed be satisfied in a model with only one agent, as the reader can check.

It can be shown that every satisfiable formula $\varphi \in \mathcal{L}_{cla+p}$ is satisfiable in a MaSP with set of agents $Ag(\varphi) \cup \{k\}$, for k being a newly introduced agent. k takes the role of all opponents of $Ag(\varphi)$ collapsed into one. This means that k gets the ability to perform exactly the actions that agents not occurring in φ can perform as a group. When transforming a model satisfying φ into one with set of agents $Ag(\varphi) \cup \{k\}$, we do not need to change the effects of actions or the abilities of agents in $Ag(\varphi)$. This is the main fact that makes the proof of Theorem 4.1 go through. Moreover, note that the preferences of agent k do not have any influence on the truth of φ since k does not occur in φ.

Theorem 4.1 *Every satisfiable formula $\varphi \in \mathcal{L}_{cla+p}$ is satisfiable in the class of MaSP's with at most $|Ag(\varphi)| + 1$ many agents.*

Proof. Assume that $M^{\preceq} = \langle S, Ac, (\rightarrow)_{A\subseteq Ac}, Ag, \mathsf{act}, \{\preceq_i\}_{i\in Ag}, V\rangle$ satisfies φ. If $Ag \supset Ag(\varphi)$, we construct $M'^{\preceq'} = \langle S, Ac, (\rightarrow)_{A\subseteq Ac}, Ag(\varphi) \cup \{k\}, \mathsf{act}', \{\preceq'_i\}_{i\in Ag(\varphi)\cup\{k\}}, V\rangle$, with $\mathsf{act}'(k) = \bigcup_{j\in Ag\setminus Ag(\varphi)} \mathsf{act}(j)$ and $\mathsf{act}'(i) = \mathsf{act}(i)$ for $i \neq k$. The preferences are defined as follows: $\preceq'_i = \preceq_i$ for $i \in Ag(\varphi)$ and $\preceq'_k = S \times S$. By induction, it can be shown that $M^{\preceq}, s \vDash \varphi$ iff $M'^{\preceq'}, s \vDash \varphi$. The interesting case is the one where φ is of the form $\langle\langle G\rangle\rangle\alpha$. Here, the claim follows from the definition of act'. The other cases involving coalition modalities follow. QED

Next, we would like to know how many actions a model needs in order to satisfy some formula. As an example, consider the formula

$$\varphi' = \langle\langle G\rangle\rangle(p \wedge q) \wedge \langle\langle G\rangle\rangle(\neg p \wedge q) \wedge \langle\langle G\rangle\rangle(\neg p \wedge \neg q).$$

It can only be satisfied in models with $|Ac| \geq 2$. The main task is to find "witnesses" for formulas of the form $\langle\langle G\rangle\rangle\psi$ in terms of concurrent actions that tell us how exactly G can achieve ψ. We can show that every satisfiable formula φ can be satisfied in a MaSP whose set of actions consists of the actions occurring in φ, one additional atomic action, and for every subformula of the forms $\langle\langle G\rangle\rangle\psi$ or $\langle\langle G^{\lhd_i}\rangle\rangle\psi$, one atomic action for each of G's members. The one additional action is a dummy that serves for making sure that every agent can perform some action.

The key step in transforming a model satisfying a formula φ into one whose set of actions satisfies the above condition is to define the action distribution and the accessibility relations in an appropriate way. For every action expression α occuring in φ, we have to ensure that two states are accessible by an α-transition in the new model iff they were in the original one. Additionally, for any formula of the forms $\langle\langle G\rangle\rangle\psi$ or $\langle\langle G^{\lhd_i}\rangle\rangle\psi$, the set of actions that we introduced for that formula serves for making explicit how G can force φ. Note that we do not need to introduce any additional actions for making explicit how a group can force an action expression α. This results from the fact that in order to force α, agents only need to perform actions that already occur in α.

Theorem 4.2 *Every satisfiabe formula $\varphi \in \mathcal{L}_{cla+p}$ is satisfiable in a MaSP with at most $|Ac(\varphi)| + (\sum_{\langle\langle G\rangle\rangle\psi\in Sub(\varphi)} |G|) + (\sum_{\langle\langle G^{\preceq_i}\rangle\rangle\psi\in Sub(\varphi)} |G|) + (\sum_{\langle\langle G^{\prec_i}\rangle\rangle\psi\in Sub(\varphi)} |G|) + 1$ many actions.*

Proof. Assume that $M^{\preceq} = \langle S, Ac, (\rightarrow)_{A\subseteq Ac}, Ag, \mathsf{act}, \{\preceq_i\}_{i\in Ag}, V\rangle$ satisfies φ. We construct a model $M'^{\preceq'} = \langle S, Ac', (\rightarrow')_{A'\subseteq Ac'}, Ag, \mathsf{act}', \{\preceq'_i\}_{i\in Ag}, V\rangle$ as follows.

$$Ac' := \quad Ac(\varphi) \quad \cup \quad \bigcup_{\langle\langle G\rangle\rangle\psi \in Sub(\varphi)} A_{G\psi} \cup \bigcup_{\langle\langle G^{\preceq_i}\rangle\rangle\psi \in Sub(\varphi)} A_{G^{\preceq_i}\psi} \cup \bigcup_{\langle\langle G^{\prec_i}\rangle\rangle\psi \in Sub(\varphi)} A_{G^{\prec_i}\psi} \cup \{\hat{a}\}.$$

$A_{G\psi}$ and $A_{G^{\lhd_i}\psi}$ consist of newly introduced actions $a_{G\psi j}$, and $a_{G^{\lhd_i}\psi j}$ respectively, for each $j \in G$. Action abilities are distributed as follows:

$$\mathsf{act}'(i) := \quad (\mathsf{act}(i) \cap Ac(\varphi)) \cup \{\hat{a}\} \cup \{a_{Gi} | \langle\langle G\rangle\rangle\psi \in Sub(\varphi) \text{ or } \langle\langle G^{\lhd_i}\rangle\rangle\psi \in Sub(\varphi), \text{ for } i \in G\}.$$

For defining the accessibility relation $\rightarrow_{A' \subseteq Ac'}$, we first define for any state s its set of successors.

$t \in T^s_{A'}$ iff

1. $\forall[\alpha]\psi \in Sub(\varphi)$ such that $A' \vDash \alpha$: If $M^{\preceq}, s \vDash [\alpha]\psi$, then $M^{\preceq}, t \vDash \psi$,
2. $\forall[\alpha]^{\lhd_i}\top \in Sub(\varphi)$ such that $A' \vDash \alpha$: If $M^{\preceq}, s \vDash [\alpha]^{\lhd_i}\top$, then $s \lhd_i t$,
3. $\forall\langle\langle G\rangle\rangle\psi \in Sub(\varphi)$ such that $A' \vDash \bigwedge \Phi(A_{G\psi}, G)$, there is some $\bar{A} \subseteq \mathsf{act}(G)$ such that $s \rightarrow_A t$ for some $A \subseteq Ac$ such that $A \vDash \bigwedge \Phi(\bar{A}, G)$, and if $M^{\preceq}, s \vDash \langle\langle G\rangle\rangle\psi$ then $M^{\preceq}, s \vDash [\bigwedge \Phi(\bar{A}, G)]\psi$,
4. $\forall\langle\langle G^{\lhd_i}\rangle\rangle\psi \in Sub(\varphi)$ such that $A' \vDash \bigwedge \Phi(A_{G^{\lhd_i}\psi}, G)$, there is some $\bar{A} \subseteq \mathsf{act}(G)$ such that $s \rightarrow_A t$ for some $A \subseteq Ac$ such that $A \vDash \bigwedge \Phi(\bar{A}, G)$, and if $M^{\preceq}, s \vDash \langle\langle G^{\lhd_i}\rangle\rangle\psi$ then $M^{\preceq}, s \vDash [\bigwedge \Phi(\bar{A}, G)]\psi$ and $M^{\preceq}, s \vDash [\bigwedge \Phi(\bar{A}, G)]^{\lhd_i}\top\}$.

For any $t \in T^s_{A'}$, we set $s \rightarrow'_{A'} t$.

Then we can show by induction on $\psi \in Sub(\varphi)$ that $M^{\preceq}, s \vDash \psi$ iff $M'^{\preceq'}, s \vDash \psi$. QED

The next step is to show that every satisfiable formula φ is also satisfiable in a model with a certain number of states. Such results are usually obtained by transforming a model into a smaller one using a transformation that preserves the truth of subformulas of φ. Here, the irreflexivity of the strict preferences is causing problems and thus we restrict our investigations to formulas that do not involve strict preferences. We denote this fragment of $\mathcal{L}_{cla+p}$ by $\mathcal{L}^{\not\prec}_{cla+p}$ and the corresponding fragment of CLA+P by CLA+P$^{\not\prec}$.

Using the method of filtration [5], we show that any satisfiable formula $\varphi \in \mathcal{L}^{\not\prec}_{cla+p}$ is satisfiable in a model with exponentially many states. Note that formulas of the form $\langle\langle G\rangle\rangle\psi$ and $\langle\langle G^{\preceq_i}\rangle\rangle\psi$ correspond to formulas of the form $\bigvee_{A \subseteq \mathsf{act}(G)}[\bigwedge \Phi(A, G)]\psi$ and $\bigvee_{A \subseteq \mathsf{act}(G)}([\bigwedge \Phi(A, G)]\psi \wedge [\bigwedge \Phi(A, G)]^{\preceq_i}\top)$, respectively – for $\bigwedge \Phi(A, G)$ as in the proof of Observation 3.11. During the filtration, the underlying agents model is not changed and therefore the truth of formulas of the form $\langle\langle G\rangle\rangle\alpha$ is preserved.

Theorem 4.3 *Every satisfiable $\varphi \in \mathcal{L}^{\not\prec}_{cla+p}$ is also satisfiable in a MaSP with $\leq 2^{|\varphi|}$ many states.*

Proof. Given that $M^{\preceq}, s \vDash \varphi$ for some $M^{\preceq} = \langle S, Ac, (\rightarrow)_{A \subseteq Ac}, Ag, \mathsf{act}, \{\preceq_i\}_{i \in Ag}, V\rangle$, $s \in S$, we obtain $M^{f\preceq^f} = \langle S, Ac, (\rightarrow^f)_{A \subseteq Ac}, Ag, \mathsf{act}^f, \{\preceq^f_i\}_{i \in Ag}, V^f\rangle$ by filtrating $M^{\preceq}$ through $Sub(\varphi)$, where the accessibility relations for actions and preferences are defined as follows:

$|s| \rightarrow^f_A |t|$ iff

1. $\forall[\alpha]\psi \in Sub(\varphi)$ such that $A \vDash \alpha$: if $M^{\preceq}, s \vDash [\alpha]\psi$, then $M^{\preceq}, t \vDash \psi$,
2. $\forall[\alpha]^{\preceq_i}\top \in Sub(\varphi)$ such that $A \vDash \alpha$: if $M^{\preceq}, s \vDash [\alpha]^{\preceq_i}\top$, then $s \preceq_i t$,
3. $\forall\langle\langle G\rangle\rangle\psi \in Sub(\varphi)$ such that $A \vDash \bigwedge \Phi(A', G)$ for some $A' \subseteq \mathsf{act}(G)$: if $M^{\preceq}, s \vDash [\bigwedge \Phi(A', G)]\psi$, then $M^{\preceq}, t \vDash \psi$,
4. $\forall\langle\langle G^{\preceq_i}\rangle\rangle\psi \in Sub(\varphi)$ such that $A \vDash \bigwedge \Phi(A', G)$ for some $A' \subseteq \mathsf{act}(G)$: if $M^{\preceq}, s \vDash [\bigwedge \Phi(A', G)]\psi$ and $M^{\preceq}, s \vDash [\bigwedge \Phi(A', G)]^{\preceq_i}\top$, then $M^{\preceq}, t \vDash \psi$ and $s \preceq_i t$.

$|s| \preceq^f_i |t|$ iff

1. $\forall\Diamond^{\preceq_i}\psi \in Sub(\varphi)$: if $M^{\preceq}, t \vDash \psi \vee \Diamond^{\preceq_i}\psi$ then $M^{\preceq}, s \vDash \Diamond^{\preceq_i}\psi$,
2. If there is some $[\alpha]^{\preceq_i}\top \in Sub(\varphi)$, then $s \preceq_i t$,
3. If there is some $\langle\langle G^{\preceq_i}\rangle\rangle\psi \in Sub(\varphi)$, then $s \preceq_i t$.

$V^f(p) := \{|s| | M, s \vDash p\}$, for all propositional letters $p \in Sub(\varphi)$. We show by induction that for all $\psi \in Sub(\varphi)$ and $s \in S$ it holds that $M^{\preceq}, s \vDash \psi$ iff $M^{\preceq^f}, |s| \vDash \psi$. This follows from the definitions of $(\rightarrow^f)_{A \subseteq Ac}$ and $\preceq^f$, and the fact that the filtration does not change the underlying agents model.

By definition of $S_{Sub(\varphi)}$, $|S_{Sub(\varphi)}| \leq 2^{|\varphi|}$. QED

Applying the constructions in the proofs of Theorems 4.1, 4.2 and 4.3 successively, we obtain the following:

Corollary 4.4 *Every satisfiable formula $\varphi \in \mathcal{L}^{\not\prec}_{cla+p}$ is satisfiable in a MaSP of size exponential in $|\varphi|$ satisfying the* *conditions* $|Ag| \leq | \ Ag(\varphi)| + 1$ *and* $|Ac| \leq | \ Ac(\varphi)| + \sum_{\langle\langle G \rangle\rangle \psi \in Sub(\varphi)} |G| + (\sum_{\langle\langle G^{\preceq_i} \rangle\rangle \psi \in Sub(\varphi)} |G|) + 1$.

Having non-deterministicly guessed a model of size exponential in $|\varphi|$, we can check in time exponential in $|\varphi|$ whether this model satisfies φ. This then gives us a NExpTime upper bound.

Theorem 4.5 *The satisfiability problem of CLA+P*$^{\not\prec}$ *is in NExpTime.*

Proof. Given φ, we non-deterministically choose a model $M^{\preceq}$ of size exponential in $|\varphi|$ satisfying the conditions $|Ag| \leq |Ag(\varphi)| + 1$ and $|Ac| \leq |Ac(\varphi)| + \sum_{\langle\langle G \rangle\rangle \psi \in Sub(\varphi)} |G| + (\sum_{\langle\langle G^{\preceq_i} \rangle\rangle \psi \in Sub(\varphi)} |G|) + 1$. Then, given this model, we can check in time $\mathcal{O}(|\varphi|||M^{\preceq}||)$, for $||M^{\preceq}||$ being the size of $M^{\preceq}$, whether $M^{\preceq}$ satisfies φ. Thus, given a model of size exponential in $|\varphi|$ that also satisfies the conditions on its sets of agents and actions explained earlier, it can be computed in time exponential in $|\varphi|$ whether it satisfies φ. Since it can be checked in time linear in the size of the model whether it is a proper MaSP, we conclude that the satisfiability problem of CLA+P$^{\not\prec}$ is in NExpTime. QED

This section has shown that the satisfiability problem of CLA+P$^{\not\prec}$ is in NExpTime. As the reader might expect, it has however a rather high computational complexity caused by the environment logic. The next section shows that the satisfiability problem of the environment logic is already NExpTime-hard and therefore adding agents as actors and preferences does not increase the complexity significantly.

4.2. Lower Bound

Establishing a NExpTime lower bound for the satisfiability problem of CLA+P can be done by reducing that of the Boolean modal logic $\mathbf{K}_m^{\neg\cup}$ [10] to it, which is known to be NExpTime-complete [9].

Models of $\mathbf{K}_m^{\neg\cup}$ have a set of accesibility relations $R_1, \ldots, R_m$ and the associated language $\mathcal{L}_m^{\neg\cup}$ that is used for talking about the models contains corresponding basic modal parameters $\mathcal{R}_1, \ldots, \mathcal{R}_m$. Using the operations $\neg$ and $\cup$, more complex modal parameters can be built. The modalities then run along the corresponding sets of accessibility relations in the models.

Then a model M of $\mathbf{K}_m^{\neg\cup}$ with set of states W can be translated into an environment model $\tau_1(M)$ with set of states $W \cup \{u\}$ for some newly introduced state u and set of actions $Ac = \{a_1, \ldots, a_m\}$. The accessibility relation $(\rightarrow)_{A \subseteq Ac}$ is defined as

$$w \rightarrow_A w' \text{ iff } A = \{a_i | (w, w') \in \mathcal{R}_i\} \text{ or } w' = u.$$

Thus, u is accessible from everywhere by any transition $\rightarrow_A$. This ensures that each $\rightarrow_A$ is serial. Formulas $\varphi \in \mathcal{L}_m^{\neg\cup}$ can be translated into $\tau_2(\varphi) \in \mathcal{L}_e$ in a straightforward way: Inside the modalities, modal parameters $\mathcal{R}_i$ are translated into atomic actions a_i and complex parameters are translated into action expressions ($\neg$ and $\cup$ correspond to $\neg$ and $\vee$ respectively).

Theorem 4.6 *For any formula* $\varphi \in \mathcal{L}_m^{\neg\cup}$ *and any model* M *of* $\mathbf{K}_m^{\neg\cup}$*, for any state* w *in* M*:*

$$M, w \models \varphi \textit{ iff } \tau_1(M), w \models \tau_2(\varphi).$$

The reduction is polynomial and hence the satisfiability problems of CLA+P$^{\not\prec}$ and CLA+P are NExpTime-hard.

Corollary 4.7 *The satisfiability problem of CLA+P*$^{\not\prec}$ *is NExpTime-complete.*

This section has shown that the satisfiability problem of CLA+P without strict preferences is NExpTime-complete. This rather high complexity is due to the environment logic which itself is already NExpTime-complete. Adding agents with nonstrict preferences as actors to the environment logic does not increase the complexity significantly. Due to the undefinability of irreflexivity extending the complexity results of CLA+P$^{\not\prec}$ to full CLA+P cannot be done using standard techniques such as filtration as we did in Theorem 4.3.

5. Conclusions and Future Work

We developed a modular modal logic that allows for reasoning about the coalitional power of agents, actions and their effects, and agents' preferences. The current approach is based on the logic CLA [14] which is combined with a preference logic [15, 16]. The resulting logic CLA+P, which is shown to be sound and complete, allows us to make explicit how groups of agents can achieve certain results. Additionally, we can express how a group can achieve that a transition takes place that is an improvement for some agent. In the framework of CLA+P, it can be expressed how the abilities to perform certain actions are distributed among the agents, what are the effects of the concurrent performance of these actions and what are the agents' preferences over those effects. Moreover, in CLA+P, we can distinguish between different ways how groups can achieve some result – not only with respect to the actions that lead to some result, but also with respect to the preferences. We can for instance express that a group can achieve some result in a way that is 'good' for its members in the sense that after the achievement all of them are better off. Thus, CLA+P provides a framework for reasoning about interactive situations in an explicit way that gives us more insights into the cooperative abilities of agents.

The satisfiability problem of CLA+P without strict preferences is shown to be NExpTime-complete. It remains to be investigated whether the same holds for CLA+P. From a computational viewpoint, it seems to be appealing to change the environment logic in order to decrease computational complexity.

There are two immediate ways to extend the logic developed in this paper. First of all, we can follow the ideas of Ågotnes et al. [3] and add a restricted form of quantification that allows statements of the form $\langle\langle P^{\preceq_i}\rangle\rangle\psi$ saying that there is some group G that has property P and $\langle\langle G^{\preceq_i}\rangle\rangle\psi$.

Moreover, it might be promising to develop a cooperation logic with actions and preferences based on a logic for reasoning about complex plans such as the the one developed by Gerbrandy and Sauro [7].

Acknowledgements

I am very grateful to Ulle Endriss, Eric Pacuit, Luigi Sauro, Johan van Benthem and Olivier Roy for their help and their useful comments.

References

[1] T. Ågotnes, P. E. Dunne, W. van der Hoek, and M. Wooldridge. Logics for coalitional games. In *LORI '07: Proceedings of the Workshop on Logic, Rationality and Interaction*, Beijing, China, 2007. to appear.

[2] T. Ågotnes, W. van der Hoek, and M. Wooldridge. On the logic of coalitional games. In *AAMAS '06: Proceedings of the fifth International Joint Conference on Autonomous Agents and Multi-agent Systems*, pages 153–160, Hakodate, Japan, 2006.

[3] T. Ågotnes, W. van der Hoek, and M. Wooldridge. Quantified coalition logic. In *IJCAI '07: Proceedings of the twentieth international joint conference on Artificial Intelligence*, pages 1181–1186, Hyderabad, India, 2007.

[4] R. Alur, T. A. Henzinger, and O. Kupferman. Alternating-time temporal logic. *Lecture Notes in Computer Science*, 1536:23–60, 1998.

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

[6] S. Borgo. Coalitions in action logic. In *IJCAI '07: Proceedings of the twentieth international joint conference on Artificial Intelligence*, pages 1822–1827, Hyderabad, India, 2007.

[7] J. Gerbrandy and L. Sauro. Plans in cooperation logic: a modular approach. In *Proceedings of the IJCAI Workshop on Nonmontonic Reasoning, Action and Change (NRAC 2007)*, Hyderabad (India), 2007.

[8] L. Kurzen. Logics for Cooperation, Actions and Preferences. Master's thesis, Universiteit van Amsterdam, the Netherlands, 2007.

[9] C. Lutz and U. Sattler. The complexity of reasoning with boolean modal logics. In F. Wolter, H. Wansing, M. de Rijke, and M. Zakharyaschev, editors, *Advances in Modal Logics Volume 3*. CSLI Publications, Stanford, 2001.

[10] C. Lutz, U. Sattler, and F. Wolter. Modal logics and the two-variable fragment. In *Annual Conference of the European Association for Computer Science Logic CSL'01*, LNCS, Paris, France, 2001. Springer Verlag.

[11] M. J. Osborne and A. Rubinstein. *A course in game theory*. MIT Press, Cambridge, MA, 1994.

[12] M. Pauly. A modal logic for coalitional power in games. *Journal of Logic and Computation*, 12(1):149–166, 2002.

[13] M. Pauly. On the complexity of coalitional reasoning. *International Game Theory Review*, 4:237–254, 2002.

[14] L. Sauro, J. Gerbrandy, W. van der Hoek, and M. Wooldridge. Reasoning about action and cooperation. In *AAMAS '06: Proceedings of the fifth International Joint Conference on Autonomous Agents and Multi-agent Systems*, pages 185–192, Hakodate, Japan, 2006.

[15] J. van Benthem, O. Roy, and P. Girard. Everything else being equal: A modal logic approach to ceteris paribus preferences, 2007.

[16] J. van Benthem, S. van Otterloo, and O. Roy. Preference logic, conditionals and solution concepts in games. In *Festschrift for Krister Segerberg*. University of Uppsala, 2005.

PDL$_{||}$and its relation to PDL

Fahad Khan
University of Nottingham
afk@cs.nott.ac.uk

Abstract

In this report we examine results pertaining to Karl Abrahamson's PDL$_{||}$, namely PDL with an interleaving operator, $\|$, with respect to an agent programming point of view. We first establish its usefulness in such contexts, before defining a syntax and semantics for the logic, looking at its relation to the regular expression shuffle operator and to PDL itself. We also look at the practical implications of this relation between PDL andPDL$_{||}$and of its PDL$_{||}$'s relation to BPDL another logic defined by Abrahamson over a quarter of a century ago.

1. Introduction

Interleaved PDL, denoted as PDL$_{||}$, was first defined by Karl Abrahamson in his 1980 PhD thesis, *Decidability and expressiveness of logics of processes*[2], as an extension of PDL (propositional dynamic logic) with a new operator, the interleaving operator $\|$. In fact, the $\|$ operator makes a useful addition to the regular four PDL operators[1]allowing the easy expression of the interleaving of two or more PDL$_{||}$programs – and hence facilitating reasoning about the effects of such interleavings. We give more details of the interleaving operator below, but in order to illustrate why reasoning about the interleaving of representations of programs is of particular importance, at least from an agent programming point of view, consider the example of SimpleAPL, a programming language explored in *A logic of agent programs*[3] and used to implement a particular model of basic agents with beliefs, goals, and plans[2].

In SimpleAPL an agent has beliefs, whose role it is to encode various aspects of its environment, and goals, which encode representations towards the realisation of which the agent works by adopting plans which are selected in turn via planning goal rules. Both beliefs and goals are represented by literals; plans on the other hand are composites built up from a set of basic actions via sequencing, conditional choice and conditional iteration operators. In [3] Alechina et al. detail two execution strategies for executing an agent program, the first of which allows either for an agent with no plan to select a planning goal rule and choose a single plan, or for an agent with a plan to execute the next step in the single plan which it carries; in the second an agent can amass a number of plans at any single juncture, consequently interleaving the execution of these plans or selecting another planning goal rule. So for example, with the first strategy, an agent would have to carry out the plans 'make coffee' and 'make toast' one after the other – potentially leaving it with a cold cup of coffee or piece of toast – whereas the second strategy would allow for multitasking as it were, allowing it to carry out actions associated with either of these plans in their correct order within the plan, but otherwise in whatever order was preferred.

Furthermore, Alechina et al., develop a sound and complete variant of PDL with which they are able to reason about the safety and liveness properties of agent programs in SimpleAPL. Crucially it turns out that the interleaved strategy admits of a straightforward formulation through the use of the $\|$ operator – which, as we will show, can ultimately be eliminated, thereby allowing any formula of PDL$_{||}$to be equivalently formulated in PDL.

It is clear, at least from the foregoing example, that the usefulness of PDL$_{||}$in the context of agent programming and modelling lies, among other things, in the fact that it allows for the succint expression of agent program execution strategies that incorporate the interleaving of agent plans (incidentally Abrahamson's own original motivation for defining PDL$_{||}$in [2] related to modelling and formally verifying claims about the behaviour of concurrent computer programs).

Given this agent based motivation, the purpose of this report is to collect and elucidate some important relevant technical results involving PDL$_{||}$. To summarise the rest of the paper, we begin by elaborating on the syntax and semantics of PDL$_{||}$as well as on the correspondence between the

[1] Recall that regular PDL programs are built up using nondeterministic union($\cup$), concatenation(;), iteration(*) and query(?) operators.

[2] Note that SimpleAPL is, as the name suggests, a simplified fragment of the more extensive agent programming language, 3APL. See [3] for more details.

shuffle regular expression operator and the interleaving operator and how we can use this equivalence to eliminate the interleaving operator from PDL programs, before describing how we can improve on the size of the resulting formula by using BPDL – another logic defined by Abrahamson in [2]. Finally in the conclusion we briefly consider some of the possible future directions for research suggested by the preceeding results.

2. An Inductive Definition of PDL$_{\|}$Syntax and a PDL$_{\|}$Semantics

The following series of definitions serve to define the syntax and semantics of PDL$_{\|}$.

Definition 2.1 Given a fixed set of proposition symbols, $\Phi = \{p, q, ...\}$, we can inductively define the set of PDL$_{\|}$formulae as follows:

- each proposition symbol $p \in \Phi$ is a formula,
- if ϕ and ψ are formulae, then so too are $\neg\phi$, $\phi \wedge \psi$ and $\langle\rho\rangle\phi$ where $\rho \in \Psi$ is a program.

Assuming a fixed set of basic programs $\Psi_0 = \{a, b, ...\}$ we inductively construct the set of programs, Ψ, used in the previous definition, in the following manner:

- each basic program $a \in \Psi_0$ is a program,
- if α and β are programs, then so too are $\alpha \cup \beta$, $\alpha; \beta$, α^* and $\alpha \parallel \beta$, where the latter is the interleaving operator,
- if ϕ is a formula of PDL$_{\|}$then $\phi?$ is a program.

◁

We now come to define models for PDL$_{\|}$:

Definition 2.2 Let M be a structure, $M = (W, \tau, V)$, then M is a model for PDL$_{\|}$if:

- W is a set of states,
- $V(p) \subseteq W$ is a function that for each $p \in \Phi$ gives us the set of states in W at which p holds. We can extend this in the obvious way so that $V(\phi)$ gives us the set of states where the PDL$_{\|}$formula ϕ holds: given the PDL$_{\|}$formulae ϕ, ψ and the program ρ, $V(\neg\phi) = W - V(\phi)$, $V(\phi \vee \psi) = V(\phi) \cup V(\psi)$ and $V(\langle\rho\rangle\phi)$ equals the set $U \subseteq W$ consisting of all the states of $u \in W$ such that there exists a computation sequence $\sigma \in \tau(\rho)$ (we define τ below) where either $(u, u_1), ..., (u_n, v)$ where $v \in V(\phi)$ and σ is a legal sequence, or where $\sigma = \epsilon$ and $u \in V(\phi)$. Note that by legal computational sequences we are referring to sequences ρ such that whenever $(s_1, s_2)(s_3, s_4)$ is a subword of ρ, then $s_2 = s_3$,
- $\tau(a) \subseteq (W \times W)$ gives us the set of state transitions for a. We can extend this inductively to give us a set of paths $\tau(\rho) \subseteq (W \times W)^*$ corresponding to any PDL$_{\|}$program expression ρ in M:
  - $\tau(\phi?) = \{(u, u) : u \in V(\phi)\}$,
  - $\tau(\rho_1 \cup \rho_2) = \{z : z \in \tau(\rho_1) \cup \tau(\rho_2)\}$,
  - $\tau(\rho_1; \rho_2) = \{z_1 \circ z_2 : z_1 \in \tau(\rho_1),\ z_2 \in \tau(\rho_2)\}$, where $\circ$ is a concatenation of paths operator,
  - $\tau(\rho^*)$ is the set of all paths consisting of zero or finitely many concatenations of paths in $\tau(\rho)$,
  - $\tau(\rho_1 \parallel \rho_2)$ is the set of all paths obtained by interleaving atomic actions and tests from $\tau(\rho_1)$ and $\tau(\rho_2)$.

  Note that the set of paths $\tau(\rho) \subseteq (W \times W)^*$ may contain non-legal sequences – in fact, to do otherwise would be to place a severe restriction on our ability to interleave sets of paths.

◁

3. Shuffling and Interleaving

From the foregoing series of definitions it is easy to see that the program constructors $\cup, ;,$ and * correspond to the regular expression (RE) operators $+, \cdot$, and *, respectively. However, given that under our definition of PDL$_{\|}$basic programs are indivisible, we are in a position to define another RE operator, shuffle, which corresponds to our interleaving operator and which we will also denote using $\|$.

Let $x, y \in \Sigma^*$, where Σ is a finite alphabet, and x, y are strings over Σ. Then the shuffle of x and y, namely, the set $x \parallel y$, is defined (in for example, [5]) as:

- $\epsilon \parallel y = \{y\}$,
- $x \parallel \epsilon = \{x\}$,
- $xa \parallel yb = (x \parallel yb) \cdot \{a\} \cup (xa \parallel y) \cdot \{b\}$.

Furthermore we define the shuffle of two languages X, Y as follows:

$$X \parallel Y = \bigcup_{\substack{x \in X \\ y \in Y}} x \parallel y.$$

Since for any two RE's α and β, we intend the language $L(\alpha \parallel \beta)$ to accept all strings x such that x belongs to the shuffle of the languages $L(\alpha)$ and $L(\beta)$ – where $L(\alpha)$ and $L(\beta)$ are the languages of α, β respectively – we define $L(\alpha \parallel \beta)$ as $L(\alpha) \parallel L(\beta)$.

As an example, take the shuffle of the two sets $\{ab\}$ and $\{cd\}$, namely $\{ab\} \parallel \{cd\}$, which gives us the set $\{abcd, acbd, acdb, cabd, cadb, cdab\}$ or the shuffle of the two RE's a^* and $(b;c)$, $a^* \parallel (b;c)$, which results in the set of strings of arbitrary length including the string bc in which b and c are inserted within a series of one or more repetitions of the character a: in other words the set of strings satisfying the RE $a^*(b)a^*(c)a^*$.

Given the correspondence of shuffle with our interleaving operator (indeed we will use the terms 'shuffle operator' and 'interleaving operator' interchangeably from hereon in) and the fact, which we will presently demonstrate, that any instance of the shuffle operator in a given regular expression can be eliminated, it is clear that we can translate any PDL$_{\parallel}$formula ϕ into a PDL formula ϕ'. And we do this by simply replacing each program ρ that occurs in a subexpression $\langle\rho\rangle\chi$ of ϕ with its equivalent RE which we consequently translate, using the method we summarise below, from a shuffle RE into a shuffle free RE before translating it back into a program ρ' and replacing the subexpression $\langle\rho\rangle\chi$ with its equivalent $\langle\rho'\rangle\chi$.

In fact, as we now show, we can directly translate any formula of PDL$_{\parallel}$containing no instances of the * operator into PDL using RE equivalences – the other method which we detail below and which can be applied to any formula of PDL$_{\parallel}$involves a detour into automata theory. Along with the usual RE equivalences this direct translation requires the following regular expression equivalences (the proofs are trivial and are therefore omitted):

Proposition 3.1
- *(i) For all regular expressions α,β, and γ, we have that $\alpha \parallel (\beta + \gamma) \equiv (\alpha \parallel \beta) + (\alpha \parallel \gamma)$ and $(\alpha + \beta) \parallel \gamma \equiv (\alpha \parallel \gamma) + (\beta \parallel \gamma)$.*
- *(ii) For all strings $x, y \in \Sigma^*$ and $a, b \in \Sigma$, where Σ is some alphabet we have that $xa \parallel yb \equiv (x \parallel yb)a + (xa \parallel y)b$.*

Now given a formula χ in *-free PDL$_{\parallel}$we can use the following algorithm to remove the instances of shuffle embedded in χ:

- Step 1: list all of the subformulae ϕ of χ,
- Step 2: let ϕ' be a maximal such subformula of χ, maximal in that it does not contain a subformula of the form $\lambda_1 \parallel \lambda_2$ and there is no shuffle free subformula of χ of which it is a proper subformula,
- Step 3: rewrite ϕ' in the form $\phi_1 + \phi_2 + ... + \phi_n$, where each ϕ_i is a concatenation of characters. We can do this through repeated application of the regular expression equivalences $\alpha(\beta + \gamma) \equiv \alpha\beta + \alpha\gamma$ and $(\alpha + \beta)\gamma \equiv \alpha\gamma + \beta\gamma$. Replace each such maximal subfomula ϕ' with its rewriting,
- Step 4: now we rewrite each subformula ψ of the form $\psi = x \parallel y$ where x and y are shuffle free (and which thanks to our previous operations are in the form we require) in terms of its equivalent of the form $\psi' = \psi_1 + ...\psi_k$ where each ψ_i is of the form $x_i \parallel y_i$ where x_i, y_i contain only concatenations of symbols via repeated applications of the the equivalences proved in Proposition 3.1 (i). Next we get rid of shuffle from each ψ_i by rewriting ψ_i using the equivalence proved in Proposition 3.1 (ii),
- we end up with a new formula χ' with which we can repeat the previous steps until we've gotten rid of all instances of shuffle.

Note that each rewriting of a subformula ψ_i of size n using the equivalence proved in Proposition 3.1(ii) in Stage 4 results in a formula ψ_i' of size $O(2^{p(n)})$ – meaning that the method we will now detail for the elimination of shuffle in any formula of PDL$_{\parallel}$and which guarantees us a double exponential bound on the size of the formula resulting from the translation is preferable in most cases.

In fact this next method (also known as the "brute force" method) seems to be the most straightforward means of translating any regular expression, α, containing one or more instances of the shuffle operator into an equivalent RE constructed solely in terms of the RE operators $\cup$,* and ;, and it proceeds in two steps. We begin by translating α into a nondeterministic finite automaton (NFA) M using a special cross-product construction, this is the first step; the second step consists of translating M back into an RE. Unfortunately the combination of these two operations entails, in the worst case, a double exponential blowup in the size of the resulting RE. We now describe in greater detail both of the steps comprising this translation method.

For the first step, we proceed inductively starting with an instance of a regular expression $\alpha = \alpha_1 \parallel \alpha_2$ consisting of a shuffle operator applied to two shuffle free RE's α_1 and α_2, as our base case.

Now, we can convert the two shuffle free RE's α_1 and α_2 into two NFA's $M_1 = (Q_1, \Sigma, \delta_1, s_1, F_1)$ and $M_2 = (Q_2, \Sigma, \delta_2, s_2, F_2)$, respectively where $L(\alpha_1) = L(M_1)$ and $L(\alpha_2) = L(M_2)$. Note that each of these conversions gives us an NFA that is linear in the size of our original RE, since the conversion algorithm we will use – and which can be found in, for example, Hopcroft and Ullman's famous Automata textbook [4], and in Kozen's textbook on the subject [5] – will only add 2 states for each subexpression of our original RE.

The important thing for us now is to be able to show that the set $L(M_1) \parallel L(M_2)(= L(\alpha_1) \parallel L(\alpha_2))$ can be converted into an NFA M such that $L(M) = (M_1) \parallel L(M_2)$ which we will then convert back into an RE. Obviously it would be simpler – though perhaps not preferable in terms of the size of the resulting formula – if we had a way of generating $L(\alpha_1) \parallel L(\alpha_2)$ directly via an RE as we did with the translation of the $^*-$free fragment of PDL$_\|$, rather than by taking this circuitous route.

To generate a machine whose language is $L(M_1) \parallel L(M_2)$ we use the following cross-product construction on M_1 and M_2, the result of which is an NFA M_3 which accepts a string x if and only if $x = x_1 \parallel x_2$, where $x_1 \in L(M_1), x_2 \in L(M_2)$:

$M_3 = (Q_1 \times Q_2, \Sigma, \delta, (s_1, s_2), F)$, where the transition function δ is defined as $\delta((q_1, q_2), a) = (\delta_1(q_1, a) \times \{q_2\}) \cup (\{q_1\} \times \delta_2(q_2, a))$, and the set of accepting states F is defined as $F = \{(q_1, q_2) \in Q_1 \times Q_2 : q_1 \in F_1, q_2 \in F_2\}$.

A proof of the correctness of the construction can be easily produced on the template of the proofs given for the equivalence of NFAs and DFAs in terms of the class of languages accepted by either, in for example Kozen [5].

Now, given an RE α containing a number of nested $\|$ operators, we can iterate through the subexpressions of α until we reach subexpressions α_i that match our base case, building up a series of NFA's which we can combine either using our cross product constructor or via the rules for building NFA's inductively from RE's – again as set out in for example [4] in the algorithm for converting an RE into an NFA. The upshot is that we have defined an NFA M_3 such that $L(M_3) = L(M_1) \parallel L(M_2) = L(\alpha_1) \parallel L(\alpha_2) = L(\alpha_1 \parallel \alpha_2)$.

Sadly in the worst case this means our NFA M is exponential in the size of our original RE α, i.e., the size of M_3 is $2^{(O|r|)}$ where $r = |\alpha|$. To understand why this is so consider that each basic program constituent, a of α can be translated into an NFA of size 2 and given $r = |\alpha|$ where by necessity $r > 3$, we may potentially need to use the cross product construction kr times, where $1 \leq k \leq \frac{r}{2}$. (Meyer and Stockmeyer use this fact to prove an upper bound for the complexity of PDL$_\|$'s satisfiability problem [6]).

Worse is to come. It seems that the best known algorithms we have for translating an n state NFA into an RE entail, in the worst case, an exponential blow up in the size of the resulting RE, e.g., the algorithm given in [4] gives us an RE of size $O(n^3 4^n)$. This means that after having translated our original RE α of size r into an NFA of size $O(2^r)$ we then end up with an RE of size $O(2^{2^r})$.

4. BPDL

Abrahamson, who first defined PDL$_\|$in his PhD thesis [2], writes in the self same that "[a]ny axiom system for PDL$_\|$which ultimately relies on reducing away concurrency by expressing it in terms of $\cup$,;, or *... is misguided." He suggests introducing auxiliary variables into PDL in order to improve on the double exponential size of the formula that results from the brute force method. In fact, it is relatively simple to see how we can do this in relation to BPDL, an extension of normal PDL that incorporates boolean variables and which is also defined by Abrahamson in his PhD thesis.

BPDL structures feature an additional set, Q, of boolean variables, which we refer to when defining the set of well-defined BPDL formulae – and note that these boolean variables are treated completely separately from the propositional symbols. The definition of the syntax of BPDL is similar to that for PDL$_\|$.

Definition 4.1 Given a fixed set of proposition symbols, $\Psi_0 = \{p, q, ...\}$ and a fixed set of boolean variables $Q = \{x, y, ...\}$, we can construct the set of BPDL programs as follows:

- each basic program $a \in \Psi_0$ is a program,
- for each variable $x \in Q, \uparrow x$ and $\downarrow x$ are programs,
- if α and β are programs, then so too are $\alpha \cup \beta, \alpha; \beta$, and α^*,
- if ϕ is a formula of BPDL then ϕ? is a program.

◁

The set of formulae of BPDL are defined as for PDL$_\|$, again with reference to a set Φ of propositional variables. We can now define models for BPDL.

Definition 4.2 Let M be a structure such that $M = (W, V_B, \tau_B, Q)$, then M is a model for BPDL if

- W is a set of states,
- Q is a set of boolean variables,
- $V_B(p) \subseteq \wp(W \times \wp(Q))$, where $\wp(Q)$ denotes the power set of Q, is a function that gives us, for each $p \in \Phi$, the cross product with the power set of Q of the set of states in W at which p holds, namely, the set $V(p)$ as defined in Definition 2.2, i.e., $V_B(p) = V(p) \times \wp(Q)$. Additionally, for each $x \in Q, V_B(x) = W \times \{S \subseteq Q : x \in S\}$.

  We extend this function to all formulae of BPDL inductively: given the formulae ϕ, ψ and the program ρ with $V_B(\phi), V_B(\psi) \subseteq W \times \wp(Q)$ and $\tau_B(\rho) \subseteq (W \times \wp(Q))^2$ (we will define τ_B below), the definition runs as follows:

  - $V_B(\neg\phi) = (W \times \wp(Q)) - V_B(\phi)$,
  - $V_B(\phi \vee \psi) = V_B(\phi) \cup V_B(\psi)$,
  - $V_B(\langle\rho\rangle)\phi = \{(u, S) \in (W \times \wp(Q))$: there exists $v \in W, T \subseteq Q$ such that $((u, S), (w, T)) \in \tau_B(\rho)$ and $(w, T) \in V_B(\phi)\}$

- $\tau_B(a) \subseteq \wp((W \times \wp(Q))^2)$, is a function that gives us, for each $a \in \Psi_0$, the set $\{((u,S),(w,S)) : (u,w) \in \tau(a), S \subseteq W\}$ where $\tau(a) \subseteq (W \times W)^*$, the set of state transitions of a is defined similarly to the function τ of Definition 2.2.

  Additionally, for each $x \in Q$ we have that $\tau_B(\uparrow x) = \{(u,S),(u,S') \in (W \times \wp(Q))^2 : S' = S \cup \{x\}\}$ and $\tau_B(\downarrow x) = \{(u,S),(u,S') \in (W \times \wp(Q))^2 : S' = S - \{x\}\}$. For programs α and β we define $\alpha \cup \beta, \alpha;\beta$ and α^* as in Definition 2.2. For any BPDL formula ϕ we define $\tau_B(\phi?)$ as $\{((u,S),(u,S)) \in (W \times \wp(Q))^2 : (u,S) \in V_B(\phi)\}$.

$\triangleleft$

It turns out that adding Boolean variables to PDL gives a strong boost to the expressiveness of the resulting language. For example, we can represent any integer in the range $0, ..., 2^{(n-1)}$ using just n Boolean variables. It is also routine to write programs of length $O(n)$ that add, subtract or compare two such "n-bit" integers. However for our purposes the most important consequence of adding Boolean variables to PDL is that we are able to drastically improve on the double exponential overhead incurred by the shuffle translation method given above.

To see how this is possible note that any NFA M consisting of n nodes can be converted to a BPDL program of length $O(n \log n + c)$ where c is the combined length of the tests on the outgoing edges of each node in M[3]. Obviously, in many cases, this allows us to improve on the exponential size of the PDL program resulting from the usual method of translating NFA's to RE's. The translation proceeds by assigning a numbering to the states of the NFA and constructing a program of the form $S;(\bigcup_i T)^*;F?$ where the subprogram S sets a counter to the number of the initial state; T_i performs the action associated with state numbered i if the counter is in the state numbered i; and finally $F?$ checks whether the counter is in an accepting state.

The easiest way to see how this works is by recourse to an example, here the NFA M illustrated below as Figure 1.

We can easily model the action of M via the following program:

$$\begin{aligned}
&I := 1;\\
&((I = 1)?;(a?;I := 2) \cup (b?;I := 3);\\
&(I = 2)?;(a?;I := 4) \cup (b?;I := 2);\\
&(I = 3)?;(a?;I := 3) \cup (d?;I := 5);\\
&(I = 4)?;(c?;I := 6))^*;\\
&(I = 5)? \cup (I = 6)?
\end{aligned}$$

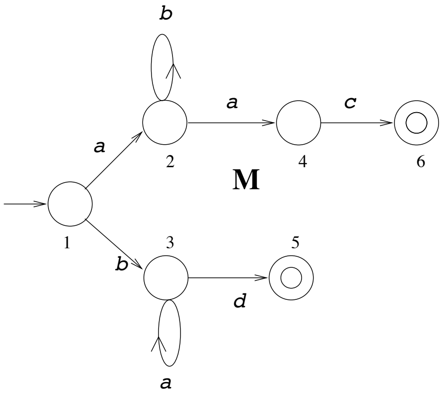


Figure 1. The NFA M

So given a program α, effectively an RE, of length n containing one or more instances of the $\|$ operator we can generate, as detailed above, an NFA M whose size is exponential in n such that $L(M) = L(\alpha)$. If we're willing to translate M into BPDL instead of PDL we end up with a formula of size $O(n2^n + d)$, where d is the length of all the tests in M. Usually this will offer a substantial improvement on the double exponential size of the formula resulting from the translation of M into PDL. Sadly we can't improve on our previous translation of PDL$_\|$formulae into PDL formulae by inserting an intermediate stage in which we translate our cross product NFA into a BPDL formula before further translating this into a PDL formula: Abrahamson determined in [1] that the translation of a formula from BPDL into PDL has an double exponential lower bound.

4.1. Conclusion

So in summary, we have surveyed a number of the main results concerning PDL$_\|$, most notably the fact that the addition of the $\|$ operator to PDL affords no increase in the expressiveness of the resulting language – however it does seem to give important, and indeed dramatic benefits in terms of the succintness of the formulae we can devise to describe the interleaving of two or more programs. We have looked at a conceptually straightforward method of translating the * free fragment of PDL$_\|$into PDL – straightforward in that it only makes use of regular expression equivalences – the original contribution of this report; and we have detailed the aptly named "brute force" method. However the double exponential size of the formulae resulting from the brute force method presents a substantial practical obstacle to the application of the interleaving operator in, for example, an agent programming context as described in the introduction. Abrahamson's BPDL – PDL enriched with binary variables – seems to provide one solution to this blow up in complexity, and indeed a further direction for investigation here is the possible use and development of BPDL tools for

[3] Note that we can exploit the nondeterminism of PDL (and hence BPDL) to model the nondeterminism of M.

the verification of agent programs.

In fact we could go further and keep the interleaving operator as a primitive: future work could investigate the effects of different axiomatisations of $\mathrm{PDL}_{||}$, or the creation of efficient $\mathrm{PDL}_{||}$ tools for verification purposes. An investigation into the kinds of regular expression featuring instances of the interleaving operator that admit of a more compact translation into a shuffle free regular expression would also yield useful practical results (it would be also interesting to see if there were other means of translating $\mathrm{PDL}_{||}$ into PDL than those we have described) . Of course further investigations could also centre on other kinds of agent execution strategies and the various logics which could be developed to describe them, including other extensions of PDL.

References

[1] K. R. Abrahamson. Boolean variables in regular expressions and finite automata. Technical report, Department of Computer Science, University of Washington, 1980.

[2] K. R. Abrahamson. *Decidability and expressiveness of logics of processes*. PhD thesis, Department of Computer Science, University of Washington, 1980.

[3] N. Alechina, M. Dastani, B. Logan, and J.-J. C. Meyer. A logic of agent programs. In *AAAI*, pages 795–800, 2007.

[4] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley, 1979.

[5] D. Kozen. *Automata and Computability*. Springer-Verlag, New York, 1997.

[6] A. J. Mayer and L. J. Stockmeyer. The complexity of PDL with interleaving. *Theoretical Computer Science*, 161(1&2):109–122, 1996.

It's all up to you:
A study of closed-world interaction and its optimal organization

Jan Broersen
Universiteit Utrecht
broersen@cs.uu.nl

Rosja Mastop
Universiteit Utrecht
rosja.mastop@phil.uu.nl

John-Jules Ch.Meyer
Univeristeit Utrecht
jj@cs.uu.nl

Paolo Turrini
Universiteit Utrecht
paolo@cs.uu.nl

Abstract

The aim of the work is to provide a deontic language to regulate closed-world interaction. To do so we use Coalition Logic enriched with a preference order over the outcomes of agents' choices. We take the perspective of a deontic language being agent-oriented, that is mandating choices that only belong to agents or coalitions. We formalize this intuition by identifying those interactions in which Nature does not play an active role. We apply the formal tools to games.

1 Introduction

Pauly's Coalition Logic has shown to be a sound formal tool to analyze the properties of strategic interactions. One issue left is to define in that language what the interesting properties of an interaction are, as possible for instance with regularity (it is never the case that a group of agents can determine that some variable p is true, while all the other agents can at the same time determine that p is false) or outcome monotonicity (if a coalition can force an outcome to lie in a set X, can also force an outcome to lie in all supersets of X).

If we think of a deontic logic as obligating agents to choose what it should ideally be the case, an intuitive property is that of *coherence*, a property of interaction that ensures players' abilities non to contradict one other and the empty coalition not to make active choices. With this property we can model a closed world interaction, such as those of a Coordination Game or of a Prisoner Dilemma, where all the outcomes are determined only by the choices of the agents that are present.

Our aim is to regulate multiagent interaction, mandating the optimal outcomes that result from the choices of the coalitions. By mandating we mean *the introduction of a normative constraint on individual and collective choices in a multiagent system.*

Column Row	White Dress	Black Dress
White Dress	$(3,3)$	$(0,0)$
Black Dress	$(0,0)$	$(3,3)$

Table 1. Clothing Conformity

We are specifically concerned with cases where the collective perspective is at odds with the individual perspective. That is, cases where we think that letting everybody pick their own best action regardless of other's interest gives a non-optimal result. The main question we are dealing with is then: how do we determine which norms, if any, are to be imposed?

To answer this question, the paper presents a language to talk about the conflict between coalitionally optimal and socially optimal choices in coherent interaction, and it expresses deontic notions referring to such circumstances.

1.0.1 Example

The toy example we would like to start with concerns conventional norms. Noms of this type are those in which players should conform to each other. In this situation (see Table 1), a legislator that wants to achieve the socially optimal state (players coordinate), should declare that discordant choices are forbidden, thereby labeling the combinations of moves (black, white), (white, black) as violations. As easy to see, these moves belong only to the set of agents taken together. A norm helping both players to reach an optimal outcome would be one that labels as violations combinations of discordant choices. However, in this kind of games Row will never know what is the best thing to choose, since the choice of Column is independent from his. In order to solve the problem a legislation should go beyond individual choice, by forcing the coalition made of Row and Column together to form and choose an efficient outcome.

Row \ Column	White Dress
White Dress	$(3,3)$
Black Dress	$(0,0)$

Table 2. Clothing Conformity Modified

1.1 Motivation

Provided the aim of regulating interactions, we ask ourselves whether it makes sense to construct a deontic logic for any type of game.

Suppose the environment (the coalition made by an empty set of agents) were active part of the game, and it could decide to transform the game of table 1 in the one of table 2.

What should then a legislator do? It is quite clear that imposing the agents to choose something should depend on the moves that are available to the players. But in a game in which Nature plays an active role, taking this statement serious would boil down to mentioning the environment in the deontic language, saying for instance "Nature should allow row to play only white" or "Nature should make it convenient for the grand coalition to form". If we think of a deontic language as a sort of "agent-oriented" language and as nature as a uncontrollable agent, the above mentioned statements do not make sense.

No legislator though would be in the condition of determining what moves Nature would play. Nature, unlike all the other players, does not have explicit preferences over the outcomes of the interaction and intuitively does not follow proper man made norms or orders. In order to have a regulation of the Multi Agent System, we need a proper agent-oriented deontic language and we should then avoid deontic statements that concern proper choices (i.e. those able to really modify the outcome of the game) to be carried out by Nature. This translates into ruling out all those interactions in which Nature plays an active role. In this paper we will pursue this idea formally, identifying all such interactions and axiomatizing their logic.

The paper is structured as follows: In the first part we introduce Coherent Coalition Logic, proving that Inability Of the Empty Coalition (IOEC) is not entailed by Pauly playable effectivity functions and it cannot even be defined in Coalition Logic. In the second part we discuss the axiomatization of the logic, giving a characterization of coherence in terms of global modality. In the third part we give application of the logic to the regulation of closed-world strategic interaction, constructing a deontic logic that tells coalitions how to behave in order to achieve socially desirable outcomes.

2 Coherent Interactions

We begin by defining the strategic abilities of agents and coalitions, introducing the concept of a dynamic Effectivity Function, adopted from [7]. Later on in the paper we will move from game forms to real games, by introducing the notion of preference.

Definition 2.1 [Dynamic Effectivity Function]
Given a finite set of agents Agt and a set of states W, a *dynamic Effectivity Function* is a function $E : W \rightarrow (2^{Agt} \rightarrow 2^{2^W})$.

◁

Any subset of Agt will henceforth be called a *coalition*.

For elements of W we use variables $u, v, w, \ldots$; for subsets of W we use variables $X, Y, Z, \ldots$; and for sets of subsets of W (i.e., elements of 2^{2^W}) we use variables $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \ldots$. The elements of W are called 'states' or 'worlds'; the subsets of Agt are called 'coalitions'; the sets of states $X \in E(w)(C)$ are called the 'choices' of coalition C in state w. The set $E(w)(C)$ is called the 'choice set' of C in w. The complement of a set $\overline{X}$ or of a choice set $\overline{\mathcal{X}}$ are calculated from the obvious domains.

A dynamic Effectivity Function assigns, in each world, to every coalition a set of sets of states. Intuitively, if $X \in E(w)(C)$ the coalition is said to be able to *force* or *determine* that the next state after w will be some member of the set X. If the coalition has this power, it can thus prevent that any state *not* in X will be the next state, but it might not be able to determine *which* state in X will be the next state. Possibly, some other coalition will have the power to refine the choice of C.

For studying closed-world interaction we isolate a set of minimally required properties, that constitute the class of *coherent Effectivity Functions*.

Definition 2.2 [Coherence]
For any world w, coalitions C,D and choice X, an Effectivity Function is *coherent* if it has the following properties:

1. coalition monotonicity: if $X \in E(w)(C)$ and $C \subseteq D$ then $X \in E(w)(D)$;
2. regularity: if $X \in E(w)(C)$ then $\overline{X} \notin E(w)(\overline{C})$;
3. outcome monotonicity: if $X \in E(w)(C)$ and $X \subseteq Y$ then $Y \in E(w)(C)$;
4. inability of the empty coalition (IOEC): $E(w)(\emptyset) = \{W\}$.

◁

The first property says that the ability of a coalition is preserved by enlarging the coalition. In this sense we do not allow new members to interfere with the preexistent capacities of a group of agents. The second property says that if a coalition is able to force the outcome of an interaction to lie in a particular set, then no possible combinations of moves by the other agents can prevent this to happen. We think that regularity is a key property to understand the meaning of ability. If an agent is properly able to do something this means that others have no means to prevent it. Outcome monotonicity is a property of all Effectivity Functions in CL, which is therefore a monotonic modal logic. It says that if a coalition is able to force the outcome of the interaction to lie in a particular set, then is also able to force the outcome to lie in all his supersets (see [7]). The last condition is IOEC, that forces the empty coalition relation to be universal. As noticed also in [2] with such a property the empty coalition cannot force non-trivial outcomes of a game.

One important class of Effectivity Functions are the *playable* ones, to which we will refer throughout the paper.

Definition 2.3 [Playability]

For any world w an Effectivity Function is *playable* if it has the following properties:

(1) $\emptyset \notin E(w)(C)$, for any C; (2) $W \in E(w)(C)$ for any C. (3) E is Agt-maximal, that is for any $X \subseteq W$, s.t. $W \setminus X \notin E(w)(\emptyset)$ implies $X \in E(w)(Agt)$ (4) E is superadditive, i.e. for $C \cap D = \emptyset$, if $X \in E(w)(C)$ and $Y \in E(w)(D)$ then $X \cap Y \in E(w)(C \cup D)$.

$\triangleleft$

The first condition imposes that games are nonempty, the second that coalitions can always choose the largest possible set, the third that the grand coalition of agents can do whatever not blocked by Nature, the fourth that coalitions can join their forces.

As proved in [7] [Theorem 2.27], nonempty strategic games exactly correspond to playable Effectivity Functions [1].

2.0.1 Playability and Coherence

What kind of interactions are coherent Effectivity Functions isolating?

In this respect, it is interesting to compare playable and coherent Effectivity Function, in order to understand the types of interactions we are considering.

Proposition 2.4 *Not all playable EF are coherent, and not all coherent EF are playable.*

Proof.

For the first part, take $W = \{x, y\}, Agt = \{i, j\}$ and the following Effectivity Function $E(\emptyset)(k) = E(\{i\})(k) = E(\{j\})(k) = E(Agt)(k) = \{W, W\backslash\{x\}\}$ for $k \in W$. Now it is just a matter of checking the conditions for playability.

For the second part take $W = \{x, y\}, Agt = \{i, j\}$ with $E(\emptyset)(k) = E(\{i\})(k) = E(\{j\})(k) = E(Agt)(k) = \{W\}$ for $k \in W$.

QED

Proposition 2.5 *Coherent Agt-maximal superadditive EF are playable.*

Proof.

It is a matter of checking the conditions of playability.

QED

3 On the axiomatization of Coherent Coalition Logic

In order to fully understand what sort of interactions we are investigating by using coherent effectivity functions we need to provide an axiomatization of their logic.

To do so we exploit some results due to Pauly and we adapt them to our framework. We recall first that Coalition Logic uses a modality $[C]\phi$ (to be read as "Coalition C can achieve ϕ") and it is interpreted in neighbourhood models with an outcome monotonic dynamic Effectivity Function as neighbourhood relation. The axioms of Coalition Logic extend propositional logic axiomatization with the Monotonicity axiom ($\phi \rightarrow \psi \Rightarrow [C]\phi \rightarrow [C]\psi$).

Consider the coalitional canonical model $C^* = ((W^*, E^*), V^*)$ and take $\overline{\phi} = \{w \in W^* | \phi \in w\}$, as the truth set of ϕ in the canonical model. The canonical relation (the rest is standard) is defined as

$$wE^*_C X \text{ iff } \exists\phi \text{ s.t } \overline{\phi} \subseteq X \text{ and } [C]\phi \in w$$

The set of formulas are closed under Modus Ponens and Monotonicity and the relation is easily proved to be monotonic. Moreover in [7] the following theorem [3.10] is proved: Every Coalition Logic Λ is sound and complete with respect to its canonical model C^*.

What we look for now is the a set of axioms and rules such that the corresponding maximally consistent sets generate a coherent Effectivity Function in the canonical models.

[1] The proof involves the definition of strategic game as a tuple $\langle N, \{\Sigma_i | i \in N\}, o, S\rangle$ where N is a set of players, each i being endowed with a set of strategies σ_i from Σ_i, an outcome function that returns the result of playing individual strategies at each of the states in S; the definition of α-Effectivity Function for a nonempty strategic game G, $E^{\alpha}_{G} : \wp(N) \rightarrow \wp\wp(S)$ defined as follows: $X \in E^{\alpha}_{G}$ iff $\exists\sigma_C \forall\sigma_{\overline{C}} o(\sigma_C; \sigma_{\overline{C}}) \in X$. The above mentioned theorem establishes that $E^{\alpha}_{G} = E$ in case E is playable and G is a nonempty strategic game.

Nevertheless IOEC is not definable in Coalition Logic. To see this it is important to notice that Coalition Logic is monotonic multimodal logic, and frame validity of formulas of monotonic modal logics is closed under taking disjoint unions. This is proven for modal satisfaction in [4][Definition 4.1, Proposition 4.2].

Definition 3.1 [[4] 4.1]

Let $M_i = (W_i, N_i, V_i), i \in I$, be a collection of disjoint models. Then we define their *disjoint union* as the model $\oplus M_i = (W, N, V)$ where $W = \bigcup_{i \in I} W_i, V(p) = \bigcup_{i \in I} V_i(p)$ and for $X \subseteq W, w \in W_i$,

$$X \in N(w) \text{ iff } X \cap W_i \in N_i(w)$$

$\triangleleft$

Without loss of generality, we can simply think of the monotonic modal logic with only the box for the empty coalition, and take frames instead of models.

Consider the following monotonic frames $F_0 = (W_0, N_0)$ and $F_1 = (W_1, N_1)$, with a domain W_j and a relation $N_j \subseteq W_j \times 2^{W_j} (j \in \{0,1\})$. Take $W_0 = \{w_0\}$, $W_1 = \{w_1\}$, $N_0(w_0) = \{w_0\}$ and $N_1(w_1) = \{w_1\}$. Now suppose ϕ is some formula true at a world w in a model $M' = (W', N', V')$ of a monotonic frame F' iff $[[\top]]^M$ is neighbour of w (if $wN'[[\top]]$) and nothing else is ($wN'X \Rightarrow X = [[\top]]$). We see that $M_0, w_0 \models \phi$ and $M_1, w_1 \models \phi$ for arbitrary M_i inside F_i ($i \in \{0,1\}$). From [4] we construct the disjoint union $\oplus(F_0, F_1) = (W, N)$ as defined. We see clearly that our formula ϕ is not true in the disjoint union, because the neighbourhoods of the single models are copied in the disjoint union even if they are smaller than the unit. We observe moreover that the disjoint union is monotonic. The conclusion is that the formula expressing inability of the empty coalition is not definable is monotonic modal language.

At this point it is clear why $[\emptyset]\phi \rightarrow [\emptyset](\phi \vee \psi)$ or also $[\emptyset]\top$ would not be decent axioms for Coherent Coalition Logic. They would both ensure the presence of the unit in the neighbourhood of $\emptyset$, but they would not say anything about the absence of all the other sets. We will give to this intuition a formal characterization, stating that in fact the ability of the empty coalition in Coherent Coalition Logic *is* a global modality.

3.1 Inability of the Empty Coalition is a global relation

We extend the language of Coalition Logic with a global modality, defined as follows:

$$M, w \models E\phi \Leftrightarrow \exists w' \in W \text{ s.t. } M, w' \models \phi$$

The dual $A\phi$ is defined as $\neg E \neg \phi$. We claim that in Coalition Logic plus the global modality IOEC is definable.

Proposition 3.2 *$A(\phi) \leftrightarrow [\emptyset]\phi$ defines IOEC. That is,*
$\models_C A(\phi) \leftrightarrow [\emptyset]\phi \Leftrightarrow E(w)(\emptyset) = \{W\}$ for every w in the coalitional frames C.

Proof. ($\Rightarrow$) Assume that $\models_C A\phi \leftrightarrow [\emptyset]\phi$ while not $E(w)(\emptyset) = \{W\}$ for every w in any frame F in the class of Coalitional Frames C. Then there is an F in which there is a w such that $E(w)(\emptyset) \neq \{W\}$. Notice that both W and $E(w)(\emptyset)$ are nonempty. So there is a $W' \neq W$ s.t $W' \in E(w)(\emptyset)$ and $W' \subset W$. Take an atom p to be true in all $w' \in W'$ and false in $W \setminus W'$. Now we have model M based on a coalitional frame C for which $M \not\models Ap \leftrightarrow [\emptyset]p$. Contradiction.

($\Leftarrow$) Assume $E(w)(\emptyset) = \{W\}$ for a given w in an arbitrary model M of a coalition frame in C, and that $w \models A\phi$. Then $[[\phi]]^M = W$ and $w \models [\emptyset]\phi$ follows. Assume now that $w \models [\emptyset]\phi$. It has to be the case that $[[\phi]]^M = W$ by assumption. So also that $w \models A\phi$, which concludes the proof.

QED

3.2 Axiomatization for the Global Modality plus a new inclusion axiom

The global relation induces an equivalence class in the models, therefore it is axiomatizable by an $S5$ modality interpreted on a global relation.

However this does not ensure that the underlying relation - that we indicate with $R_\exists$ - is globally connected. Global connectedness is not definable in basic modal language [1] [2].

As suggested in [1][p.417-418], taken a set of maximally consistent formulae Σ^+ we can simply take a generated submodel of the canonical model in such a way that the formulae in Σ^+ are invariant and the relation is (it follows by construction) a global relation.

Taken the canonical model $M^* = ((W^*, E^*, R^*_\exists), V^*)$, its submodel

$M^{*'} = ((W^{*'}, E^{*'}, R^{*'}_\exists), V^*)$ generated by Σ^+ using the $R^*_\exists$ relation should ensure that $R^{*'}_\exists = W^{*'} \times W^{*'}$.

Nevertheless in taking the generated submodel we should ensure that the coalitional relation is not altered. One way to do it is to guarantee that the canonical coalitional relation is included in the global relation and that the generated submodel for the second relation is also a generated submodel for the first.

We begin with some definitions:

[2] The reason is also the invariance under taking disjoint unions. This fact sheds light on the relation between IOEC and Global Relation, in fact now we see clearly that the ability of the empty coalition in Coherent Coalition Logic *is* a global modality.

Definition 3.3 [Generated Submodels for Basic Modal Language, [1]]

Let $M = (W, R, V)$ and $M' = (W', R', V')$ be two models; we say that M' is a submodel of M if $W \subseteq W'$, R' is the restriction of R to W', that is $R' = R \cap (W' \times W')$ and V' is the restriction of V to M'. We say that M' is a *generated submodel* of M ($M' \mapsto M$)if M' is a submodel of M and for all points the following closure condition holds:

$$\text{if } w \text{ is in } M' \text{ and } Rwv, \text{ then } v \text{ is in } M'$$

◁

Modal satisfaction is invariant under taking generated submodels [1].

Now the definition for monotonic modal logic.

Definition 3.4 [Generated Submodels for Monotonic Modal Language, [4]] Given a monotonic model M, M' is a submodel of M if $W' \subseteq W, V'(p) = V(p) \cap W'$ for p atomic, and $N' = N \cap (W' \times 2^{W'})$, that is

$$\forall s \in W' : N'(s) = \{X \subseteq W' | X \in N(s)\}$$

In neighbourhood semantics given M' submodel of M, M' is also a generated submodel of M if the identity mapping $i : W \rightarrow W'$ is a bounded morphism, that is, for all $w' \in W'$ and all $X \subseteq W$

$$i^{-1}[X] = X \cap W' \in N'(w') \text{ iff } X \in N(w')$$

◁

For all states of the generated submodels, truth of modal formulas is preserved [4].

Now the question is, is the submodel generated a maximally consistent set of formulas Σ^+ using the existential global modality relation (making the canonical model strongly connected with respect to this relation) also a generated submodel with respect to the coalitional relation?

The answer is: it depends on the extra axioms. Usually when we have a K and a global modality it is sufficient to include the diamond relation in the global modality relation. But we cannot simply have:

$$[C]\phi \rightarrow E\phi$$

because the coalitional canonical relation may cross $S5$ equivalence classes. Instead the good candidate for our attempt is just the following:

$$A\phi \leftrightarrow [\emptyset]\phi$$

We claim that taking a generated submodel with respect to the global relation, given this axiom, ensures the condition of taking also a generated submodel with respect to the neighbourhood modality.

This is easy to see, because all the neighbourhoods of all coalitions are of the form $X \subseteq W$ and W is covered by the global modality.

Proposition 3.5 *The axiom $A\phi \leftrightarrow [\emptyset]\phi$ guarantees inclusion of the canonical relation in the global relation*

Proof.

Take a maximally consistent set of formulas Σ^+ that extends a consistent set of formulas Σ according to the axioms and the rules that we have just defined (for the global and the coalitional modality). Suppose now $A\phi$ is in Σ^+ for some ϕ. This means that $W^* = [[\phi]]^{C^*}$. Now take a given $[C]\psi$ in the same maximally consistent set of formulas. This means that $[[\phi]]^{C^*} \in E^*(\Sigma^+)(C)$. But by definition, $[[\phi]]^{C^*} \subseteq W^*$ which proves that all neihbourhoods are covered by the global modality relation.

QED

Now, let us take a generated submodel, as described in [1] for basic modal logic, using the maximally consistent set Σ^+ looking only at the global modality.

Proposition 3.6 *The generated canonical submodel under Σ^+ preserves both global modality and monotonic Coalition Logic formulas satisfaction.*

Proof.

It is just a matter of verifying that the generated submodel for the global relation is also a generated submodel for the coalitional relation.

QED

It follows that we have an axiomatization for the Coherent Coalition Logic.

3.3 A sound and complete axiomatization

Take now the maximally consistent sets $w \in W^*$, closed under the proof system depicted in the table.

We take the following conditions to describe coherence of the Effectivity Function on the canonical relation.

- $wE^*_C X$ iff $\exists\overline{\phi} \subseteq X : [C]\phi \in w$ and $\forall\overline{\psi} \subseteq (W^* \backslash X) : [\overline{C}]\psi \notin w$ (for $C \neq \emptyset$)
- $E^*_C \subseteq E^*_D$ (for $C \subseteq D$)
- $wE^*_C X$ iff $X = W^*$ (for $C = \emptyset$)
- $wR_\exists v$ iff $w, v \in W^*$

Proposition 3.7 *The canonical Coherent frame for Coalition Logic with $A\phi \leftrightarrow [\emptyset]\phi$ as axiom has the property that $E(w)(\emptyset) = \{W^*\}$ for any MCS w and $A\phi \leftrightarrow [\emptyset]\phi$ is valid in the class of frames with that property.*

It is a consequence of the previous propositions and the canonical relation definition.

Proposition 3.8 *The set of axioms and rules in the table are sound and complete with respect to Coherent Coalition Frames*

Proof.

We need just to check the statement with respect to $M^{*'}$. We omit the detailed proof.

QED

Proof System	
A1	$[C]\phi \to [D]\phi$ (for $C \subseteq D$)
A2	$[C]\phi \to \neg[\overline{C}]\neg\phi$
A3	$A\phi \leftrightarrow [\emptyset]\phi$
A4	$\phi \to E\phi$
A5	$EE\phi \to E\phi$
A6	$\phi \to AE\phi$
A7	$A(\phi \to \psi) \to (A\phi \to A\psi)$
R1	$\phi \land (\phi \to \psi) \Rightarrow \psi$
R2	$\phi \to \psi \Rightarrow [C]\phi \to [C]\psi$
R3	$\phi \Rightarrow A\phi$

3.4 On Agt-Maximal Coherent Games

Notice that if we add Agt-maximality to Coherent Games, the following holds:

$$M, w \models [Agt]\phi \leftrightarrow E\phi$$

This suggest, at the expressivity level, that Coherent Coalition Logic is powerful enough to reason on global properties of the models. These results are useful to apply the language to the study of multiagent interactions.

4 A Deontic Logic for Efficient Interactions

Any deontic language comes along with an idea of how a certain world state should be.

Once we view a deontic language as regulating a Multi Agent System, we can say that a set of commands promote a certain interaction (or social state), prohibiting certain others. Following this line of reasoning it is possible, given a notion of optimality or efficiency, to construct a deontic language that requires this notion to hold.

If we want to consider what it is socially optimal, as we do here, we can see obligations and prohibitions as resulting from one general norm saying that all actions of coalitions that do not take into account the interests of the society as a whole, are forbidden.

From the practical point of view, one way to view our logic is to say that it can be used to derive obligations, permission and prohibitions from conflicting group preferences, and use these as *suggestions* for norm introduction in the society.

This last part of the paper is devoted to formalize this derivation. Here we will introduce a notion of preference in the strategic interaction scenario, to be lifted to coalitional choice, in order to define what it is best for a society to choose. We will then move to study the property of the enriched language focusing on the regulation of coherent interactions. We will show that Nature can be obliged to do something when and only when it is not avoidable, that is it will be assigned only trivial obligations.

4.1 Preference

As already noticed by von Wright, the notion of preference can be understood and modeled in many ways [9]. This is especially true in strategic interaction, in which players, in order to choose what is best to do, need to have preferences over the possible outcomes of the game. Thus those are the preferences that constitute our main concern.

The claim is thus that players do have a fixed ordering over the domain of discourse (what we call *preferences*), and that generate their strategic preference considering where the game may end (called *choices domination*, or simply *domination*).

We start from a preference relation for individuals over states working our way up to preferences for coalitions over sets. A similar view is taken in [3].

Definition 4.1 [Individual preferences for states] A preference ordering $(\geq_i)_{i \in Agt}$ consists of a partial order (reflexive, transitive, antisymmentric) $\geq_i \subseteq W \times W$ for all agents $i \in Agt$, where $v \geq_i w$ means that v is 'at least as nice' as w for agent i. The corresponding strict order is defined as usual: $v >_i w$ if, and only if, $v \geq_i w$ and not $w \geq_i v$. ◁

Definition 4.2 [Individual preferences for sets of states] Given a preference ordering $(\geq_i)_{i \in Agt}$, we lift it to an ordering on nonempty sets of states by means of the following principles.

1. $\{v\} \geq_i \{w\}$ iff $v \geq_i w$; (Singletons)
2. $(X \cup Y) \geq_i Z$ iff $X \geq_i Z$ and $Y \geq_i Z$; (Left weakening)

3. $X \geq_i (Y \cup Z)$ iff $X \geq_i Y$ and $X \geq_i Z$. (Right weakening)

$\lhd$

These are some properties that seem minimally required for calling some relation a preference relation. The first ensures that preferences are copied to possible choices. The properties of left and right weakening ensure a lifting from singletons to sets.

The lifting enables us to deal with preference under uncertainty or indeterminacy. The idea is that if an agent were ever confronted with two choices X, Y he would choose X over Y provided $X >_i Y$. preferences do not consider any realizability condition, they are simply basic aspirations of individual players, on which to construct a more realistic order on the possible outcomes of the game, which are by definition dependent on what all the agents can do together.

Out of agents' preferences, we can redefine on choices the classical notion of Pareto Efficiency.

Definition 4.3 [Strong Pareto efficiency] Given a choice set $\mathcal{X}$, a choice $X \in \mathcal{X}$ is *Strongly Pareto efficient* for coalition C if, and only if, for no $Y \in \mathcal{X}$, $Y \geq_i X$ for all $i \in C$, and $Y >_i X$ for some. When $C = Agt$ we speak of *Strong Pareto Optimality.* $\lhd$

We will use the characterization of Pareto Efficiency and Optimality to refer to the notions we have just defined, even though the classical definitions (compare [5]) are weaker [3].

We now construct a preference relation on choices. To do so we first need to look at the interaction that agents' choices have with one another.

Definition 4.4 [Subchoice] If E is an Effectivity Function, and $X \in E(w)(\overline{C})$, then the X-*subchoice set* for C in w is given by $E^X(w)(C) = \{X \cap Y \mid Y \in E(w)(C)\}$. $\lhd$

Considering subchoices allows to reason on a restriction of the game and to consider possible moves looking from a coalitional point of view, i.e. what is best for a coalition to do provided the others have already moved.

When agents interact therefore they make choices on the grounds of their own preferences. Nevertheless the moves at their disposal need not be all those that the grand coalition has. We can reasonably assume that preferences are filtered through a given coalitional Effectivity Function. That is we are going to consider what agents prefer among the things they can do.

Definition 4.5 [Domination] Given an Effectivity Function E, X is *undominated* for C in w (abbr. $X \rhd_{C,w}$) if, and only if, (i) $X \in E(w)(C)$ and (ii) for all $Y \in E(w)(\overline{C})$, $(X \cap Y)$ is Pareto efficient in $E^Y(w)(C)$ for C. $\lhd$

The idea behind the notion of domination is that if X' and X'' are both members of $E(w)(C)$ then, in principle, C will not choose X'', if X' dominates X''. This property ensures that a preference takes into account the possible moves of the other players. This resembles the notion of Individual Rationality in Nash solutions [5], according to which an action is chosen reasoning on the possible moves of the others.

If we take the Coordination Game previously discussed, we have the following cases:

- $(White_R, White_C) \rhd_{Agt,w}$ for any w.
- $(Black_R, Black_C) \rhd_{Agt,w}$ for any w.
- not $(Black_C) \rhd_{C,w}$

The preceding three definitions capture the idea that 'inwardly' coalitions reason Pareto-like, and 'outwardly' coalitions reason strategically, in terms of strict domination. A coalition will choose its best option given all possible moves of the opponents. Looking at the definition of Optimality we gave, we can see that undomination collapses to individual rationality when we only consider individual agents, and to Pareto efficiency when we consider the grand coalition of agents.

Proposition 4.6

$X \rhd_{Agt,w}$ iff X is a standard Pareto Optimal Choice in w.

$X \rhd_{i,w}$ iff X is a standard Dominating Choice in w for i.

Proof For the first, notice that since $E(w)(\emptyset) = \{W\}$, then X is undominated for Agt in w iff it is Pareto efficient in $E(w)(Agt)$ for Agt (i.e., it is Pareto optimal in w). The second is due to the restriction of undomination to singleton agents.Q.E.D.

4.1.1 Violation

A way to impose normative constraints in a Multi Agent System is to look at the optimality of the strategic interaction of such system. In particular the presence of possible outcomes in which agents could not unanimously improve (Pareto Efficient) can be a useful guide line for designing a new set of norms to be imposed.

Following this line we define a a set of violation sets as the set of those choices that are not a Pareto Efficient interaction.

Definition 4.7 [Violation] If E is an Effectivity Function and $C \subseteq C'$, then the choice $X \in E(w)(C)$ is a C'-violation in w ($X \in VIOL_{C,C',w}$) iff there is a $Y \in$

[3] The last definition is clearer when we consider the case $\mathcal{X} = E(w)(C)$. But it is formulated in a more abstract way in order to smoothen the next two definitions.

$E(w)(C' \setminus C)$, $(X \cap Y)$ that is not undominated for C' in w. $\lhd$

In words, X is a violation if it is not safe for the other agents, in the sense that not all the moves at their disposal yield an efficient outcome.

We indicate with $VIOL_{C,w}$ the set violations by C at w towards Agt.

5 Logic

We now introduce the syntax of our logic, an extension of the language of Coalition Logic [7] with modalities to talk about ideal states in a closed-world interaction.

5.1 Language

Let Agt be a finite set of agents and $Prop$ a countable set of atomic formulas. The syntax of our Logic is defined as follows:

$$\phi ::= p|\neg\phi|\phi \vee \phi|[C]\phi|E\phi|P(C,\phi)|F(C,\phi)|O(C,\phi)|[rational_C]\phi$$

where p ranges over $Prop$ and C ranges over the subsets of Agt. The other boolean connectives are defined as usual. The informal reading of the modalities is: "Coalition C can choose ϕ", "There is a state that satisfies ϕ","It is permitted (/forbidden/obligated) for coalition C to choose ϕ", "It is rational for coalition C to choose ϕ".

5.2 Structures

Definition 5.1 [Models] A *model* for our logic is a tuple

$$(W, E, R_{\exists}, \{\geq_i\}_{i \in Agt}, V)$$

where:

- W is a nonempty set of states;
- $E : W \longrightarrow (2^{Agt} \longrightarrow 2^{2^W})$ is a Coherent Effectivity Function.
- $R_{\exists} = W \times W$ is a global relation.
- $\geq_i \subseteq W \times W$ for each $i \in Agt$, is the preference relation. Out of this relation we define the undomination relation $\rhd \subseteq 2^{Agt} \times W \times 2^W \times 2^{2^W}$ as previously specified.
- $V : W \longrightarrow 2^{Prop}$ is a valuation function.

$\lhd$

5.3 Semantics

The satisfaction relation of modal formulas (the rest is standard) with respect to a pointed model M, w is defined as follows:

$$\begin{array}{rll}
M,w \models [C]\phi & \text{iff} & [[\phi]]^M \in E(w)(C) \\
M,w \models E\phi & \text{iff} & \exists v \text{ s.t. } M,v \models \phi \\
M,w \models [rational_C]\phi & \text{iff} & \forall X(X \rhd_{C,w} \;\Rightarrow\; X \subseteq [[\phi]]^M) \\
M,w \models P(C,\phi) & \text{iff} & \exists X \in E(w)(C) \text{ s.t.} \\
 & & X \in \overline{VIOL}_{C,w} \text{ and } X \subseteq [[\phi]]^M \\
M,w \models F(C,\phi) & \text{iff} & \forall X \in E(w)(C)(X \subseteq [[\phi]]^M \Rightarrow \\
 & & X \in VIOL_{C,w}) \\
M,w \models O(C,\phi) & \text{iff} & \forall X \in E(w)(C)(X \in \overline{VIOL}_{C,w} \Rightarrow \\
 & & X \subseteq [[\phi]]^M)
\end{array}$$

In this definition, $[[\phi]]^M =_{def} \{w \in W \mid M,w \models \phi\}$.

The modality for coalitional ability is standard from Coalition Logic [7]. The modality for rational action requires for a proposition ϕ to be rational (wrt a coalition C in a given state w) that all undominated choices (for C in w) be in the extension of ϕ. This means that there is no safe choice for a coalition that does not make sure that ϕ will hold. It is still possible for a coalition to pursue a rational choice that may be socially not rational. The deontic modalities are defined in terms of the coalitional abilities and preferences. A choice is permitted whenever is safe, forbidden when it may be unsafe (i.e. when it contains an inefficient choice), and obligated when it is the only safe.

6 Properties

It is now interesting to look at what we can say within our system.

Some Validities	
1	$P(C,\phi) \rightarrow \neg O(C,\neg\phi)$
2	$F(C,\phi) \leftrightarrow \neg P(C,\phi)$
3	$P(C,\phi) \vee P(C,\psi) \rightarrow P(C,\phi \vee \psi)$
4	$O(C,\phi) \rightarrow [C]\phi \rightarrow P(C,\phi)$
5	$[rational_C]\phi \wedge [rational_{Agt}]\neg\phi \rightarrow F(C,\phi)$
6	$O(C,\top)$
7	$O(\emptyset,\phi) \leftrightarrow [\emptyset]\phi$

The first validity says that permissions are consistent with obligations (the converse does not hold in general). The second that prohibition and permission are interdefinable. The third says that permission is monotonic. The fourth that the obligation to choose ϕ for an agent plus the ability to do something entails the permission to carry out ϕ. The validity number 5 says that the presence of a safe state that is rational for the grand coalition of agents is a norm for every coalition, even in case of conflicting preferences, i.e. in case of conflict the interest of the grand coalition prevails. The sixt one that there are no empty normative systems. The last validity says that obligations for Nature coincides with its ability. Notice that in Coherent Coalition Logic this means that obligation for Nature can only be a trivial choice.

6.1 Back to the Game

Norms of Conformity Consider the model M_c of the Game of Conformity described in Table 1.

Nature is obligated only a trivial choice:

$$M_c \models O(\emptyset, \phi) \leftrightarrow A\phi$$

What is interesting also is that also players are individually permitted only nontrivial choices:

$M_c \models \neg P(R, white_R) \wedge \neg P(R, black_R)$
$\wedge \neg P(C, black_C) \wedge \neg P(C, black_C)$.

But as coalition they are:

$$M_c \models P(\{R, C\}, white_{R,C}) \wedge M_c \models P(\{R, C\}, black_{R,C})$$

No precise indication of the choices is given by the resulting obligation:

$$M_c \models O(\{R, C\}, (white_R, C) \vee (black_{R,C}))$$

This is revealing of the form of the game: no equilibrium can be achieved by the agents acting independently, but only as a coalition [4]. As a matter of fact, looking at the obligations for this game tells us more than just a static fact about coalitional choice. In Coordination games only the grand coalition can make an optimal choice, which suggest that the grand coalition is in fact obligated to form.

7 Conclusion and Future Work

In this paper we studied those interactions in which Nature does not play an active role, and we proposed a deontic logic to indicate their optimal solutions. We provided an axiomatization of the resulting logic, switching from gameform interactions to interactions with preferences in order to analyze gametheoretical examples like Coordination Game. The work here described allows for several developments. Among the most interesting ones is the study of the relation between imposed outcomes and steady states that describe where the game will actually end up (i.e. Nash Solution, the Core etc...). As suggested by the last example, some obligations say something about the convenient dynamics to achieve a socially optimal outcome. One idea is to talk explicitly about such dynamics. Conversely another feature that is worth studying is those structures in which Pareto Efficiency is not always present. Agents will reckon some actions as optimal even though there is no social equilibrium that can be ever reached. This can be achieved by talking explicitly about preferences in the language as done for instance in [8]. The study of the interaction between choices and preferences has shown to have an interesting connection with deontic logic that, viewed in a multiagent perspective, allows to talk about those desirable properties that an interaction should have. As system designers, our aim is at last to construct efficient social procedures that can guarantee a socially desirable property to be reached. We think that normative system design is at last a proper part of the Social Software enterprise [6].

References

[1] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Cambridge Tracts in Theoretical Computer Science, 2001.

[2] S. Borgo. Coalitions in action logic. In *Proc. of IJCAI*, pages 1822–1827, 2007.

[3] P. Gardenfors. Rights, games and social choice. *Nous*, 15:341–56, 1981.

[4] H.H. Hansen. *Monotonic Modal Logics*. Master Thesis, ILLC, 2001.

[5] M. Osborne and A. Rubinstein. *A course in Game Theory*. The MIT Press, 1994.

[6] R. Parikh. Social software. *Synthese*, 132(3):187–211, 2002.

[7] M. Pauly. *Logic for Social Software*. ILLC Dissertation Series, 2001.

[8] J. van Benthem and F.Liu. Dynamic logic of preference upgrade. *Journal of Applied Non-Classical Logics*, 14, 2004.

[9] G.H. von Wright. *The logic of preference*. Edimburgh University Press, 1963.

[4] Notice that we have no way of detaching from this choice a more precise command: $O(C, \phi \vee \psi) \rightarrow ((O(C, \phi) \vee O(C, \psi)))$ is not a validity.

Verifying time, memory and communication bounds in systems of reasoning agents[*]

Natasha Alechina, Brian Logan, Nguyen Hoang Nga and Abdur Rakib
School of Computer science, University of Nottingham
Nottingham NG8 1BB, UK
{nza,bsl,hnn,rza}@Cs.Nott.AC.UK

Abstract

We present a framework for verifying systems composed of heterogeneous reasoning agents, in which each agent may have differing knowledge and inferential capabilities, and where the resources each agent is prepared to commit to a goal (time, memory and communication bandwidth) are bounded. The framework allows us to investigate, for example, whether a goal can be achieved if a particular agent, perhaps possessing key information or inferential capabilities, is unable (or unwilling) to contribute more than a given portion of its available computational resources or bandwidth to the problem.

1. Introduction

A distributed approach to problem solving involves the collective effort of multiple agents combine their knowledge and information to solve problems which no single agent could solve alone or to solve problems more effectively. For a given problem, for different multi agent systems, different solution strategies will be preferred depending on the relative costs of computational and communication resources for each agent. These tradeoffs may be different for different agents (e.g., reflecting their computational capabilities or network connection) and may reflect the agent's commitment to a particular problem. For a given set of agents with specified inferential abilities and resource bounds it may not be clear whether a particular problem can be solved at all, or, if it can, what computational and communication resources must be devoted to its solution by each agent.

There has been considerable work in the agent literature on distributed problem solving in general e.g., [15, 20, 22] and on distributed reasoning in particular [1, 8]. Much of this work analyses the time and communication complexity of distributed reasoning algorithms. In this paper we present a framework for reasoning about tradeoffs between time, memory and communication in systems of distributed reasoning agents. In contrast to previous work, e.g., [3] which focused primarily on memory limitations of single reasoners, our approach allows us to specify bounds on the number of messages the agents can exchange, allowing the investigation of tradeoffs between different resources. We introduce a novel epistemic logic, $BMCL$, for specifying resource-bounded reasoners. Critically, the logic allows upper bounds on the resource commitments (time, memory and communication) of each agent in the system to be specified. The logic is sound and complete and admits efficient model-checking. Using simple resolution examples, we show how to encode systems of distributed reasoning agents specified in the logic in a model checker, and verify some example properties.

2 Distributed Reasoners

We define the 'shape' of a proof in terms of the maximum space requirement at any step in the proof and the number of inference steps it contains. The lower bound on space for a given problem is then the least maximum space requirement of any proof, and the lower bound on time is the least number of inference steps of any proof. In general, a minimum space proof and a minimum time proof will be different (have different shapes). Bounding the space available for a proof will typically increase the number of inference steps required and bounding the number of steps will increase the space required.

We define the bounds on a reasoning agent in terms of its available resources expressed in terms of memory, time and communication. We assume that the memory required for a particular proof can be taken to be its space requirement (e.g., the number of formulas that must be simultaneously held in memory) times some constant. For a single threaded

[*] An extended version of this paper will be presented at AAMAS 2008, Estoril, Portugal.

agent, the number of inference steps executed times some constant can be taken as a measure of the time necessary to solve the problem. The communication requirement of a proof is taken to be the number of messages exchanged with other agents. In what follows, we ignore the constants and assume that the units of problem size and resources are the same.

In the distributed setting we distinguish between *symmetric problem distributions*, where all agents have the same premises and the same rules of inference, and *asymmetric problem distributions* where different premises may be assigned to different agents and/or the agents use different rules of inference. Similarly, we can distinguish between *symmetric resource distributions* (when all agents have the same resource bounds) and *asymmetric resource distributions*, (when different agents have different resource bounds).

Distribution does not necessarily change the shape (maximum space requirement and number of inference steps) of a proof. However, in a distributed setting the tradeoffs between memory and time bounds is complicated by communication. Unlike memory and time, communication has no direct counterpart in the proof. However like memory, communication can be substituted for time (e.g., if part of the proof is carried out by another agent), and, like time, it can be substituted for memory (e.g., if a lemma is communicated by another agent rather than having to be remembered). In the distributed setting, each agent has a minimum memory bound which is determined by its inference rules and which may be smaller than the minimum space requirement for the problem. If the memory bound for all agents taken individually is less than the minimum space requirement for the problem, then the communication bound must be greater than zero.

In the next section, we present measures of space, time and communication for distributed reasoning agents which allow us to make these tradeoffs precise.

3 Measuring Resources

We assume a set of n agents where each agent i has a set of propositional inference rules R_i (for example, R_i could contain conjunction introduction and modus ponens, or it could contain just a single rule of resolution) and a set of premises or a knowledge base K_i. The notion of a derivation, or a proof of a formula G from K_i is standard. We view the process of producing a proof of G as a sequence of configurations or states of a reasoner, starting from an empty configuration, and producing the next configuration by one of four operations: **Read** copies a formula from K_i into the current configuration; **Infer** applies a rule from R_i to formulas in the current configuration; **Skip** leaves the configuration unchanged; and **Copy** copies a formula α into the next configuration of agent j if α is in the current configuration of agent $i, j \neq i$. Note that **Read**, **Infer** and **Copy** may overwrite a formula from the previous configuration. The goal formula is derived if it occurs in the configuration of one of the agents.

We take the time complexity of a derivation to be the length of the sequence of configurations. Space complexity is taken to be the size of configurations as in [6].[1] The size of a configuration can be measured either in terms of the maximal number of formulas appearing in any configuration or in terms of the number of symbols required to represent a configuration. Clearly, for some inference systems, for example, where the set of inference rules contains both conjunction introduction and conjunction elimination, the first size measure results in constant space usage. However, for other systems, such as resolution, counting formulas results in non-trivial space complexity [13]. In this paper, we take the size of a configuration to be the maximal number of formulas, since all the reasoning systems we consider have a non-trivial space complexity for this measure.

#	Configuration	Operation
1	$\{\}$	
2	$\{A_1\}$	**Read**
3	$\{A_1, A_2\}$	**Read**
4	$\{A_1, A_1 \wedge A_2\}$	**Infer**
5	$\{A_1 \wedge A_2, A_1 \wedge A_2 \rightarrow B_1\}$	**Read**
6	$\{A_1 \wedge A_2, B_1\}$	**Infer**

Figure 1. Example derivation using $\bigwedge_I$ and MP

As an illustration, Figure 1 shows the space and time complexity of the derivation of the formula B_1 from $A_1,\ A_2,\ A_1 \wedge A_2 \rightarrow B_1$ in an inference system which contains only conjunction introduction and modus ponens. The length of the proof is 6 and the space usage is 2 (at most 2 formulas need to be present in the configuration at any given time). It is worth observing that the inference system consisting of just conjunction introduction and modus ponens does not have constant space complexity when space is measured as the number of formulas; a sequence of derivation examples requiring (logarithmically) growing space can easily be constructed starting from the example above, and continuing with a derivation of C_1 from $A_1, A_2, A_3, A_4, A_1 \wedge A_2 \rightarrow B_1, A_3 \wedge A_4 \rightarrow B_2, B_1 \wedge B_2 \rightarrow C_1$, etc.

Most research in time and space complexity of proofs

[1] We deviate from [6] in that we do not have an explicit *erase* operation, preferring to incorporate erasing (overwriting) in the *read* and *infer* operations. This obviously results in shorter proofs; however including an explicit erase operation gives proofs which are no more than twice as long as our proofs if we don't require the last configuration to contain *only* the goal formula.

#	Configuration	Operation
1	$\{\,\}$	
2	$\{A_1 \vee A_2\}$	**Read**
3	$\{A_1 \vee A_2, \neg A_1 \vee A_2\}$	**Read**
4	$\{A_1 \vee A_2, A_2\}$	**Infer**
5	$\{A_2, A_1 \vee \neg A_2\}$	**Read**
6	$\{A_2, A_1 \vee \neg A_2, \neg A_1 \vee \neg A_2\}$	**Read**
7	$\{A_2, \neg A_2, \neg A_1 \vee \neg A_2\}$	**Infer**
8	$\{\emptyset, \neg A_2, \neg A_1 \vee \neg A_2\}$	**Infer**

Figure 2. Example derivation using resolution

	Agent 1		Agent 2	
#	Configuration	Op.	Configuration	Op.
1	$\{\}$		$\{\}$	
2	$\{A_1 \vee A_2\}$	**Read**	$\{A_1 \vee \neg A_2\}$	**Read**
3	$\{A_1 \vee A_2, \neg A_1 \vee A_2\}$	**Read**	$\{\neg A_1 \vee \neg A_2, A_1 \vee \neg A_2\}$	**Read**
4	$\{A_1 \vee A_2, A_2\}$	**Infer**	$\{\neg A_2, A_1 \vee \neg A_2\}$	**Infer**
5	$\{A_1 \vee \neg A_2, A_2\}$	**Read**	$\{\neg A_2, A_2\}$	**Copy**
6	$\{A_1, A_2\}$	**Infer**	$\{\{\}, A_2\}$	**Infer**

Figure 3. Example derivation using resolution with two agents

has focused on the lower bounds for the inference system as a whole. While we are interested in the lower bounds, we are also interested in the trade-offs between time and space usage for particular derivations. For example, consider a set of premises $A_1, A_2, A_3, A_4, A_1 \wedge A_2 \rightarrow B_1, A_3 \wedge A_4 \rightarrow B_2$, $B_1 \wedge B_2 \rightarrow C_1$ and a goal formula $A_1 \wedge A_2 \wedge C$. It is possible to derive the goal from the premises using conjunction introduction and modus ponens and configurations of size 3 in 17 steps (deriving $A_1 \wedge A_2$ twice). On the other hand, with configurations of size 4 the proof is 3 steps shorter.

Different inference systems have different complexity and different tradeoffs. Figure 2 illustrates the (non-trivial) space complexity of resolution proofs in terms of the number of formulas in a configuration. The example, which is due to [13], shows the derivation of an empty clause by resolution from the set of all possible clauses of the form

$$\sim A_1 \vee \sim A_2 \vee \ldots \vee \sim A_n$$

(where $\sim A_i$ is either A_i or $\neg A_i$), for $n = 2$. Its space usage is 3 and the length of the proof is 8.

In the multiagent case, when several reasoners can communicate to derive a common goal, an additional resource of interest is how many messages the reasoners must exchange in order to derive the goal. In the distributed setting, we assume that each agent has its own set of premises and inference rules and its own configuration, and that the reasoning of the agents proceeds in lock step.

The goal formula is derived if it occurs in the configuration of one of the agents. Our model of communication complexity is based on [25], except that we count the number of formulas exchanged by the agents rather than the number of bits exchanged. The communication complexity of a joint derivation is then the (total) number of **Copy** operations in the derivation.

In general, in a distributed setting, trade-offs are possible between the number of messages exchanged and the space (size of a single agent's configuration) and time required for a derivation. The total space use (the total number of formulas in all agent's configurations) clearly cannot be less than the minimal configuration size required by a single reasoner to derive the goal formula from the union of all knowledge bases using all of the available inference rules, however this can be distributed between the agents in different ways, resulting in different numbers of exchanged messages. Similarly, if parts of a derivation can be performed in parallel, the total derivation will be shorter, though communication of the partial results will increase the communication complexity. As an illustration, figure 3 shows one possible distribution of the resolution example in figure 2. As can be seen, two communicating agents can solve the problem faster than a single agent.

4 A Bounded Memory and Communication Logic BMCL

In this section we present a temporal epistemic logic $BMCL$ which allows us to describe a set of reasoning agents with bounds on memory and on the number of messages they can exchange. In this logic, we can express statements like 'the agents will be able to derive the goal formula in n inference steps'. The bounds on memory and communication ability are expressed as axioms in the logic. In this paper, as an example, we have chosen to axiomatise a set of agents reasoning using resolution. Other reasoning systems can be axiomatised in a similar way, and we briefly sketch how to add model conditions and axioms for reasoners which reason using conjunction introduction and modus ponens to our logic at the end of this section.

Let the set of agents be $AG = \{1, 2, .., n_{AG}\}$. For simplicity, we assume that they agree on a finite set $PROP$ of propositional variables (this assumption can easily be relaxed, so that only some propositional variables are shared). Since each agent uses resolution for reasoning, we assume that all formulas of the internal language of the agents are in the form of *clauses*. For convenience, we define a clause as a set of *literals* in which a literal is a propositional variable or its negation. Then the set of literals is defined as $LPROP = \{p, \neg p | p \in PROP\}$. If L is a literal, then $\neg L$ is $\neg p$ if L is a propositional variable p, and p if L is of the form $\neg p$. Let Ω be the set of all possible clauses over

$PROP$, i.e., $\Omega = \wp(LPROP)$. Note that Ω is finite.

The only rule of inference that each agent has is the *resolution rule* which is defined as follows:

$$\frac{\alpha \ni L \qquad \beta \ni \neg L}{(\alpha \setminus \{L\}) \cup (\beta \setminus \{\neg L\})} \text{ Res}$$

which states that if there are two clauses α and β such that one contains a literal L and the other contains $\neg L$, then we can derive a new clause $(\alpha \setminus \{L\}) \cup (\beta \setminus \{\neg L\})$.

Each agent i has a memory of size $n_M(i)$ where one unit of memory corresponds to the ability to store an arbitrary clause. Agent i can read clauses from its set of premises K_i. We assume that each K_i is finite. The communication ability of the agents is expressed by the *copy* action which copies a clause from another agent's memory. The limit on each agent's communication ability is $n_C(i)$: in any valid run of the system, agent i can perform at most $n_C(i)$ copy actions.

4.1 Syntax of $BMCL$

The syntax of $BMCL$ is defined inductively as follows.

- $\top$ is a well-formed formula (wff) of $BMCL$.
- If α is a clause, then $B^r_i\alpha$ is a wff of $BMCL$, for all $i \in AG$.
- If α is a clause, then $B^c_i\alpha$ is a wff of $BMCL$, for all $i \in AG$.
- If ϕ and ψ are wff, then so are $\neg\phi$, $\phi \wedge \psi$.
- If ϕ and ψ are wff, then so are $X\phi$, $\phi U\psi$, $Y\phi$, $\phi S\psi$ and $A\phi$.

Classical abbreviations for $\vee$, $\rightarrow$, $\leftrightarrow$ and $\bot$ are defined as usual.

The language contains both temporal and epistemic modalities. For the temporal part of $BMCL$, we have $PCTL^*$, a branching time temporal logic with the past operator.[2] Intuitively, $PCTL^*$ describes infinite trees, or all possible runs of the system, over discrete time. In the temporal logic part of the language, X stands for next step, U for until, Y for previous step, S for since and A for 'on all paths'. We will also use abbreviations $F\phi \equiv \top U\phi$ for some time in the future, $P\phi \equiv \top S\phi$ for some time in the past, $E\phi \equiv \neg A\neg\phi$ for on some path and $start \equiv \neg Y\top$ for the starting state of the system. The epistemic part of the language consists of belief modalities $B^r_i\alpha$, which means that agent i has read α from its knowledge base or derived it, and $B^c_i\alpha$, which means that i has copied α from another agent. We define $B_i\alpha$ (agent i believes α) to be $B^r_i\alpha \vee B^c_i\alpha$.

4.2 Semantics of $BMCL$

The semantics of $BMCL$ is defined by $BMCL$ tree-like transition systems. A $BMCL$ transition system $M = (S, R, V^r, V^c)$ is defined as follows.

- S is a non-empty set of states.
- $R \subseteq S \times S$ is a total binary relation, i.e. for any $s \in S$, there exists $t \in S$ such that $(s, t) \in R$. Moreover, it is also required that (S, R) is a tree-frame. A branch σ is an infinite sequence $(s_0, s_1, ..)$ such that $(s_i, s_{i+1}) \in R$ for all $i \geq 0$, σ_i denotes the element s_i of σ and $\sigma_{\leq i}$ is the prefix $(s_0, s_1, .., s_i)$ of σ. The set of all branches is denoted as BR. Note that since (S, R) is a tree-frame every state s has a unique past $past(s) = \sigma_{\leq i}$ where $\sigma_i = s$.
- $V^r : S \times AG \rightarrow \wp(\Omega)$, is a mapping that defines for each state which formulas an agent believes due to reading or inference.
- $V^c : S \times AG \rightarrow \wp(\Omega)$, is a mapping that defines for each state which formulas an agent copied from the memories of other agents.

The truth of a $BMCL$ formula in a state at point n of a path σ of M is defined inductively as follows:

- $M, \sigma, n \models B^r_i\alpha$ iff $\alpha \in V^r(\sigma_n, i)$,
- $M, \sigma, n \models B^c_i\alpha$ iff $\alpha \in V^c(\sigma_n, i)$,
- $M, \sigma, n \models \neg\phi$ iff $M, \sigma, n \not\models \phi$,
- $M, \sigma, n \models \phi \wedge \psi$ iff $M, \sigma, n \models \phi$ and $M, \sigma \models \psi$,
- $M, \sigma, n \models X\phi$ iff $M, \sigma, n+1 \models \phi$,
- $M, \sigma, n \models \phi U\psi$ iff $\exists m \geq n$ such that $\forall k \in [n, m)$ $M, \sigma, k \models \phi$ and $M, \sigma, m \models \psi$,
- $M, \sigma, n \models Y\phi$ iff $n > 0$ and $M, \sigma, n-1 \models \phi$,
- $M, \sigma, n \models \phi S\psi$ iff $\exists m \leq n$ such that $\forall k \in (m, n]$ $M, \sigma, m \models \phi$ and $M, \sigma, k \models \psi$,
- $M, \sigma, n \models A\phi$ iff $\forall \sigma' \in BR$ such that $\sigma'_{\leq n} = \sigma_{\leq n}$, $M, \sigma', n \models \phi$.

Now we describe conditions on the models. The first set of conditions refers to the accessibility relation R. The intuition behind the conditions is that R corresponds to the agents executing actions $\langle a_1, \ldots, a_{n_{AG}} \rangle$ in parallel, where action a_i is a possible action (transition) for the agent i in a given state. Actions of each agent i are: $Read_{i,\alpha,\beta}$ (reading a clause α from the knowledge base and erasing β), $Res_{i,\alpha_1,\alpha_2,L,\beta}$ (resolving α_1 and α_2 and erasing β), $Copy_{i,\alpha,\beta}$ (copying α from another agent and erasing β),

[2] The reason we use $PCTL^*$ rather than CTL^* is that we need the past operator to express the bound on agent communication.

$Erase_{i,\alpha}$ (erasing α), and $Null_i$ (doing nothing), where $\alpha, \alpha_1, \alpha_2, \beta \in \Omega$ and $L \in LPROP$.[3] Intuitively, β is an arbitrary clause which may or may not be in the agent's memory, which gets overwritten in this transition. If the agent's memory is full ($|V^r(s,i)| + |V^c(s,i)| = n_M(i)$), then β has to be in $V^r(s,i) \cup V^c(s,i)$, otherwise we cannot add an extra formula to it (this would violate the condition on memory defined below). Not all actions are possible in any given state (for example, to perform a resolution step from state s, the agent has to have two resolvable clauses assigned in s). Let us denote the set of all possible actions by agent i in state s by $R_i(s)$.

Below is the definition of $R_i(s)$:

Definition 4.1 [Available actions] For every state s and agent i,

1. $Read_{i,\alpha,\beta} \in R_i(s)$ iff $\alpha \in K_i$ and $\beta \in \Omega$, or if $|V^r(s,i)| + |V^c(s,i)| = n_M(i)$ then $\beta \in V^r(s,i) \cup V^c(s,i)$.
2. $Res_{i,\alpha_1,\alpha_2,L,\beta} \in R_i(s)$ iff $\alpha_1, \alpha_2 \in \Omega$, $\alpha_1 \ni L$, $\alpha_2 \ni \neg L$, $\alpha_1, \alpha_2 \in V^r(s,i) \cup V^c(s,i)$, $\alpha = (\alpha_1 \setminus \{L\}) \cup (\alpha_2 \setminus \{\neg L\}) \notin K_i$ and β is as before.
3. $Copy_{i,\alpha,\beta} \in R_i(s)$ iff there exists $j \neq i$ such that $\alpha \in V^r(s,j) \cup V^c(s,j)$ and $past(s)$ does not contain more than $n_C(i) - 1$ transitions of the form $Copy_{i,\beta}$, and β is as before.[4]
4. $Null_i$ is always in $R_i(s)$.
5. There are no conditions on $Erase_{i,\alpha} \in R_i(s)$.

$\triangleleft$

Now we define effects of actions on the agent's state (assignments $V^r(s,i)$ and $V^c(s,i)$).

Definition 4.2 [Effects of actions] For each $i \in AG$, the result of performing an action a in state s is defined if $a \in R_i(s)$ and has the following effect on the assignment of clauses to i in the successor state t:

1. if a is $Read_{i,\alpha,\beta}$: $V^r(t,i) = (V^r(s,i) \setminus \{\beta\}) \cup \{\alpha\}$ and $V^c(t,i) = V^c(s,i) \setminus \{\beta\}$.
2. if a is $Res_{i,\alpha_1,\alpha_2,L,\beta}$: $V^r(t,i) = (V^r(s,i) \setminus \{\beta\}) \cup \{\alpha\}$ and $V^c(t,i) = V^c(s,i) \setminus \{\beta\}$, where $\alpha = (\alpha_1 \setminus \{L\}) \cup (\alpha_2 \setminus \{\neg L\})$.
3. if a is $Copy_{i,\alpha,\beta}$: $V^c(t,i) = (V^c(s,i) \setminus \{\beta\}) \cup \{\alpha\}$ and $V^r(t,i) = V^r(s,i) \setminus \{\beta\}$.
4. if a is $Null_i$: $V^r(t,i) = V^r(s,i)$ and $V^c(t,i) = V^c(s,i)$
5. if a is $Erase_{i,\alpha}$ then $V^r(t,i) = V^r(s,i) \setminus \{\alpha\}$ and $V^c(t,i) = V^c(s,i) \setminus \{\alpha\}$, where $\alpha \in \Omega$.

$\triangleleft$

Definition 4.3 $BMCM(K_1, .., K_{n_{AG}}, n_M, n_C)$ is the set of models $M = (S, R, V, C)$ such that:

1. For every s and t, $R(s,t)$ iff for some tuple of actions $\langle a_1, \ldots, a_{n_{AG}} \rangle$, $a_i \in R_i(s)$ and the assignment in t satisfies the effects of a_i for every i in $\{1, \ldots, n_{AG}\}$.
2. For every s and a tuple of actions $\langle a_1, \ldots, a_{n_{AG}} \rangle$, if $a_i \in R_i(s)$ for every i in $\{1, \ldots, n_{AG}\}$, then there exists $t \in S$ such that $R(s,t)$ and t satisfies the effects of a_i for every i in $\{1, \ldots, n_{AG}\}$.
3. The bound on each agent's memory is set by the following constraint on the mappings V^r and V^c:

$$|V^r(s,i)| + |V^c(s,i)| \leq n_M(i) \text{ for all } s \in S \text{ and } i \in AG$$

$\triangleleft$

Note that the bound $n_C(i)$ on each agent i's communication ability (no branch contains more than $n_C(i)$ $Copy$ actions by agent i) follows from the fact that $Copy_i$ is only enabled if i has performed fewer than $n_C(i)$ copy actions in the past.

4.3 Axiomatisation of $BMCL$

Before we give an axiomatisation for the set of models defined above, we need the following abbreviations for expressing that i has performed at least k copy actions in the past. A successful copying of a clause α by agent i from an agent j is defined by the following formula:

$$copied(i,j,\alpha) \equiv B_j\alpha \wedge \neg B_i\alpha \wedge XB_i^c\alpha$$

Copying of any clause from any agent by agent i is defined as follows:

$$copied_i \equiv \bigvee_{j \in AG,\ \alpha \in \Omega} copied(i,j,\alpha)$$

So to say that there are at least k copy actions in agent i's past, we can use

$$C_i^{\geq}(k) = \underbrace{(YP(copied_i \wedge YP(copied_i \wedge \ldots YP(copied_i) \ldots)}_{k \text{ times}}$$

[3] The $Erase_{i,\alpha}$ action is introduced for purely technical reasons, to obtain a simpler axiomatisation of the system. The optimal sequences of actions found by the system when verifying properties of agents will contain no $Erase$ actions so will not affect the verification process.

[4] Assume that the state contains a communication counter for each agent i, which is set to 0 in the start state and is incremented every time i performs a copy action. After the counter reaches $n_C(i)$, agent i cannot perform any more copy actions.

and to say that there are fewer than k copy actions in agent i's past, we can say

$$C_i^{\leq}(k) = \underbrace{\neg(YP(copied_i \wedge YP(copied_i \wedge \ldots YP(copied_i) \ldots)}_{k+1 \text{ times}}$$

To define exactly k copies, we can use $C_i(0) = C_i^{\leq}(0)$ and $C_i(k) = C_i^{\geq}(k) \wedge C_i^{\leq}(k)$ for $k > 0$.

Consider the following set of axiom schemata:

A1 Axioms and rules of $PCTL^*$ as given in [24].

A2 $\bigwedge_{\alpha_q \in Q} B_i\alpha_q \wedge C_i(n) \rightarrow EX(\bigwedge_{\alpha_q \in Q} B_i\alpha_q \wedge C_i(n) \wedge B_i^r\alpha)$ for all $\alpha \in K_i$, $i \in AG$, $Q \subseteq \Omega$ with $|Q| < n_M(i)$, and $n \geq 0$.

A3 $\bigwedge_{\alpha_q \in Q} B_i\alpha_q \wedge C_i(n) \wedge B_i\alpha_1 \wedge B_i\alpha_2 \rightarrow EX(\bigwedge_{\alpha_q \in Q} B_i\alpha_q \wedge C_i(n) \wedge B_i\alpha)$ for any α_1 and α_2 such that $\alpha_1 \ni L$ and $\alpha_2 \ni \neg L$ for some literal L, $\alpha = (\alpha_1 \setminus \{L\}) \cup (\alpha_2 \setminus \{\neg L\}) \notin K_i$, $Q \subseteq \Omega$ with $|Q| < n_M(i)$, and $n \geq 0$.

A4 $\bigwedge_{\alpha_q \in Q} B_i\alpha_q \wedge C_i(n) \wedge B_j\alpha \wedge C_i^{\leq}(n_C(i)) \rightarrow EX(\bigwedge_{\alpha_q \in Q} B_i\alpha_q \wedge C_i(n+1) \wedge B_i^c\alpha)$ for all $i \neq j \in AG$, $Q \subseteq \Omega$ with $|Q| < n_M(i)$, and $n \geq 0$.

A5 $EX(B_i\alpha_1 \wedge B_i\alpha_2) \rightarrow B_i\alpha_1 \vee B_i\alpha_2$

A6 $EX(\neg B_i\alpha_1 \wedge \neg B_i\alpha_2) \rightarrow (\neg B_i\alpha_1 \vee \neg B_i\alpha_2)$

A7 $EX(B_i^r\alpha \wedge C_i(n)) \rightarrow B_i^r\alpha \vee (\neg B_i^r\alpha \wedge C_i(n))$ for all $\alpha \in K_i$

A8 $EX(B_i^r\alpha \wedge C_i(n)) \rightarrow B_i^r\alpha \vee (\neg B_i^r\alpha \wedge \bigvee_{(\alpha_1,\alpha_2) \in Res(\alpha)}(B_i\alpha_1 \wedge B_i\alpha_2 \wedge C_i(n)))$ for all $\alpha \notin K_i$ and $Res(\alpha) = \{(\alpha_1, \alpha_2) \in \Omega \times \Omega | \alpha_1 \ni L, \alpha_2 \ni \neg L \text{ and } \alpha = (\alpha_1 \setminus \{L\}) \cup (\alpha_2 \setminus \{\neg L\})\}$ for some literal L and $n \geq 0$

A9 $EX(B_i^c\alpha \wedge C_i(n)) \rightarrow B_i^c\alpha \vee (\neg B_i^c\alpha \wedge C_i(n-1) \wedge (\bigvee_{j \in AG} B_j\alpha))$

A10 $B_i\alpha_1 \wedge .. \wedge B_i\alpha_{n_M} \rightarrow \neg B_i\alpha_{n_M+1}$ where $i = 1, .., n_{AG}$, and $\alpha_i \neq \alpha_j$ for all $i \neq j$

A11 $C_i^{\leq}(n_C(i))$

A12 $\bigwedge_{j \in J} EX(\bigwedge_{q \in Q} B_j\alpha_q \wedge C_j(k_j)) \rightarrow EX \bigwedge_{j \in J}(\bigwedge_{q \in Q} B_j\alpha_q \wedge C_j(k_j))$ where $J \subseteq AG$ and all indices j are distinct, and $Q \subseteq \Omega$.

A13 $\phi \rightarrow EX\phi$

Let $BMCL(K_1, .., K_{n_{AG}}, n_M, n_C)$ be the logic defined by the our axiomatization. Then we have the following result.

Theorem 4.4 $BMCL(K_1, .., K_{n_{AG}}, n_M, n_C)$ *is sound and weakly complete with respect to* $BMCM(K_1, .., K_{n_{AG}}, n_M, n_C)$.

Proof. The proof of soundness is standard. Due to lack of space, we only prove validity of the first $BMCL$ axiom.

Let us consider **A2** and a model $M = (S, R, V^r, V^c)$ of $BMCM$
$(K_1, \ldots, K_{n_{AG}}, n_M)$. Let $\sigma = (s_0, s_1, \ldots) \in BR$, it is required to prove for any m, that if $M, \sigma, m \models \bigwedge_{\alpha_q \in Q} B_i\alpha_q \wedge C_i(n)$, then $M, \sigma, m \models EX(\bigwedge_{\alpha_q \in Q} B_i\alpha_q \wedge C_i(n) \wedge B^r\alpha)$ where $\alpha \in K_i$, $i \in AG$, $Q \subseteq \Omega$ with $|Q| < n_M(i)$ and $n \geq 0$. Since $\alpha \in K_i$, $Read_{i,\alpha,\beta} \in R_i(\sigma_n)$ for some $\beta \in \Omega \setminus Q$. Therefore, there exists $t \in S$ such that $R(s, t)$ and t satisfies the effects of $Read_{i,\alpha,\beta}$. In other words, we obtain $V^r(t, i) = V^r(s, i) \cup \{\alpha\} \setminus \{\beta\}$ and $V^c(t, i) = V^c(s, i) \setminus \{\beta\}$, this shows $V^r(t, i) \ni \alpha$. Since $M, \sigma, m \models \bigwedge_{\alpha_q \in Q} B_i\alpha_q$, $V^r(s, i) \cup V^c(s, i) \ni \alpha_q$ for all $q \in Q$. Moreover, since $|Q| < n_M(i)$, we have $\beta \in \Omega \setminus Q$, therefore $V^r(t, i) \cup V^c(t, i) \ni \alpha_q$. Then, $M, \sigma', m + 1 \models \bigwedge_{\alpha_q \in Q} B_i\alpha_q \wedge B^r\alpha$ for some $\sigma' \in BR$ such that $\sigma'_{\leq m+1} = (\sigma_1, \ldots, \sigma_m, t)$.

Since $M, \sigma, m \models C_i(n)$, we have that agent i has performed exactly n copy actions on the prefix $(\sigma_1, \ldots, \sigma_m)$. Moreover, the action that agent i performs between σ_m and t is to read α from K_i, therefore it has still performed exactly n copy actions on the prefix $(\sigma_1, \ldots, \sigma_m, t)$. Then, it is straightforward that $M, \sigma', m + 1 \models C_i(n)$. That gives us $M, \sigma', m + 1 \models \bigwedge_{\alpha_q \in Q} B_i\alpha_q \wedge B^r\alpha \wedge C(i, n)$. Since $\sigma'_{\leq m} = \sigma_{\leq m}$ and $R(\sigma_m, t)$, we obtain $M, \sigma, m \models EX(\bigwedge_{\alpha_q \in Q} B_i\alpha_q \wedge C_i(n) \wedge B^r\alpha)$.

To prove completeness, a satisfying model for a consistent formula is constructed as in the completeness proof of $PCTL^*$ from [24]. Then we use the axioms to show that this model is in $BMCM(K_1, \ldots, K_{n_{AG}}, n_M, n_C)$. QED

4.4 Systems of Heterogeneous Reasoners

Changing the logic to accommodate reasoners which reason using a different set of inference rules rather than resolution is relatively straightforward. As an illustration, we show how to add model conditions and axioms for reasoners which use modus ponens and conjunction introduction. We assume that the knowledge base of these reasoners contains literals and implications of the form $L_1 \wedge \ldots \wedge L_n \rightarrow L$.

First of all, we need to change the conditions on models so that instead of using the Res action, a reasoner could change the state by performing MP and AND actions. Let i be an (MP, AND) reasoner. Define Ω_i as K_i closed under subformulas and the following conjunction introduction: if Q is a set of distinct literals from K_i, then $\wedge Q \in \Omega_i$. An agent i has actions $Read_{i,\phi,\beta}$ for any formula ϕ in K_i,

$Copy_{i,\phi,\beta}$ for any formula $\phi \in \Omega_i$, $Null_i$, $Erase_i$, and instead of Res it has $MP_{i,\phi_1,\phi_1 \rightarrow \phi_2,\beta}$ and $AND_{i,\phi_1,\phi_2,\beta}$.

Definition 4.5 [Availability of MP and AND] For any $s \in S$:

1. $MP_{i,\phi_1,\phi_1 \rightarrow \phi_2,\beta} \in R_i(s)$ iff $\phi_1, \phi_1 \rightarrow \phi_2 \in V^r(s,i) \cup V^c(s,i)$ and $\beta \in \Omega_i$.
2. $AND_{i,\phi_1,\phi_2,\beta} \in R_i(s)$ iff $\phi_1, \phi_2 \in V^r(s,i) \cup V^c(s,i)$ and $\beta \in \Omega_i$.

$\triangleleft$

Definition 4.6 [Effects of MP and AND] For every $s \in S$, the result of performing action a is defined if $a \in R_i(s)$ and has the following effect on the resulting state t:

1. if a is $MP_{i,\phi,\phi \rightarrow \phi_2,\beta}$ then $V^r(t,i) = V^r(s,i) \cup \{\phi_2\} \setminus \{\beta\}$ and $V^c(s,i) = V^c(t,i) \setminus \{\beta\}$.
2. if a is $AND_{i,\phi_1,\phi_2,\beta}$ iff $V^r(t,i) = V^c(s,i) \cup \{\phi_1 \wedge \phi_2\} \setminus \{\beta\}$ and $V^c(s,i) = V^c(t,i) \setminus \{\beta\}$.

$\triangleleft$

The corresponding axioms for the (MP, AND) reasoner are as follows:

A13 $\bigwedge_{q \in Q} B_i\phi_q \wedge C_i(n) \wedge B_i\phi_1 \wedge B_i(\phi_1 \rightarrow \phi_2) \rightarrow EX(\bigwedge_{q \in Q} B_i\phi_q \wedge C_i(n) \wedge B_i\phi_2)$ where $Q \subseteq \Omega_i$ with $|Q| < n_M(i)$

A14 $\bigwedge_{q \in Q} B_i\phi_q \wedge C_i(n) \wedge B_i\phi_1 \wedge B_i\phi_2 \rightarrow EX(\bigwedge_{q \in Q} B_i\phi_q \wedge C_i(n) \wedge B_i(\phi_1 \wedge \phi_2))$ where $Q \subseteq \Omega_i$ with $|Q| < n_M(i)$, and $\phi_1, \phi_2 \in \Omega_i$.

A15 $EX(B_i(\phi_1 \wedge \phi_2) \wedge C_i(n)) \rightarrow (B_i(\phi_i \wedge \phi_2) \vee (\neg B_i(\phi_i \wedge \phi_2) \wedge B_i\phi_1 \wedge B_i\phi_2 \wedge C_i(n)))$

A16 $EX(B_i\phi_2 \wedge C_i(n)) \rightarrow (B_i\phi_2 \vee (\neg B_i\phi_2 \wedge C_i(n) \wedge \bigvee_{\phi_1 \rightarrow \phi_2 \in K_i}(B_i\phi_1 \wedge B_i(\phi_1 \rightarrow \phi_2)))$ for all $\phi_2 \notin K_i$.

Now we can add the conditions and axioms for the (MP, AND) reasoner to the system for resolution reasoners and obtain an axiomatisation for the heterogeneous system of reasoners.

5 Verifying Resource Bounds

The logic $BMCL$ allows us to express precisely how beliefs of a set of resource-bounded agents change over time, and, given a memory and communication bound for each agent, to verify formulas which state that a certain belief will or will not be acquired within a certain number of steps. For example, given a system of two agents with premises $K_1 = \{\{p_1, p_2\}, \{\neg p_1, p_2\}\}$ and $K_2 = \{\{p_1, \neg p_2\}, \{p_1, \neg p_2\}\}$, with bounds $n_M(1) = 2$, $n_M(2) = 2$ (both agents have 2 memory cells) and $n_C(1) = 0$, $n_C(2) = 1$ (agent 1 cannot copy anything and agent 2 can copy one clause), we can prove that $start \rightarrow EX^5 B_2(\{\})$ (i.e., from the start state, the agents can derive the empty clause in 5 steps).

However, rather than deriving such properties by hand, it is more convenient to use an automatic method to verify them. In this section, we describe how the models in $BMCM(K_1, .., K_{n_{AG}}, n_M, n_C)$ can be encoded as an input to a model-checker to allow the automatic verification of the properties expressing resource bounds.

5.1 Model Checker Encoding

It is straightforward to encode a $BMCM$ model of such a system for a standard model checker, and to verify resource bounds using existing model checking techniques. For the examples reported here, we have used the Mocha model checker [7].

States of the $BMCM$ models correspond to an assignment of values to state variables in the model-checker. The state variables representing an agent's memory are organised as a collection of 'cells', each holding at most one clause. For an agent i with memory bound $n_M(i)$, there are $n_M(i)$ cells. Each cell is represented by a pair of bitvectors, each of length $k = |PROP|$, representing the positive and negative literals in the clause in some standard order (e.g., lexicographic order). For example, if $PROP$ contains the propositional variables A_1, A_2 and A_3 with index positions 0, 1 and 2 respectively, the clause $A_1 \vee \neg A_3$ would be represented by two bitvectors: "100" for the positive literals and "001" for the negative literals. This gives reasonably compact states.

Actions by each agent such as reading a premise, resolution and communication with other agents are represented by Mocha *atoms* which describe the initial condition and transition relation for a group of related state variables. Reading a premise ($Read_{i,\alpha,\beta}$) simply sets the bitvectors representing an arbitrary cell in agent i's memory to the appropriate values for the clause α. Resolution ($Res_{i,\alpha_1,\alpha_2,L,\beta}$) is implemented using simple bit operations on cells containing values representing α_1 and α_2, with the results being assigned to an arbitrary cell in agent i's memory. Communication ($Copy_{i,\alpha,\beta}$) is implemented by copying the values representing α from a cell of agent j to an arbitrary cell of agent i. To express the communication bound, we use a counter for each agent which is incremented each time a copy action is performed by the agent. After the counter for agent i reaches $n_C(i)$, the $Copy_{i,\alpha,\beta}$ action is disabled.

Mocha supports hierarchical modelling through composition of *modules*. A module is a collection of atoms and a specification of which of the state variables updated by

# agents	Distrib.	Memory	Comm.	Time
1	Symmetric	3	–	8
2	Symmetric	2, 2	1, 0	6
2	Symmetric	3, 3	1, 0	6
2	Symmetric	3, 3	0, 0	8
2	Symmetric	2, 1	1, 1	9
2	Asymmetric	2, 2	2, 1	7
2	Asymmetric	3, 3	2, 1	7
2	Asymmetric	3, 1	1, 0	8

Table 1. Tradeoffs between resource bounds

those atoms are visible from outside the module. In our encoding, each agent is represented by a module. A particular distributed reasoning system is then simply a parallel composition of the appropriate agent modules.

The specification language of Mocha is ATL, which includes CTL. We can express properties such as 'agent i may derive belief ϕ in n steps' as $EF\ tr(B_i\alpha)$ where $tr(B_i\alpha)$ is a suitable encoding of the fact that a clause α is present in the agent's memory (either as a disjunction of possible values of cell bitvectors, or as a special boolean variable which becomes true when one of the cells contains a particular value, for example all 0s for the empty clause). To obtain the actual derivation we can verify the negation of a formula, for example $AG\ \neg tr(B_i\alpha)$—the counterexample trace will show how the system reaches the state where α is proved.

5.2 Examples

Consider a single agent (agent 1) whose knowledge base contains all clauses of the form $\sim A_1 \vee \sim A_2$ where $\sim A_i$ is either A_i or $\neg A_i$, and which has the goal of deriving the empty clause. We can express the property that agent 1 will derive the empty clause at some point in the future as $EF\ B_1\{\}$.

Using the model checker, we can show that deriving the empty clause requires a memory bound of 3 and 8 time steps (see Figure 2).[5] We can also show that these space and time bounds are minimal for a single agent; i.e., increasing the space bound does not result in a shorter proof.

With two agents and a symmetric problem distribution (i.e., each agent has all the premises $\sim A_1 \vee \sim A_2$), we can show that a memory bound of 2 (i.e., the minimum required for resolution) and a communication bound of 1 gives a proof of 6 steps (see Figure 3). Reducing the communication bound to 0 results in no proof, as, with a memory bound of 2 for each agent, at least one clause must be communicated from one agent to the other. Increasing the space bound to 3 (for each agent) does not shorten the proof, though it does allow the communication bound to be reduced to 0 at the cost of increasing the proof length to 8 (i.e., the single agent case). Reducing the total space bound to 3 (i.e., 2 for one agent and 1 for the other, equivalent to the single agent case) increases the number of steps required to find a proof to 9 and the communication bound to 1 for each agent. In effect, one agent functions as a cache for a clause required later in the proof, and this clause must be copied in both directions.

If the problem distribution is asymmetric, e.g., if one agent has premises $A_1 \vee A_2$ and $\neg A_1 \vee \neg A_2$ and the other has premises $\neg A_1 \vee A_2$ and $A_1 \vee \neg A_2$, then with a memory bound of 2 for each agent, we can show that the time bound is 7, and the communication bound is 2 for the first agent and 1 for the second. Increasing the memory bound for each agent to 3 does not reduce the time bound. However the memory bound can be reduced to 1 and the communication bound reduced to 1 for one agent and 0 for the other, if the time bound is increased to 8 (again this is equivalent to the single agent case, except that one agent copies the clause it lacks from the other rather than reading it). These tradeoffs are summarised in Table 1.

Increasing the size of the problem increases the number of possible tradeoffs, but similar patterns can be seen to the 2-variable case. For example, if the agent's knowledge base contain all clauses of the form $\sim A_1 \vee \sim A_2 \vee \sim A_3$, then a single agent requires a memory bound of 4 and 16 steps to achieve the goal. In comparison, two agents, each with a memory bound of 2, require 13 steps and 4 messages to derive the goal.

While extremely simple, these examples serve to illustrate the interaction between memory, time and communication bounds, and between the resource distribution and the problem distribution.

6 Related Work

There exist several approaches to epistemic logic which model reasoners as resource-bounded (not logically omniscient), including deduction model of belief [21], step logic and active logic [12, 17], algorithmic knowledge [18, 14, 23], and other syntactic epistemic logics [11, 2, 5, 19] where each inference step takes the agent into the next (or some future) moment in time. A logic where the depth of belief reasoning is limited is studied in [16].

A considerable amount of work has also been done in the area of model-checking multi-agent systems (see, e.g., [10, 9]). However, this work lacks a clear connection between the way agent reasoning is modelled in agent theory (which typically assumes that the agents are logically omniscient) and the formalisations used for model checking, and emphasises correctness rather than the interplay between time, memory, bounds on communications and the ability

[5] The space required for problems of this form is known to be logarithmic in the number of premises [13].

of agents to derive a certain belief.

The current paper extends the work of [3] which proposed a method of verifying memory and time bounds in a single reasoner which reasons in classical logic using natural deduction rather than resolution. We also extend the work in [4] which analyses a system of communicating rule-based reasoners and verifies time bounds for those systems, but assumes unlimited memory. As far as we are aware, the logic we propose in this paper is the first attempt to analyse time, space and communication bounds of reasoners in one logical system, and verify properties relating to all three resources using a model-checker.

7 Conclusions and Future Works

In this paper, we analyse the time, space and communication resources required by a system of reasoning agents to achieve a goal. We give a rigorous definition of the measures for each of those resources, and introduce an epistemic logic $BMCL$ where we can express properties of a system of resource-bounded reasoning agents. In particular, we can express bounds on memory and communication resources as axioms in the logic. We axiomatise a system of agents which reason using resolution (other reasoning systems can be axiomatised in a similar way), prove that the resulting logic is sound and complete, and show how to express properties of the system of reasoning agents in $BMCL$. Finally, we show how $BMCL$ transition systems can be encoded as input to the Mocha model-checker and how properties, such as existence of derivations with given bounds on memory, communication, and the number of inference steps, can be verified automatically.

In future work, we plan to consider logical languages containing primitive operators which would allow us to state the agents' resource limitations as formulas in the language rather than axioms, and consider agents reasoning about each other's resource limitations. We also would like to consider agents reasoning in a simple epistemic or description logic.

Acknowledgements: This work was partially supported by the EPSRC project EP/E031226/1.

References

[1] P. Adjiman, P. Chatalic, F. Goasdoué, M.-C. Rousset, and L. Simon. Distributed reasoning in a peer-to-peer setting. In *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI'2004)*, pages 945–946. IOS Press, 2004.

[2] T. Ågotnes and M. Walicki. Strongly complete axiomatizations of "knowing at most" in standard syntactic assignments. In *Proceedings of the 6^{th} International Workshop on Computational Logic in Multi-agent Systems (CLIMA VI)*, 2005.

[3] A. Albore, N. Alechina, P. Bertoli, C. Ghidini, B. Logan, and L. Serafini. Model-checking memory requirements of resource-bounded reasoners. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 2006)*, pages 213–218. AAAI Press, 2006.

[4] N. Alechina, M. Jago, and B. Logan. Modal logics for communicating rule-based agents. In *Proceedings of the 17th European Conference on Artificial Intelligence (ECAI'2006)*, pages 322–326. IOS Press, 2006.

[5] N. Alechina, B. Logan, and M. Whitsey. A complete and decidable logic for resource-bounded agents. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2004)*, pages 606–613. ACM Press, 2004.

[6] M. Alekhnovich, E. Ben-Sasson, A. A. Razborov, and A. Wigderson. Space complexity in propositional calculus. *SIAM J. of Computing, 31(4)*, pages 1184–1211, 2002.

[7] R. Alur, T. A. Henzinger, F. Y. C. Mang, S. Qadeer, S. K. Rajamani, and S. Tasiran. MOCHA: Modularity in model checking. In *Computer Aided Verification*, pages 521–525, 1998.

[8] E. Amir and S. A. McIlraith. Partition-based logical reasoning for first-order and propositional theories. *Artificial Intelligence, 162(1-2)*, pages 49–88, 2005.

[9] M. Benerecetti, F. Giunchiglia, and L. Serafini. Model checking multiagent systems. *J. Log. Comput., 8(3)*, pages 401–423, 1998.

[10] R. Bordini, W. V. M. Fisher, and M. Wooldridge. State-space reduction techniques in agent verification. In *Proc. of the Third International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS-2004)*, pages 896–903. ACM Press, 2004.

[11] H. Duc. Reasoning about rational, but not logically omniscient, agents. *J. Log. Comput., 5*, pages 633–648, 1997.

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

[13] J. L. Esteban and J. Torán. Space bounds for resolution. In *Proc. of the 16th Annual Symposium on Theoretical Aspects of Computer Science (STACS 99)*, pages 551–560. Springer, 1999.

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

[15] B. Faltings and M. Yokoo. Introduction: Special issue on distributed constraint satisfaction. *J. of Artif. Intell., 161(1-2)*, pages 1–5, 2005.

[16] M. Fisher and C. Ghidini. Programming Resource-Bounded Deliberative Agents. In *Proc. of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI'99)*, pages 200–206. Morgan Kaufmann, 1999.

[17] J. Grant, S. Kraus, and D. Perlis. A logic for characterizing multiple bounded agents. *Autonomous Agents and Multi-Agent Systems, 3(4)*, pages 351–387, 2000.

[18] J. Y. Halpern, Y. Moses, and M. Y. Vardi. Algorithmic knowledge. In *Proc. of the 5th Conference on Theoretical Aspects of Reasoning about Knowledge*, pages 255–266. Morgan Kaufmann, 1994.

[19] M. Jago. *Logics for Resource-Bounded Agents*. PhD thesis, University of Nottingham, 2006.

[20] H. Jung and M. Tambe. On communication in solving distributed constraint satisfaction problems. In *Multi-Agent Systems and Applications IV, Proc. 4th International Central and Eastern European Conference on Multi-Agent Systems, CEEMAS 2005*, pages 418–429. Springer, 2005.

[21] K. Konolige. *A Deduction Model of Belief.* Morgan Kaufmann, San Francisco, Calif., 1986.

[22] G. M. Provan. A model-based diagnosis framework for distributed embedded systems. In *Proc. of the Eighth International Conference on Principles and Knowledge Representation and Reasoning (KR-02)*, pages 341–352. Morgan Kaufmann, 2002.

[23] R. Pucella. Deductive algorithmic knowledge. In *AI&M 1-2004, Eighth International Symposium on Artificial Intelligence and Mathematics*, 2004.

[24] M. Reynolds. An axiomatization of $PCTL^*$. *Inf. Comput., 201(1)*, pages 72–119, 2005.

[25] A. C.-C. Yao. ome complexity questions related to distributive computing(preliminary report). In *Conference Record of the Eleventh Annual ACM Symposium on Theory of Computing*, pages 209–213. ACM, 1979.

ETL, DEL, and Past Operators

Tomohiro Hoshi
Stanford University
thoshi@stanford.edu

Audrey Yap
University of Victoria
ayap@uvic.ca

Abstract

[8] merges the semantic frameworks of Dynamic Epistemic Logic DEL *([1, 3]) and* Epistemic Temporal Logic ETL *([2, 6]). We consider the logic* TDEL *on the merged semantic framework and its extension with the labeled past-operator "P_ϵ" ("The event ϵ has happened before which..."). To axiomatize the extension, we introduce a method to transform a given model into a* normal *form in a suitable sense. These logics suggest further applications of* DEL *in the theory of agency, the theory of learning, etc.*

1. Introduction

[8] provides a framework for generating the models of *Epistemic Temporal Logic* (ETL, [2, 6]) from the models of *Dynamic Epistemic Logic* (DEL, [1, 3]). In the framework, the temporal transitions in DEL are captured by sequences of event models, called *DEL-protocols*, and each transition made by a product update is encoded into the tree structures of ETL. This allows us to say that DEL-models *generate* ETL-models. The framework allows for a systematic comparison between the two major trends, DEL and ETL, in describing agents' intelligent interactions, and suggests a direction for the studies of new logics that are hybrids of the two.

The main objective of the present paper is to push that investigation further. [8] studies the logic TPAL of ETL-models generated by protocols consisting of public announcements. However, public announcements are just one kind of event model. Thus we might ask what the logic would be like if we extend the setting of TPAL to the full class of event models. In Section 2, we apply the basic methods in TPAL and obtain an axiomatization of the class of the ETL-models generated from the class of all DEL-protocols. We call this extended system TDEL.

After axiomatizing TDEL, in Section 4 we will study the extension of TDEL with the labeled past-operator P_ϵ, where P_ϵ reads as "the event ϵ has occurred before which φ." We call the resulting system TDEL+P. This is a very natural operator to add to the context of TDEL, where all successive updates by event models are encoded as tree structures. A similar operator has been investigated in [12] in the original DEL-context; our objective in the present paper can be characterized as investigating that operator in the TDEL-context.

The axiomatization of TDEL+P will be based on one distinctive feature of the DEL-generated ETL-models. Given a set X of event models, DEL-generated ETL-models can be transformed into the models that consist only of the event models in X or event models with trivial preconditions, and this transformation preserves the truth of formulas whose only event models are those in X. We call this model transformation *normalization*. In Section 3, we will show that DEL-generated ETL-models can be normalized in this sense, and will apply this fact to the axiomatization of TDEL+P.

TDEL and its extension TDEL+P suggest further interesting applications in the theory of agency and the theory of learning. In modeling agency, some systems model intentionality in terms of agents' goals to bring about certain states. And, for instance in [7], for an agent to intend to bring about a state at which φ holds, it is not sufficient for her just to bring about that state. In the history leading up to that state, she must also have believed that her actions would lead to a φ state (so she does not bring it about by accident). This seems exactly to call for a way to express what an agent used to believe, about what was then her future. Also, when expressing that an agent learned something from an event, we want to be able to say something like, "After ϵ took place, i knew that φ. But before ϵ, i did not know φ." Expressing this sentence requires both a future and a past modality. We will discuss these issues further in Section 5.

2. TDEL

We start by generating ETL-models from DEL-models, though a detailed exposition for ETL and DEL is omitted. Readers who are not familiar with the systems are invited to refer to e.g. [2, 6] for ETL and to e.g. [10] for DEL. Below, we fix a finite set $\mathcal{A}$ of agents and a countable set At of propositional letters.

2.1. DEL-Generated ETL-Models

Definition 2.1 An *epistemic model* $\mathcal{M}$ is a tuple $\langle W, \sim, V\rangle$, where W is a nonempty set, $\sim : \mathcal{A} \to W \times W$, and $V : At \to 2^W$. The set W represents the set of possible situations, $\sim$, the indistinguishability relation over the possible situations for an agent i, and V, the valuation function. We denote W, $\sim$ and V by $Dom(\mathcal{M})$, $Rel(\mathcal{M})$, and $Val(\mathcal{M})$ respectively. Also, we write $\sim_i$ for $\sim(i)$ by convention. ◁

Definition 2.2 An *event model* $\mathcal{E}$ is a tuple $\langle E, \to, \mathsf{pre}\rangle$, where E is a nonempty set, $\to : \mathcal{A} \to E \times E$, and $\mathsf{pre} : E \to \mathcal{L}_{EL}$, where $\mathcal{L}_{EL}$ is the set of epistemic formulas. E represents the set of possible events, $\to_i$, the indistinguishability relation over the possible events for an agent i, and pre assigns the preconditions for the possible events. We denote the domain E of $\mathcal{E}$ by $Dom(\mathcal{E})$, and write $\to_i$ for $\to(i)$ by convention. ◁

Let $\mathbb{E}$ be the class of pointed event models $(\mathbb{E}, e)$. Let $\mathbb{E}^*$ be the class of finite sequences of pointed event models.

Definition 2.3 A *DEL-protocol* is a set $\mathsf{P} \subseteq \mathbb{E}^*$, which is closed under finite prefix. Let $ptcl(\mathbb{E})$ be the class of DEL-protocols. Given an epistemic model $\mathcal{M}$, a *state-dependent DEL-protocol* is a function $\mathsf{p} : Dom(\mathcal{M}) \to ptcl(\mathbb{E})$. ◁

Given a sequence $\sigma = \epsilon_1 \ldots \epsilon_n \in \mathbb{E}^*$, we write $\sigma_{(n)}$ for the initial segment of σ of length n ($n \leq \mathsf{len}(\sigma)$), and σ_n for the nth component of σ. When $n > \mathsf{len}(\sigma)$ or $n = 0$, σ_n and $\sigma_{(n)}$ are empty. If $\sigma = (\mathcal{E}_1, e_1)(\mathcal{E}_2, e_2)\ldots(\mathcal{E}_n, e_n) \in \mathbb{E}^*$, we write σ^L and σ^R for $\mathcal{E}_1 \cdots \mathcal{E}_n$ and $e_1 \cdots e_n$ respectively. Thus, for example, if $\sigma = (\mathcal{E}_1, e_1)\ldots(\mathcal{E}_n, e_n)$, then $(\sigma^L)_{(3)} = \mathcal{E}_1\mathcal{E}_2\mathcal{E}_3$ and $(\sigma^R)_3 = e_3$. Clearly, $(\cdot)^L, (\cdot)^R$ on the one hand and $(\cdot)_n, (\cdot)_{(n)}$ on the other commute. Thus, we omit parentheses when there is no danger of ambiguity.

Definition 2.4 (σ^L-Generated Model) Let $\mathcal{M} = \langle W, \sim, V\rangle$ be an epistemic model and p, a state-dependant DEL-protocol on $\mathcal{M}$. Given a sequence $\sigma \in \mathbb{E}^*$, the σ^L-*generated model*, $\mathcal{M}^{\sigma^L, \mathsf{p}} = \langle W^{\sigma^L, \mathsf{p}}, \sim_i^{\sigma^L, \mathsf{p}}, V^{\sigma^L, \mathsf{p}}\rangle$, is defined by induction on the initial segment of σ^L:

- $W^{\sigma^L_{(0)}, \mathsf{p}} := W$, for each $i \in \mathcal{A}$, $\sim_i^{\sigma^L_{(0)}, \mathsf{p}} := \sim_i$ and $V^{\sigma^L_{(0)}, \mathsf{p}} := V$.
- $w\tau \in W^{\sigma^L_n, \mathsf{p}}$ iff
  1. $w \in W$,
  2. $\sigma^L_{(n)} = \tau^L$,
  3. $w\tau_{(n-1)} \in W^{\sigma^L_{(n-1)}, \mathsf{p}}$,
  4. $\tau \in \mathsf{p}(w)$, and
  5. $\mathcal{M}^{\sigma^L_{(n-1)}, \mathsf{p}}, w\tau_{(n-1)} \models \mathsf{pre}(\tau^R_n)$
- For each $w\tau, v\tau' \in H_n$ ($0 < n < \mathsf{len}(\sigma^L)$), $w\tau \sim^{\sigma^L_{(n)}} v\tau'$ iff $w\tau_{(n-1)} \sim_i^{\sigma^L_{(n-1)}, \mathsf{p}} v\tau'_{(n-1)}$ and $\tau^R_n \to_i (\tau')^R_n$ in τ^L_n.
- For each $p \in \mathsf{At}$, $V^{n+1, \mathsf{p}}(p) = \{w\sigma \in W^{n+1, \mathsf{p}} \mid w \in V(p)\}$.

Note that, in the definition of $\sim_i$, $\tau^L = (\tau')^L = \sigma^L_n$, and thus $\sigma^L = (\sigma')^L$. ◁

Definition 2.5 (DEL-Generated ETL-Model) Let $\mathcal{M} = \langle W, \sim, V\rangle$ be an epistemic model and p a state-dependent DEL-protocol on $\mathcal{M}$. *An* ETL-*model* $\mathsf{Forest}(\mathcal{M}, \mathsf{p}) = \langle \mathsf{H}, \sim, U\rangle$ *generated from* $\mathcal{M}$ *by* p is defined as follows:

- $\mathsf{H} := \{h \mid \exists w \in W, \sigma \in \bigcup_{w \in W} p(w)$ such that $h = w\sigma \in W^{\sigma^L, \mathsf{p}}\}$.
- For all $h, h' \in \mathsf{H}$ with $h = w\sigma$ and $h' = v\sigma'$, $h \sim_i h'$ iff $w\sigma \sim_i^{\sigma^L, \mathsf{p}} v\sigma'$.
- For each $p \in \mathsf{At}$ and $h = w\sigma \in \mathsf{H}$, $h \in V'(p)$ iff $h \in V^{\sigma^L, \mathsf{p}}(p)$.

We define the class $\mathbb{F}_{st}(\mathbb{E})$ to be the class of all ETL-models of the form $\mathsf{Forest}(\mathcal{M}, \mathsf{p})$. ◁

Given $X \subseteq \mathbb{E}$, we denote by $\mathbb{F}_{sd}(X)$ the class of ETL-models generated from epistemic models $\mathcal{M}$ by state-dependent protocols p consisting only of elements in X, i.e., for every w in $\mathcal{M}$, if $\sigma \in \mathsf{p}(w)$, $\sigma \subseteq X^*$.

Example 2.6 (Public Announcements) We illustrate the above construction in public announcement logic with each event model denoting an announcement or observation of some true formula. Let $\mathcal{M}$ be a model that consists of w, v, u, each of which are indistinguishable (the $\sim$ relation in $\mathcal{M}$ is an equivalence relation on w, v, u), where $V(p) = \{w, v\}$ and $V(q) = \{v\}$. This model is represented by the three points labeled with w, v, u, respectively at the bottom of Figure 1. Consider the protocol p where $\mathsf{p}(w) = \{p, pq, \neg q\}$, $\mathsf{p} = \{p, pq, \neg q\}$ and $\mathsf{p} = \{\neg q, \neg q\top, p\}$. The DEL-generated ETL-model $\mathsf{Forest}(\mathcal{M}, \mathsf{p})$ can be visualized as follows:

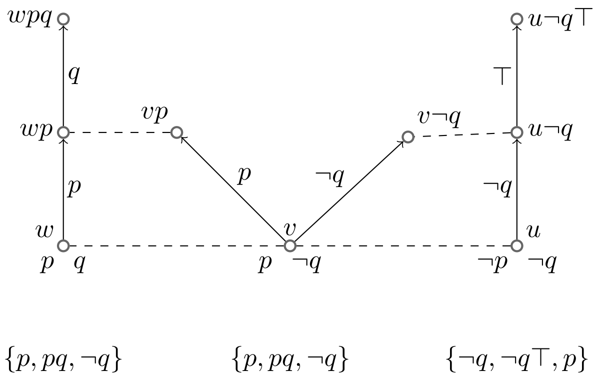


Figure 1. A DEL-generated ETL model.

2.2. Axiomatization of TDEL

The language $\mathcal{L}_{TDEL}$ of TDEL extends the language $\mathcal{L}_{EL}$ of epistemic logic by the operator $\langle\epsilon\rangle$, where $\epsilon \in \mathbb{E}$. The dual of $\langle\epsilon\rangle$ is $[\epsilon]$ defined by $\neg\langle\epsilon\rangle\neg$ as usual.

Let $\mathcal{H} \in \mathbb{F}_{sd}(\mathbf{E})$ with

$$\mathcal{H} = \mathsf{Forest}(\mathcal{M}, \mathsf{p}) = \langle \mathsf{H}, \{\sim_i\}_{i\in\mathcal{A}}, V\rangle.$$

The semantics of the knowledge operator and the event model operator are defined by:

- $\mathcal{H}, h \models K\varphi$ iff for all h' such that $h \sim_i h'$, $\mathcal{H}, h' \models \varphi$.
- $\mathcal{H}, h \models \langle\epsilon\rangle\varphi$ iff $h\epsilon \in \mathsf{H}$ and $\mathcal{H}, h\epsilon \models \varphi$.

The boolean cases are defined in the standard way.

Example 2.7 (Semantics in TDEL) Let $\mathcal{H}$ be the model $\mathsf{Forest}(\mathcal{M}, \mathsf{p})$ in Figure 1. For instance, we have $\mathcal{H}, w \models \langle p\rangle\langle q\rangle K(p\wedge q)$ but $\mathcal{H}, w \not\models \langle p\wedge q\rangle K(p\wedge q)$. This illustrates the fact that in TDEL we cannot treat sequences of events as single events, while in DEL we can. Also the fact that we have $\mathcal{H}, w \models (p \wedge q) \wedge \neg\langle p \wedge q\rangle\top$ violates the schema $\langle\epsilon\rangle\top \leftrightarrow \mathsf{pre}(\epsilon)$, which is valid in DEL. In TDEL, we only have $\langle\epsilon\rangle\top \rightarrow \mathsf{pre}(\epsilon)$.

Definition 2.8 The axiomatization TDEL of $\mathbb{F}_{sd}(\mathbb{E})$ is given by the following axiom schemes and inference rules.

Axioms

PC Propositional validities

K_i $K_i(\varphi \rightarrow \psi) \rightarrow (K_i\varphi \rightarrow K_i\psi)$

F1 $\langle\epsilon\rangle p \leftrightarrow \langle\epsilon\rangle\top \wedge p$

F2 $\langle\epsilon\rangle\neg\varphi \leftrightarrow \langle\epsilon\rangle\top \wedge \neg\langle\epsilon\rangle\varphi$

F3 $\langle\epsilon\rangle(\varphi \wedge \psi) \leftrightarrow \langle\epsilon\rangle\varphi \wedge \langle\epsilon\rangle\psi$

F4 $\langle\epsilon\rangle K_i\varphi \leftrightarrow \langle\epsilon\rangle\top \wedge \bigwedge_{\{(\epsilon')^R \in Dom(\epsilon^L) | \epsilon^R \rightarrow_i (\epsilon')^R\}} K_i(\langle\epsilon'\rangle\top \rightarrow \langle\epsilon'\rangle\varphi)$

A1 $\langle\epsilon\rangle(\varphi \rightarrow \psi) \rightarrow (\langle\epsilon\rangle\varphi \rightarrow \langle\epsilon\rangle\psi)$

A2 $\langle\epsilon\rangle\top \rightarrow \mathsf{pre}(\epsilon^R)$

Inference Rules

MP If $\vdash \varphi \rightarrow \psi$ and $\vdash \varphi$, then $\vdash \psi$.

k-Nec If $\vdash \varphi$, then $\vdash K_i\varphi$.

e-Nec If $\vdash \varphi$, then $\vdash [\epsilon]\varphi$.

$\lhd$

Readers are invited to verify that these are sound with respect to $\mathbb{F}_{sd}(\mathbb{E})$.

2.3. Completeness Proof

The proof is given by a variant of the Henkin-style construction. The basic construction is the same as the one in [8] with minor modifications.

Definition 2.9 (Legal Histories) Let W_0 be the set of all TDEL-maximal consistent sets. We define λ_n and H_n ($0 \leq n \leq d(\Sigma)$) as follows:

- Define $H_0 = W_0$ and for each $w \in H_0$, $\lambda_0(w) = w$.
- Let $H_{n+1} = \{h\epsilon \mid h \in H_n \text{ and } \langle\epsilon\rangle\top \in \lambda_n(h)\}$. For each $h = h'\epsilon \in H_{n+1}$, define $\lambda_{n+1}(h) = \{\varphi \mid \langle\epsilon\rangle\varphi \in \lambda_n(h')\}$.

Given $h \in H_n$, we write $\lambda(h)$ for $\lambda_n(h)$. $\lhd$

The following can be straightforwardly verified by appealing to the construction and F2.

Lemma 2.10 *For each $n \geq 0$, for each $\sigma \in H_n$, $\lambda_n(\sigma)$ is a maximally consistent set.*

Let $\mathcal{H}_0^{can} = (H_0, \sim^0, V^0)$, where $\sim^0$ and V^0 are defined by

- $w \sim_i^0 v$ iff $\{\varphi \mid K_i\varphi \in w\} \subseteq v$.
- For each $p \in \mathsf{At}$ and $w \in H_0$, $p \in V(w)$ iff $p \in w$.

Definition 2.11 (Canonical Model) The canonical model $\mathcal{H}^{can}$ is a triple $\langle \mathsf{H}^{can}, \{\sim_i^{can}\}_{i\in\mathcal{A}}, V^{can}\rangle$, where each item is defined as follows:

- $\mathsf{H}^{can} =_{\text{def}} \bigcup_{i=0}^{\infty} H_i$.
- For each $w\sigma, w'\sigma' \in \mathsf{H}^{can}$, $w\sigma \sim_i^{can} w'\sigma'$ iff$_{\text{def}}$ $w\sigma \sim_i^{\sigma^L} w'\sigma'$, where $\sim^{\sigma^L}$ is defined by induction in the following way:

- $\sim_i^{\sigma^L_{(0)}} = \sim_i^0$
- For each $w\tau, v\tau' \in H_n$ $(0 < n < \mathsf{len}(\sigma^L))$, $w\tau \sim^{\sigma^L_{(n)}} v\tau'$ iff $w\tau_{(n-1)} \sim_i^{\sigma^L_{(n-1)}} v\tau'_{(n-1)}$ and $\tau_n^R \rightarrow_i (\tau')_n^R$ in τ_n^L.

- For every $P \in \mathsf{At}$ and $h = w\sigma \in \mathsf{H}^{can}$, $w\sigma \in V^{can}(P)$ iff $w \in V^0(P)$.

◁

Proposition 2.12 *Let $w\sigma \sim_i^{can} v\tau$ with $w, v \in W^0$, $\sigma = \sigma_1 \ldots \sigma_n$ and $\tau = \tau_1 \ldots \tau_n$. If $K_i\varphi \in \lambda(w\sigma)$, then* $K_i(\langle\tau_1\rangle\top \rightarrow \langle\tau_1\rangle(\langle\tau_2\rangle\top \rightarrow \langle\tau_2\rangle(\ldots(\langle\tau_n\rangle\top \rightarrow \langle\tau_n\rangle\varphi)\ldots) \in \lambda(w)$.

Proof. By induction on n. When $n = 0$, σ, τ are empty and thus the claim clearly holds. For the inductive step, assume that $K_i\varphi \in \lambda(\sigma)$. Then, by the construction of $\mathcal{H}^{can}$, $\langle\sigma_n\rangle K_i\varphi \in \lambda(w\sigma_{n-1})$. By F4, for all events e in σ_n^L such that $\sigma_n^R \rightarrow_i e$:

$$K_i(\langle\sigma_n^L, e\rangle\top \rightarrow \langle\sigma_n^L, e\rangle\varphi) \in \lambda(w\sigma_{(n-1)}).$$

Here, by the construction of $\mathcal{H}^{can}$, $\sigma_n \rightarrow_i \tau_n$. By applying the IH, we are done. QED

Lemma 2.13 (Truth Lemma) *For every $\varphi \in \mathcal{L}_{\mathsf{TDEL}}$ and $h \in \mathsf{H}^{can}$,*

$$\varphi \in \lambda(h) \quad \textit{iff} \quad \mathcal{H}^{can}, h \models \varphi.$$

Proof. We show by induction on the structure of $\varphi \in \mathcal{L}_{\mathsf{TDEL}}$ that for each $h \in \mathsf{H}^{can}$, $\varphi \in \lambda(h)$ iff $\mathcal{H}^{can}, h \models \varphi$. The base and the boolean cases are straightforward.

For the knowledge modality, let $h \in \mathsf{H}^{can}$ with $h = w\sigma_1 \cdots \sigma_n$ $(w \in W_0)$ and assume $K_i\psi \in \lambda(h)$. Suppose $h' \in \mathsf{H}^{can}$ with $h \sim_i^{can} h'$. By construction of the canonical model, we know that $h' = v\tau_1 \cdots \tau_n$ for some $v \in H_0$ and $\tau_1 \ldots \tau_n \in \mathbb{E}^*$ with $w \sim_i^0 v$. By Proposition 2.12, we have $K_i(\langle\tau_1\rangle\top \rightarrow \langle\tau_1\rangle(\langle\tau_2\rangle\top \rightarrow \langle\tau_2\rangle(\cdots\langle\tau_{n-1}\rangle(\langle\tau_n\rangle\top \rightarrow \langle\tau_n\rangle\psi)\cdots)) \in \lambda(w)$.
Since $w \sim_i^0 v$, we have by the construction of $\mathcal{H}^{can}$,
$\langle\tau_1\rangle\top \rightarrow \langle\tau_1\rangle(\langle\tau_2\rangle\top \rightarrow \langle\tau_2\rangle(\cdots\langle\tau_{n-1}\rangle(\langle\tau_n\rangle\top \rightarrow \langle\tau_n\rangle\psi)\cdots) \in \lambda(v)$.
Now note that
$\langle\tau_1\rangle\top \in \lambda(v), \langle\tau_2\rangle\top \in \lambda(v\tau_1), \ldots, \langle\tau_n\rangle\top \in \lambda(v\tau_1 ... \tau_{n-1})$.
Thus, we have
$\langle\tau_2\rangle\top \rightarrow \langle\tau_2\rangle(\cdots\langle\tau_{n-1}\rangle(\langle\tau_n\rangle\top \rightarrow \langle\tau_n\rangle\psi)\cdots) \in \lambda(v\tau_1)$
$\langle\tau_3\rangle\top \rightarrow \langle\tau_3\rangle(\cdots\langle\tau_{n-1}\rangle(\langle\tau_n\rangle\top \rightarrow \langle\tau_n\rangle\psi)\cdots) \in \lambda(v\tau_1\tau_2)$
$\vdots$
$\langle\tau_n\rangle\psi \in \lambda(v\tau_1 \cdots \tau_{n-1})$

Therefore, $\psi \in \lambda(v\tau_1 \cdots \tau_n) = \lambda(h')$. By the induction hypothesis, $\mathcal{H}^{can}, h' \models \psi$. Therefore, $\mathcal{H}^{can}, h \models K_i\psi$, as desired.

For the other direction, let $h \in \mathsf{H}^{can}$ and assume $K_i\psi \notin \lambda(h)$. For simplicity, let $h = w\sigma_1$ with $w \in W_0$ and $\sigma_1 \in \mathbb{E}$. The argument can easily be generalized to deal with the general case along the lines of the argument above. Since $\lambda(h)$ is a maximally consistent set, we have $\neg K_i\psi \in \lambda(h)$. Thus, by Definition 2.9, $\langle\sigma_1\rangle\neg K_i\psi \in \lambda(w)$. Using axiom $F2$, $\neg\langle\sigma_1\rangle K_i\psi \in \lambda(w)$; and so, by Axiom $F4$, $\neg\langle\sigma_1\rangle\top \vee \neg\bigwedge_{\{\tau|\sigma_1 \rightarrow_i \tau \text{ in } \sigma_1^L\}} K_i(\langle\tau\rangle\top \rightarrow \langle\tau\rangle\psi) \in \lambda(w)$. Since $\langle\sigma_1\rangle\top \in \lambda(w)$ by construction, it follows that $\neg\bigwedge_{\{\tau|\sigma_1 \rightarrow_i \tau \text{ in } \sigma_1^L\}} K_i(\langle\tau\rangle\top \rightarrow \langle\tau\rangle\psi) \in \lambda(w)$.

Now consider the set $v_0 = \{\theta \mid K_i\theta \in \lambda(w)\} \cup \{\neg\bigwedge_{\{\tau|\sigma_1 \rightarrow_i \tau \text{ in } \sigma_1^L\}}(\langle\tau\rangle\top \rightarrow \langle\tau\rangle\psi)\}$. We claim that this set is consistent. Suppose not. Then, there are formulas $\theta_1, \ldots, \theta_m$ such that $\vdash \bigwedge_{j=1}^m \theta_j \rightarrow \bigwedge_{\{\tau|\sigma_1 \rightarrow_i \tau \text{ in } \sigma_1^L\}}(\langle\tau\rangle\top \rightarrow \langle\tau\rangle\psi)$ and for $j = 1, \ldots, m$, $K_i\theta_j \in \lambda(w)$.

By standard modal reasoning, $\vdash \bigwedge_{j=1}^m K_i\theta_j \rightarrow \bigwedge_{\{\tau|\sigma_1 \rightarrow_i \tau \text{ in } \sigma_1^L\}} K_i(\langle\tau\rangle\top \rightarrow \langle\tau\rangle\psi)$. This implies that $\bigwedge_{\{\tau|\sigma_1 \rightarrow_i \tau \text{ in } \sigma_1^L\}} K_i(\langle\tau\rangle\top \rightarrow \langle\tau\rangle\psi) \in \lambda(w)$. However, this contradicts the fact that $\neg\bigwedge_{\{\tau|\sigma_1 \rightarrow_i \tau \text{ in } \sigma_1^L\}} K_i(\langle A\rangle\top \rightarrow \langle A\rangle\psi) \in \lambda(w)$, since $\lambda(w)$ is a maximally consistent set.

Now using standard arguments (Lindenbaum's lemma), there exists a maximally consistent set v with $v_0 \subseteq v$. By the construction of v, we must have $w \sim_i^0 v$. Also, since v is an mcs such that $\neg\bigwedge_{\{\tau|\sigma_1 \rightarrow_i \tau \text{ in } \sigma_1^L\}}(\langle A\rangle\top \rightarrow \langle A\rangle\psi) \in \lambda(v)$, there is some τ_1 such that $\neg(\langle\tau_1\rangle\top \rightarrow \langle\tau_1\rangle\psi \in \lambda(v)$. Otherwise, v is inconsistent. Therefore, for such τ_1, we have $\langle\tau_1\rangle\top \in \lambda(v), \neg\langle\tau_1\rangle\psi \in \lambda(v)$. Here, by axiom $F2$, $\langle\tau_1\rangle\neg\psi \in \lambda(v)$. Hence $\neg\psi \in \lambda(v\tau_1)$ and therefore $\psi \notin \lambda(v\tau)$. By the induction hypothesis, $\mathcal{H}^{can}, v\tau_1 \not\models \psi$. This implies $\mathcal{H}^{can}, w\tau_1 \not\models K_i\psi$, as desired.

For the event model operator, assume that $\langle\epsilon\rangle\psi \in \lambda(h)$. Since $\langle\epsilon\rangle\top \in \lambda(h)$ (for $\neg\langle\epsilon\rangle\top \in \lambda(h)$ makes $\lambda(h)$ inconsistent), $\psi \in \lambda(h\epsilon)$. By the induction hypothesis, we have $\mathcal{H}^{can}, h\epsilon \models \psi$, which implies $\mathcal{H}^{can}, h \models \langle\epsilon\rangle\psi$.

For the other direction, assume $\mathcal{H}^{can}, h \models \langle\epsilon\rangle\psi$. Then, $\mathcal{H}^{can}, h\epsilon \models \psi$. By the inductive hypothesis, we have $\psi \in \lambda(h\epsilon)$ and thus $\langle\epsilon\rangle\psi \in \lambda(h)$. QED

All that remains is to show is that $\mathcal{H}^{can}$ is in the class of intended models (i.e., is an element of $\mathbb{F}_{sd}(\mathbb{E})$).

Lemma 2.14 *The canonical model $\mathcal{H}^{can}$ is in $\mathbb{F}_{sd}(\mathbb{E})$. That is, there is an epistemic model $\mathcal{M}$ and local protocol p on $\mathcal{M}$ such that $\mathcal{H}^{can} = \mathsf{Forest}(\mathcal{M}, \mathsf{p})$.*

Proof. Let $\mathcal{M}_{can} = (W_0, \{\sim_i^0\}_{i \in \mathcal{A}}, V^0)$ and define $\mathsf{p}_{can} : W_0 \rightarrow \mathbb{E}^*$ so that $\mathsf{p}_{can}(w) = \{\sigma \mid w\sigma \in \mathsf{H}^{can}\}$. Suppose that $\mathcal{H}^{p_{can}} = \mathsf{Forest}(\mathcal{M}_{can}, \mathsf{p}_{can})$. We claim that $\mathcal{H}^{can}$ and $\mathcal{H}^{p_{can}}$ are the same model. For this, it suffices to show that for all $w \in W_0$ and $\sigma \in \mathbb{E}^*$ we have $w\sigma \in \mathsf{H}^{can}$ iff $w\sigma \in W^{\sigma,p_{can}}$. For this implies $\mathsf{H}^{can} = \mathsf{H}^{p_{can}}$, where $\mathsf{H}^{p_{can}}$ is the domain of $\mathcal{H}^{p_{can}}$. Then, by inspecting the construction of Forest and Definition 2.11, we see that $\mathcal{H}^{can}$ and $\mathcal{H}^{p_{can}}$ are the same model .

We will show by induction on the length of $\sigma \in \mathcal{E}^*$ that for any $w \in W_0$, $w\sigma \in \mathsf{H}^{can}$ iff $w\sigma \in W^{\sigma,p_{can}}$. The base case ($\mathsf{len}(\sigma) = 0$) is clear. Assume that the claim holds for all σ with $\mathsf{len}(\sigma) = n$.

Given any $\sigma \in \mathbb{E}^*$ with $\mathsf{len}(\sigma) = n$, we first show by subinduction (on the structure of A) that, for all $A \in \mathcal{L}_{EL}$, $\mathcal{H}^{can}, w\sigma \models A$ iff $\mathcal{M}^{\sigma,p_{can}}, w\sigma \models A$. The base and boolean cases are straightforward. Suppose that $\mathcal{H}^{can}, w\sigma \models K_iB$. We must show $\mathcal{M}^{\sigma,p_{can}}, w\sigma \models K_iB$. Let $v\sigma \in W^{\sigma,p_{can}}$ with $w\sigma \sim_i^{\sigma,p} v\sigma$. By the main induction hypothesis, we have both $v\sigma \in \mathsf{H}^{can}$ and $w\sigma \in W^{\sigma,p_{can}}$. By construction, since $w\sigma \sim_i^{\sigma,p_{can}} v\tau$, we have $w \sim_i^0 v$. Furthermore, $w\sigma \sim_i^{can} v\tau$. Hence, $\mathcal{H}^{can}, v\sigma \models B$. By the subinduction hypothesis, $\mathcal{M}^{\sigma,p_{can}}, v\sigma \models B$. Therefore, $\mathcal{M}^{\sigma,p_{can}}, w\sigma \models K_iB$.

Coming back to the main induction, assume $w\sigma_{(n)}\sigma_{n+1} \in \mathsf{H}_{can}$. This implies that $\langle\sigma_{n+1}\rangle\top \in \lambda(w\sigma_{(n)})$. By truth lemma, we have $\mathcal{H}^{can}, w\sigma_{(n)} \models \langle\sigma_{n+1}\rangle\top$. This, together with axiom $A2$, implies $\mathcal{H}^{can}, w\sigma \models \mathsf{pre}(\sigma^R_{n+1})$. From the above subinduction, it follows that $\mathcal{M}^{\sigma_{(n)},p_{can}}, w\sigma_{(n)} \models \mathsf{pre}(\sigma^R_{n+1})$ (recall that $\mathsf{pre}(e) \in \mathcal{L}_{EL}$ for all events e by definition). Thus, by the construction of p_{can}, we have $w\sigma_{(n)}\sigma_{n+1} \in W^{\sigma_{(n)},p_{can}}$. This shows that if $w\sigma_{(n)}\sigma_{n+1} \in \mathsf{H}^{can}$ then $w\sigma_{(n)}\sigma_{n+1} \in W^{\sigma_{(n)}\sigma_{n+1},p_{can}}$. The other direction is similar. QED

The proof of the completeness theorem follows from Lemma 2.13 and Lemma 2.14 using a standard argument.

Theorem 2.15 TDEL *is sound and complete with respect to* $\mathbb{F}_{sd}(\mathbb{E})$.

2.4. TDEL Restricted to Some Class of Protocols

TDEL axiomatizes the class $\mathbb{F}_{sd}(\mathbb{E})$. However, note that the completeness proof above does not depend on the fact that $\mathbb{E}$ is the set of *all* pointed event models, but only the fact that $\mathbb{F}_{sd}(\mathbb{E})$ contains the ETL-models generated from epistemic models $\mathcal{M}$ by the protocol p that allows *all possible finite sequences of* $\mathbb{E}$ at each w in $\mathcal{M}$, i.e $\mathsf{p}(w) = \mathbb{E}^*$.

Thus, even if we restrict our attention to some $X \subseteq \mathbb{E}$, the proof should work as well for the class $\mathbb{F}_{sd}(X)$. However, here we have to be careful that such an X must at least contain all the "relevant" pointed event models: if $(\mathcal{E}, e) \in X$, then $(\mathcal{E}, f) \in X$ for all f such that $e \to f$ in $\mathcal{E}$. Otherwise the knowledge modality case of Lemma 2.13 since we need all the "relevant" histories in the present sense must be included in the canonical model.

Let $X \subseteq \mathbb{E}$. Call X *e-closed* if, for all $\mathcal{E}$, if there is $\epsilon \in X$ such that $\epsilon^L = \mathcal{E}$, then for every event e in $\mathcal{E}$, (ϵ^L, e) is in X. Denote by $\mathcal{L}_{\mathsf{TDEL}}(X)$ the fragment of $\mathcal{L}_{\mathsf{TDEL}}$ that only allows the event model operators $\langle\epsilon\rangle$ such that $\epsilon \in X$. Also, let $\mathsf{TDEL}(X)$ be the axiomatization as above except that the axiom schema and the $[\epsilon]$-necessitation rule can be instantiated by the event models in X. The following is a corollary of our completeness proof.

Corollary 2.16 *For all e-closed subsets* X *of* $\mathbb{E}$, $\mathsf{TDEL}(X)$ *is complete with respect to* $\mathbb{F}_{sd}(X)$.

Thus, by changing the parameter X, we could have axiomatizations for various kinds of logic of protocols. In fact, the logic of public announcement protocols, as is presented in [8] is a particular version of $\mathsf{TDEL}(X)$. We could also consider the logics of secret message protocols, etc.

3. Normalization of DEL-Generated ETL-Models

Before we study the proposed extension, we need to turn our attention to a distinctive property of DEL-generated ETL-models. The rough idea is that, given a set X of event models, DEL-generated ETL-models can be transformed into the models that consist of the event models in X and the event models with trivial preconditions in such a way that the truth of the formulas expressed with event models in X is preserved. We call this model transformation *normalization*. To formulate this notion here, we need some definitions.

Definition 3.1 We say that two event models $(E, \to, \mathsf{pre})$ and $(E', \to', \mathsf{pre}')$ are *isomorphic*, if $(E, \to)$ and $(E', \to')$ are isomorphic. Clearly, such an isomorphic relation partitions the set of event models. Given an event model $\mathcal{E}$, let $[\mathcal{E}]$ be the class of event models isomorphic to $\mathcal{E}$. We call $[\mathcal{E}]$ *the type of* $\mathcal{E}$. Also given a finite e-closed subset X of $\mathbb{E}$, we denote by PRE_X the conjunction of the preconditions of the events that occur in X. ◁

Definition 3.2 (Normalization Function) Let X be a finite e-closed subset of $\mathbb{E}$. The *normalization function with respect to* X is a function $f_X : \mathbb{E} \to \mathbb{E}$ such that, for every pointed event model $(\mathcal{E}, e)$ with $\mathcal{E} = (E, \to, \mathsf{pre})$, $f_X((\mathcal{E}, e)) = (\mathcal{E}', e)$, where $\mathcal{E}' = (E', \to', \mathsf{pre}')$ is defined by:

- $E' = E$
- $\to' (i) = \to (i)$

- $\mathsf{pre}'(e) = \mathsf{pre}(e) \vee \neg\mathsf{pre}(e) \vee PRE_X$.

$\lhd$

The purpose of having this function is to replace certain pointed event models ϵ with isomorphic pointed models with tautologous preconditions. Therefore, this role of the normalization function does not turn on the particular form $(\mathsf{pre}(e) \vee \neg\mathsf{pre}(e))$ of the tautology, as given in the third clause of the definition. However, having the tautology of such a form, we can guarantee that, if $\epsilon \neq \epsilon'$, then $f_X(\epsilon) \neq f_X(\epsilon')$. Also the third disjunct in the third clause guarantees that, for all $\epsilon \in \mathbb{E}$, $f_X(\epsilon) \notin X$.

Definition 3.3 Given a finite e-closed subset X of $\mathbb{E}$, a *substitution function for X* is a function $\sigma_X : \mathbb{E} \to \mathbb{E}$ such that, for all $\epsilon \in \mathbb{E}$,

$$\sigma_X(\epsilon) = \begin{cases} \epsilon & \text{if } \epsilon \in X \\ f_X(\epsilon) & \text{otherwise} \end{cases}$$

Given a DEL-generated ETL-model $\mathcal{H}$ and a history $h = w\epsilon_1 \ldots \epsilon_n$ in $\mathcal{H}$, we denote $w\sigma_X(\epsilon_1) \ldots \sigma_X(\epsilon_n)$ by $\sigma_X(h)$. $\lhd$

Definition 3.4 (Normalization) Let X be an e-closed subset of $\mathbb{E}$. The *normalization* $\mathcal{H}\sigma_X$ of a DEL-generated ETL-model $\mathcal{H} = (H, \sim, V)$ with respect to X is a tuple $(H', \sim', V')$. σ_X that satisfies the following conditions:

$\mathsf{H}' := \{\sigma(h) \mid h \in \mathsf{H}\}$

$\sigma(h) \sim'_i \sigma(g)$ iff $h \sim_i g$.

$V'(p) := \{\sigma(h) \mid h \in V(p)\}$

$\lhd$

Example 3.5 (Normalization) We can now illustrate the manner in which a model can be normalized, and how that process depends on the set of event models we are interested in. The process uniformly replaces any event not in the set with an event that has tautological preconditions. Let our initial model be the one from Figure 1. If we normalized this model with respect to the set $\{p, q, \neg q, \top\}$, the model would not change, since this is the set of all events in the model. For the other extreme case, if we normalized with respect to the set $\emptyset$, indicating tautologous preconditions by indexed $\top$'s, we would obtain the following:

On the other hand, if we normalized with respect to some subset of the expressions in the model, we would replace some events and keep others.

Proposition 3.6 *Let $\mathcal{H}$ be a* DEL*-generated* ETL*-model. Then $\mathcal{H}\sigma_X$ is a* DEL*-generated* ETL*-model.*

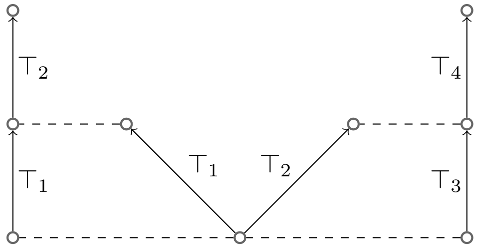


Figure 2. Normalizing Figure 1 with respect to $\emptyset$.

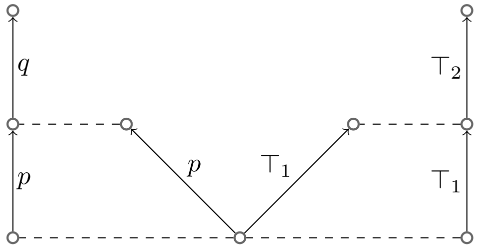


Figure 3. Normalizing Figure 1 with respect to $\{p, q\}$.

Proof. Let $\mathcal{H} = \mathsf{Forest}(\mathcal{M}, \mathsf{p}) = (H, \sim, V)$ and $\mathcal{H}\sigma_X = (H', \sim', V'))$. Let p_0^N be such that for all w in $\mathcal{M}$, $\mathsf{p}_0(w) = \{\sigma \mid w\sigma \in H'\}$. Then $\mathcal{H}\sigma_X = \mathsf{Forest}(\mathcal{M}, \mathsf{p}_0)$. The rest of the proof goes by an argument similar to the proof of Lemma 2.14. QED

Now it is straightforward to show that the normalization with respect to a given X preserves the truth of the formulas in which only the event operators from X occur.

Proposition 3.7 (Normalization) *Let X be an e-closed subset of $\mathbb{E}$. Then, for every DEL-generated model $\mathcal{H}$ and every formula φ in $\mathcal{L}_{DEL}(X)$ (the fragment of $\mathcal{L}_{DEL+P}$ that only allows the event models in X),*

$$\mathcal{H}, h \models \varphi \textit{ iff } \mathcal{H}\sigma_X, \sigma_X(h) \models \varphi.$$

Proof. We proceed by induction on φ. The base and boolean cases are clear. For the knowledge modality case, assume $\mathcal{H}, h \models K_i\psi$. Then, for all $h \sim h'$, $\mathcal{H}, h' \models \psi$. By IH, $\mathcal{H}\sigma_X, \sigma_X(h') \models \psi$. By Definition 3.4, we have $\mathcal{H}\sigma_X, \sigma_X(h) \models K_i\varphi$. The other direction is similar.

For the event modality, assume that $\mathcal{H}, h \models \langle\epsilon\rangle\psi$, where $\epsilon \in X$. Then $\mathcal{H}, h\epsilon \models \psi$. By the IH, $\mathcal{H}\sigma_X, \sigma_X(h\epsilon) \models \psi$. However, since $\epsilon \in X$, we have $\mathcal{H}\sigma_X, \sigma(h)\epsilon \models \psi$. This gives $\mathcal{H}\sigma_X, \sigma_X(h) \models \langle\epsilon\rangle\psi$, as desired. The other direction is similar. QED

Note that, if we also replaced the pointed event models in X that occur in the given model, the truth of the formulas might not be preserved, since the truth definitions of the event model operator explicitly refer to given event models. To see this, suppose $\mathcal{H}, h\epsilon \models \langle\epsilon\rangle\varphi$. If we replaced ϵ in the model with the pointed event model ϵ' of the same type, but distinct from ϵ, $\langle\epsilon\rangle\varphi$ cannot be true by definition, simply because $\epsilon \neq \epsilon'$.

4. Extending TDEL with the Past Modality

One fact about TDEL is that it only has forward-looking operators $\langle\epsilon\rangle$. However, given that, in TDEL, we have the forest structres that encodes all successive stages of update by event models, we can naturally think about the operator that states what *was* the case prior to a given temporal point. In this section, we extend TDEL with a past-looking modality P_ϵ with $\epsilon \in \mathbb{E}$. This extension will be called TDEL+P. Also, given an e-closed subset X of $\mathbb{E}$, we denote the corresponding fragment of TDEL+P by TDEL+P(X).

Let $\mathcal{H} = (\mathsf{H}, \sim, V)$ be an ETL-model generated from an epistemic model and a state-dependent protocol. The semantics of the operator P_ϵ is defined as follows:

$$\mathcal{H}, h \models P_\epsilon\varphi \text{ iff } \exists h' \text{ such that } h = h'\epsilon \text{ and } \mathcal{H}, h' \models \varphi.$$

The dual of P_ϵ is denoted by $\hat{P}_\epsilon$. The reading of P_ϵ is "the event ϵ has happened, before which φ". The dual $\hat{P}_\epsilon$ reads as "Before the event ϵ, φ".

Let t_{PAL} be the type of event models consisting of single reflexive events. Below we show that, given an e-closed subset X of $\mathbb{E}$ such that X is a union of a finite number of types including t_{PAL}, TDEL+P(X) is axiomatizable. For this, we first observe that the normalization results hold for TDEL+P(X).

Proposition 4.1 *Let Y be an e-closed subset of $\mathbb{E}$. Then, for every DEL-generated model $\mathcal{H}$ and every formula φ in* TDEL+P*(X),*

$$\mathcal{H}, h \models \varphi \textit{ iff } \mathcal{H}\sigma_Y, \sigma_Y(h) \models \varphi.$$

Proof. We proceed by induction on φ. The cases other than P_ϵ are as in Lemma 3.7. Thus, assume $\mathcal{H}, h \models P_\epsilon\psi$. Then there must be some h' such that $h'\epsilon$ and $\mathcal{H}, h' \models \psi$. By the IH, $\mathcal{H}\sigma_Y, \sigma_Y(h') \models \psi$. Since $\epsilon \in Y$, $\sigma_Y(h'\epsilon) = \sigma(h\prime)\epsilon$. Thus, $\mathcal{H}\sigma_Y, \sigma_Y(h'\epsilon) \models P_\epsilon\psi$. The other direction is similar. QED

To present the axiomatization of TDEL+P, we need some definitions.

Definition 4.2 Given a formula φ, the *past depth* $d(\varphi)$ of the formula φ is defined as follows:

- $d(p) = 0$ for p propositional.
- $d(\neg\varphi) = d(\varphi)$
- $d(\varphi \wedge \psi) = max\{d(\varphi), d(\psi)\}$
- $d(K_i\varphi) = d(\varphi)$
- $d(\langle\epsilon\rangle\varphi) = d(\varphi) - 1$
- $d(P_\epsilon\varphi) = \max(d(\varphi), 0) + 1$

◁

The intuition behind this definition is that if a formula has a depth n, we would have to go n-steps into the past from the current point of the ETL-tree in order to verify it. Thus, the final clause reflects the intended meaning. Had the definition instead been $d(P_\epsilon\varphi) = d(\varphi) + 1$, this would not have worked for, $P_{(\mathcal{E}_1,e_1)}\langle\mathcal{E}_2, e_2\rangle\langle\mathcal{E}_3, e_3\rangle P$. That definition would mistakenly have set the past depth as -1 instead of 1.

Let X be a union of a finite number of types such that $t_{PAL} \subseteq X$, so X is a class of event models.

Definition 4.3 Given a finite set Σ of expressions in $\mathcal{L}_{\mathsf{TDEL+P}}$ and a type t, define $\mathbb{E}(\Sigma) := \bigcup_{\varphi\in\Sigma}\mathbb{E}(\varphi)$. Also denote by PRE_Σ the conjunction of the preconditions of the events in $\mathbb{E}(\Sigma)$. ◁

Definition 4.4 Given a type $t \subseteq X$, let $\mathcal{E}^t_\Sigma$ be a distinguished event of the type t in which the precondition of each event is the tautologous formula of the form $Pre_\Sigma \vee \neg Pre_\Sigma$. The role of $\mathcal{E}^t_\Sigma$ is to pick up one event model of the type t, whose precondition is tautologous and whose pointed event model is not in Σ. The form of the precondition is to prevent the pointed event model formed by $\mathcal{E}^t_\Sigma$ from being in Σ. ◁

Definition 4.5 Further, define the set $N_X(\Sigma)$ by:

$$N_X(\Sigma) := \{(\mathcal{E}^t, e) \mid t \subseteq X \text{ is a type and } e \text{ in } \mathcal{E}^t_\Sigma\}.$$

◁

Here, given the definition of $\mathcal{E}^t_\Sigma$, there are infinitely many event models that can be specified as $\mathcal{E}^t_\Sigma$, since there are infinitely many event models of the type t in which the preconditions of events are $Pre_\Sigma \vee \neg Pre_\Sigma$. By definition, isomorphic event models are distinct when they consist of distinct events. Therefore, clearly, there are infinitely many pair-wise *disjoint* sets defined to be $N_X(\Sigma)$ as defined above, depending on which event model will be taken as $\mathcal{E}^t_\Sigma$.

Let $A_1, A_2, \ldots$ be an infinite sequence of such sets, i.e. (1) A_i is of the form defined by $N_X(\Sigma)$ and (2) A_i, A_j are disjoint for every i, j. Define $N^n_X(\Sigma)$ be the union of $A_1, \ldots, A_n$. Clearly, $N^n_X(\Sigma)$ is finite, since A_i is finite for all i and $N^n_X(\Sigma)$ is a finite union of such sets.

Definition 4.6 Also, given a finite set Σ of expressions and a formula φ, define $\epsilon^{\top}_{\Sigma}$ to be an pointed event model in t_{PAL} in which the precondition of the event in the model is $PRE_{\Sigma} \vee PRE_{\Sigma}$. Given the form of the precondition in the definition, $\epsilon^{\top}_{\Sigma}$ does not occur in Σ. $\lhd$

Definition 4.7 The axiomatization of TDEL+P extends that of epistemic logic with necessitation for $[\epsilon]$ and $\hat{P}_{\epsilon}$ and the following axioms and inference rules:

F5 $\langle\epsilon\rangle P_{\epsilon'}\varphi \to \bot$ if $\epsilon \neq \epsilon'$

F6 $\langle\epsilon\rangle P_{\epsilon}\varphi \leftrightarrow \langle\epsilon\rangle\top \wedge \varphi$

A3 $P_{\epsilon}(\varphi \to \psi) \to (P_{\epsilon}\varphi \to P_{\epsilon}\psi)$

R(X) If $\vdash [\epsilon_1]\ldots[\epsilon_{d(\varphi)}]\varphi$ for all $\epsilon_1 \ldots \epsilon_{d(\varphi)}$ such that, for all k $(1 \leq k \leq d(\varphi))$, $\epsilon_k \in \mathbb{E}(\varphi) \cup N_X^{d(\varphi)}(\mathbb{E}(\varphi) \cup \{\epsilon^{\top}_{\mathbb{E}(\varphi)}\}$, then $\vdash \varphi$.

$\lhd$

Note that $\mathbb{E}(\varphi) \cup N_X^n(\mathbb{E}(\varphi)) \cup \{\epsilon^{\top}_{\mathbb{E}(\varphi)}\}$ is finite. Also, to show the soundness of **R(X)**, it suffices to show the following:

Lemma 4.8 *If φ is satisfiable, then $\langle\epsilon_1\rangle\ldots\langle\epsilon_{d(\varphi)}\rangle\varphi$ is satisfiable for some sequence $\epsilon_1\ldots\epsilon_{d(\varphi)}$ of the specified form in* R(X).

To show this lemma, we need some definitions. Let p be a state-dependent protocol on $\mathcal{M}$.

Definition 4.9 Given $n \in \mathbb{N}$, we define a local protocol $\mathsf{p}_{n<}$ on $\mathcal{M}^{n,\mathsf{p}}$ so that $p_{n<}(w\sigma_1\ldots\sigma_n) = \{\tau \mid w\sigma_1\ldots\sigma_n\tau \in \mathsf{p}(w)$ where $w \in Dom(\mathcal{M})\}$. $\lhd$

Given an ETL-model Forest$(\mathcal{M}, \mathsf{p})$, the model Forest$(\mathcal{M}^{n,\mathsf{p}}, \mathsf{p}_{n<})$ can be seen as a submodel of Forest$(\mathcal{M}, \mathsf{p})$ that describes what happens in Forest$(\mathcal{M}, \mathsf{p})$ after the $n+1$-th stage, with the histories up to the $n+1$-th stage taken as the elements of the base epistemic model.

Now we prove Lemma 4.8. The idea behind the proof is as follows. Assuming $\mathcal{H}, h \models \varphi$, we first apply the normalization method based on Proposition 4.1. Then, if φ is satisfied in the model at a sufficiently long history (i.e. strictly longer than $d(\varphi)$), then we can satisfy $\langle\epsilon_1\rangle\ldots\langle\epsilon_{d(\varphi)}\rangle\varphi$ by tracing the history using the truth definition of the future operator. If any ϵ_i in the sequence is not of the form specified in $R(X)$, then in the model $\mathcal{H}$ we can replace it with an event model of the same type with tautologous preconditions. Such a replacement does not affect the structure of the model, and $\langle\epsilon_1\rangle\ldots\langle\epsilon_{d(\varphi)}\rangle\varphi$ will be satisfied at the corresponding node in the resulted model.

However, if the history is not long enough, then we construct a new model from the original, by lifting the roots of the trees with a sequence of single reflexive event models $\epsilon^{\top}_{\mathbb{E}(\varphi)}$ with the tautologous precondition. The new model preserves the structures above the sequence of such events and there is a sufficiently long history at which φ is satisfied. The preservation result follows because iteratively performing single reflexive events with tautologous preconditions (uniformly at every world) keeps the structure of the original model unchanged.

To illustrate this, consider the evaluation of the formula $\varphi = P_{\sigma}\neg P_{\tau}\top$, with past depth 2, in Figure 4. Notice that we can satisfy this formula at world $w\sigma$ in Figure 4, even though $\mathsf{len}(w\sigma) = 2$. To obtain a length of 3 for the history at which the formula in question is satisfied, we add a public announcement with a tautologous precondition, $\epsilon^{\top}_{\varphi}$. This is represented in Figure 5. We now proceed to the proof.

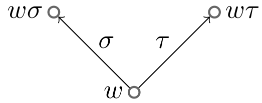


Figure 4. A formula with depth 2 can be satisfied at $w\sigma$. This is a case in which we need to extend the history.

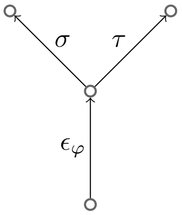


Figure 5. Extending the history with ϵ_{φ}.

Proof. Let $\mathcal{H}, h \models \varphi$. Apply Proposition 4.1 by setting $Y := \mathbb{E}(\varphi)$. Then we obtain $\mathcal{H}\sigma_Y, \sigma_Y(h) \models \varphi$.

Assume $\mathsf{len}(h) > d(\varphi)$. Then for some $g, \epsilon_1, \ldots \epsilon_{d(\varphi)}$, $h = g\epsilon_1\ldots\epsilon_{d(\varphi)}$. In $\mathcal{H}\sigma_Y$, for every $\sigma_Y(\epsilon_i) \notin Y$ $(1 \leq i \leq d(\varphi))$, replace $\sigma_Y(\epsilon_i)$ with an isomorphic event model $\epsilon \in N_X(\mathbb{E}(\varphi))$. Given that the preconditions of the event models are tautologous, such a model transformation does not affect the truth value of φ. That is, denoting by $\mathcal{H}'$ and h' the model and the history (corresponding to h) that are obtained by the replacements, we have $\mathcal{H}', h' \models \varphi$. By $\mathsf{len}(h) = \mathsf{len}(h') > d(\varphi)$ and the construction of h', we have some g' and $\epsilon'_1, \ldots, \epsilon'_n$ such that $\mathcal{H}', g' \models\rangle\epsilon'_1\rangle\ldots\langle\epsilon'_{d(\varphi)}\rangle\varphi$, where $\epsilon'_1, \ldots, \epsilon'_{d(\varphi)}$ are of the specified form in $R(X)$.

Thus, assume that $\mathsf{len}(h) \leq d(\varphi)$. Let $k := d(\varphi) - \mathsf{len}(h) + 1$ (the length that we want to add to the history). Let ϵ_0 be $\epsilon^{\top}_{\mathbb{E}(\varphi)}$. Also denote by ϵ_0^k the sequence of k ϵ_0's.

Now let $\mathcal{M} = (W, \sim, V)$. Construct a local protocol p^+ on $\mathcal{M}$ so that $\mathsf{p}^+(w)$ is the set obtained by taking the closure under finite prefix on $\{\epsilon_0^k\sigma \mid \sigma \in \mathsf{p}(w)\}$. Then, by these constructions, it is the case that for all σ (possibly empty):
$\mathsf{Forest}(\mathcal{M}^k, \mathsf{p}^+_{e_0^k<}), (w\epsilon_0^k)\sigma \models \varphi$ iff
$\mathsf{Forest}(\mathcal{M}, \mathsf{p}), w\sigma \models \varphi$
where w is in $\mathcal{M}$. Thus, if we have, for all σ,
$\mathsf{Forest}(\mathcal{M}^k, \mathsf{p}^+_{e_0^k<}), (w\epsilon_0^k)\sigma \models \varphi$ iff
$\mathsf{Forest}(\mathcal{M}, \mathsf{p}^+), w\epsilon_0^k\sigma \models \varphi$.
The desired claim follows. For we can proceed as in the case of $\mathsf{len}(h) > d(\varphi)$, given that
$\mathsf{len}(w\theta\tau) = \mathsf{len}(w) + [d(\varphi) - \mathsf{len}(w\tau) + 1] + \mathsf{len}(\tau) = d(\varphi) + 1$ where $h = w\tau$.

We prove this by showing that, for all σ and formulas ψ,
$\mathsf{Forest}(\mathcal{M}^k, \mathsf{p}^+_{e_0^k<}), w\epsilon_0^k\sigma \models \psi$ iff
$\mathsf{Forest}(\mathcal{M}, \mathsf{p}^+), w\epsilon_0^k\sigma \models \psi$.

The proof is by a straightforward induction. We will only do the past-modality case. The left-to-right direction follows immediately by the IH. So assume the RHS. If h is non-empty, then by the IH we are done. If h is empty, then since $\sigma \neq \epsilon_\varphi$ by definition, we have a contradiction with the RHS. This completes the proof. QED

The completeness proof can be given based on the Henkin-style construction given for TDEL above. Let $\mathcal{H}^{can}$ be the ETL-model constructed from the set of TDEL+P maximally consistent sets in the same way as in TDEL. The lemma for the canonical model that corresponds to Lemma 2.10 can be shown in the same way. Now, we show the truth lemma stated as follows:

Lemma 4.10 (Truth Lemma) *For every formula φ and $h \in \mathcal{H}^{can}$ such that $\mathsf{len}(h) > d(\varphi)$,*

$$\varphi \in \lambda(h) \textit{ iff } \mathcal{H}^{can}, h \models \varphi$$

Proof. The boolean and knowledge modality cases are given in the same way as Lemma 2.13 above, so we will only consider the past modality case. Let $h = h'\sigma$ for some $\mathsf{len}(h) \geq d(\varphi) + 1$, where $\sigma \in \mathbb{E}$. Let φ be $P_\tau\psi$.

Assume then that $P_\tau\psi \in \lambda(h)$. By the definition of canonical model, $\langle\sigma\rangle P_\tau\psi \in \lambda(h')$. If $\sigma \neq \tau$, then by F5, $\bot \in \lambda(h')$, which contradicts the consistency of $\lambda(h')$. Thus, assume $\sigma = \tau$. Then, by F6, we have $\psi \in \lambda(h')$. By the IH, $\mathcal{H}^{can}, h' \models \psi$ (note $\mathsf{len}(h') \geq d(\psi) + 1$). Since $h'\sigma \in \mathcal{H}^{can}$ and $\sigma = \tau$, the truth definition implies that $\mathcal{H}^{can}, h \models P_\tau\psi$.

For the other direction, assume that $\mathcal{H}^{can}, h \models P_\tau\psi$. By the truth definition, we have $\sigma = \tau$, and also $\mathcal{H}, h' \models \psi$. By the IH, we have $\psi \in \lambda(h')$. And by the construction of the canonical model, we have $\langle\sigma\rangle\top \in \lambda(h')$. Thus, by F6, we have $\langle\sigma\rangle P_\sigma\psi \in \lambda(h')$, which by construction implies that $P_\sigma\psi \in \lambda(h)$. QED

We can also prove the lemma corresponding to Lemma 2.14 in the same way. Now, to conclude our proof of the completeness result, we need to prove the following theorem.

Theorem 4.11 TDEL+P *is complete with respect to* $\mathbb{F}_{sd}(\mathbb{E})$.

Proof. Let φ be consistent. Then $\langle\sigma_1\rangle \dots \langle\sigma_{d(\varphi)}\rangle\varphi$ is consistent for some $\sigma_1 \dots \sigma_{d(\varphi)} \in \mathbb{E}^*$. For suppose otherwise. Then for every $\sigma_1 \dots \sigma_{d(\varphi)} \in \mathbb{E}^*$, $\langle\sigma_1\rangle \dots \langle\sigma_{d(\varphi)}\rangle$ is inconsistent and thus $\vdash [\sigma_1] \dots [\sigma_{d(\varphi)}]\neg\varphi$. By R, $\vdash \neg\varphi$. This contradicts the consistency of φ. Thus $\langle\sigma_1\rangle \dots \langle\sigma_{d(\varphi)}\rangle\varphi$ is consistent for some $\sigma_1 \dots \sigma_{d(\varphi)}$. Let $\theta = \langle\tau_1\rangle \dots \langle\tau_{d(\varphi)}\rangle\varphi$ be one of those formulas. Since θ is consistent, by Lindenbaum's Lemma, we have a maximally consistent set containing it. Note that $d(\theta) = 0$. Thus, by the truth lemma, there is some history h of length 1 such that $\mathcal{H}^{can}, h \models \theta$. This gives us the result that $\mathcal{H}^{can}, h\tau_1 \dots \tau_{d(\varphi)} \models \varphi$. QED

The reason that we cannot conclude the result immediately from the truth lemma and the analogue of Lemma 2.14 is that we are not sure that, given a formula of depth n, we have a maximal consistent set that contains φ, which is assigned to a history long enough to apply truth lemma. This fact is guaranteed by R, as is seen in the above argument.

5. Philosophical Connections and Applications

Although the addition of a past operator to the temporal framework may seem trivial, it turns out that the resulting increase in expressive power might have several significant applications. The interaction between past and future in an epistemic context can be found in thinking about agency—more specifically, in trying to formulate a definition of an agent's intention—as well as in learning.

Both of these seem at first glance to be forward looking ideas. For instance, intending seems to refer only to something we plan to do in the future. And learning seems to have to do with an update of our state of knowledge. But notice that if we intend to bring something about, it can't already have been the case (since we can't intend to do something that's already been done). And if we want to learn something, we can't already know it. Thus, expressing both of these ideas requires talking about a *change* in our epistemic states. It is not too difficult to come up with a sentence using only the future modality and the static language stating that I am about to learn that φ, or that I do not now know φ, but will after it is announced:

$$\langle !\varphi\rangle K_i\varphi \wedge \neg K_i\varphi$$

Alternately, we can use this formalism to capture our intuitions about what is learned by a public announcement of a formula φ. For what we learn is not necessarily that φ is now the case, but rather than φ was the case before the announcement. So our general formulation of what an agent learns by a public announcement can be expressed by the formula

$$[!\varphi]K_iP_{!\varphi}\varphi.$$

So in order to say that I have in fact learned φ, I need to refer back to the past. Otherwise, all I will be able to say is that I now know φ. But the fact that I now know φ tells me nothing about whether or not I knew it in the past. Thus, in order to claim that I have learned φ, because of some event ϵ I really need to say that I now know φ, but did not know it before ϵ took place:

$$P_\epsilon\neg K_i\varphi \wedge K_i\varphi.$$

The fact that a past modality is required to express that a state of affairs has changed means that it is also related to the idea of a successful update [11]. We can call a public announcement successful when the formula announced is true after the update, and unsuccessful when the formula announced becomes false. For instance, in the familiar Muddy Children example, the announcement by all the children that they do not know their state becomes false afterwards.

Another example of announcements which result in unsuccessful updates are Moore sentences, such as $p \wedge \neg K_ip$, or "p is the case, but i doesn't know it." For after that is announced, i will know that p is the case, and the original formula will become false. So as above, all we know is that $p\wedge\neg K_ip$ was true before it was announced. So even though the formula $K_i(p\wedge\neg K_ip)$ remains inconsistent in epistemic logic, the formula

$$K_iP_{!(p\wedge\neg K_ip)}(p \wedge \neg K_ip)$$

is satisfiable in dynamic epistemic logic, for instance, in a model like the one given in Figure 6.

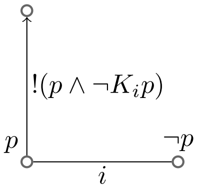


Figure 6. The public announcement of a Moore sentence. At the updated world, it is the case that $K_iP_{!(p\wedge\neg K_ip)}(p \wedge \neg K_ip)$.

So although an agent can never know that p is the case, but she herself does not know it, she can know that it *once was true* that p was the case and she then did not know it.

Now, we might think that the opposite of learning is forgetting, and wonder if this too is something that can be formalized by our models. After all, if we can express that an agent learned that φ after ϵ took place by saying $P_\epsilon\neg K_i\varphi \wedge K_i\varphi$, perhaps we could express that after ϵ, an agent forgot that φ by moving the negation:

$$P_\epsilon K_i\varphi \wedge \neg K_i\varphi.$$

But even though this sentence is expressible, the logic itself does not yet allow for a general way to model agents who can forget. For in the current models, such a sentence would only be satisfiable for a limited class of φ. For instance, it could never be true for a proposition letter. Since we have persistence for proposition letters across updates, once an agent knows that p, he can never forget it after an event. The reason for this is the fact that updates only ever erase uncertainties between worlds, or maintain existing uncertainties. In order to model forgetting, we would require an update mechanism that allowed for adding uncertainties between worlds which were not previously present. There are several different options for implementing such a mechanism, which are beyond the scope of this paper to discuss. However, this avenue seems like another fruitful path to investigate in terms of dynamic epistemic systems with temporal operators.

6. Conclusion

We have shown that, even if we extend the setting of TPAL presented in [8] to the full class of event models, the completeness proof can be given based on the proof given for TPAL in [8]. Also the extension TDEL+P can be axiomatized by the method of normalization for DEL-generated ETL-models.

But these are not the only extensions which suggest themselves for investigation. For instance, in TDEL+P, we only have *labeled* past and future operators. So natural further steps would be to add in an un-indexed past operator, expressing "yesterday", and an un-indexed future operator, expressing "tomorrow". We can look at these operators as quantifying over event models. It turns out that a system TADEL with the "tomorrow" operator can be axiomatized without too many problems, as it can be seen as a generalization of the system TAPAL studied in [4], which has an operator quantifying over public announcements. These results will be presented in forthcoming work by Hoshi, which will demonstrate the way in which the normalization method can be applied to axiomatize TADEL.

Perhaps surprisingly, though, the addition of a "yesterday" operator is not as straightforward, since the normalization method would not work as given. In particular, the method whereby we extend the history with ϵ_φ as illustrated in Figure 5 would not necessarily work for formulas in a language with a "yesterday" operator. For where we can satisfy $P_\sigma\neg P_\tau\top$ in a world with length 2, the formula $P\neg P\top$

can only be satisfied in a world with length 1. So the history could not be lengthened in a world in which the latter was satisfied without changing its truth value.

Other natural extensions include iterated past modality P^*, where the * is the Kleene star operator. In the case of the iterated future modality of the kind, say $\Diamond^*$, "There is some sequence of events after which...", the result in [5] suggests that such an operator results in incompleteness when combined with the common knowledge operator. It is interesting to see if this is also the case for the case of the iterated past-modality P^*.

There are distinct motivations also for considering an extension of TDEL+P together with a common knowledge operator. The considerations raised about learning in the previous section apply just as well to agents' common knowledge after an announcement, since we can also express what becomes common knowledge by the following formula:

$$[!\varphi]C_G P_{!\varphi}\varphi.$$

Further, the relativized common knowledge operator from [9] $C_G(\varphi, \psi)$, which expresses that every G-path which consists exclusively of φ worlds ends in a ψ world, also has a very natural interpretation in past language. One way to paraphrase this operator in natural language is "If φ were announced, it would be common knowledge among G that ψ was the case before the announcement." This is expressible in the past language.

$$C_G(\varphi, \psi) \equiv [!\varphi]C_G P_{!\varphi}\psi$$

Thus, there are many potentially fruitful extensions of the system considered here, which will certainly be the subject of future investigation.

References

[1] A. Baltag, L. Moss, and S. Solecki. The logic of public announcements, common knowledge and private suspicions. In I. Gilboa, editor, *TARK 1998*, number 43-56, 1998.

[2] R. Fagin, J. Halpern, Y. Moses, and M. Vardi. *Reasoning about Knowledge*. Synthese Library. MIT Press, Boston, 1995.

[3] J. Gerbrandy. *Bisimulations on Planet Kripke*. PhD thesis, ILLC, 1999.

[4] T. Hoshi. Logics of public announcement with constrained protocols. LOFT, 2008.

[5] J. Miller and L. Moss. The undecidability of iterated modal relativization. *Studia Logica*, 79:373–407, 2005.

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

[7] Y. Shoham and K. Leyton-Brown. Multiagent systems: Algorithmic, game-theoretic, and logical foundations. 2008.

[8] J. van Benthem, J. Gerbrandy, T. Hoshi, and E. Pacuit. Merging frameworks for interaction: DEL and ETL. 2007.

[9] J. van Benthem, J. van Eijck, and B. Kooi. Common knowledge in update logics. In *Proceedings of the 10th Conference on Theoretical Aspects of Rationality and Knowledge*. 2005.

[10] J. van Benthem, J. van Eijck, and B. J. Kooi. Logic of communication and change. *Information and Computation*, 204(11):1620–1662, 2006.

[11] H. van Ditmarsch and B. Kooi. The secret of my success. *Synthese*, 151(2):201–232, 2006.

[12] A. Yap. Dynamic epistemic logic and temporal modality. University of Victoria, 2007.

Introspective forgetting

Hans van Ditmarsch Andreas Herzig Jérôme Lang Pierre Marquis

1. Introduction

[1]There are different ways of forgetting.

Completely forgetting In the movie 'Men in Black', Will Smith makes you forget knowledge of extraterrestials by flashing you with a light in the face. After that, you have forgotten the green ooze flowing out of mock-humans and such: you not remember that you previously had these experiences. In other words, even though for some specific forgotten fact p it is now the case that $\neg Kp$ and $\neg K\neg p$, the flash victims have no memory that they previously knew the value of p. Worse, they forgot that p is an atomic proposition at all. This sort of forgetting is dual to awareness—in a logical setting it is uncommon that parameters of the language, such as the set of atoms, shrink, although there are ways to simulate that. We will leave this matter aside for now.

Remembering prior knowledge A different sort of forgetting is when you forgot which of two keys fits your office door, because you have been away from town for a while. In this case you remember that you knew which key it was, and you currently don't know which key it is. This is about forgetting the value of a atomic proposition p. Previously, either Kp or $K\neg p$, but currently $\neg Kp$ and $\neg K\neg p$. This sort of forgetting will be very central to our concerns.

Forgetting values Did it ever happen to you that you met a person whose face you recognize but whose name you no longer remember? Surely! Or that you no longer know the pincode of your bankcard? Hopefully not. But such a thing is very conceivable. This sort of forgetting means that you forgot the value of a proposition, or the assignment of two values from different sets of objects to each other. In the case of a bankcard you the four-number code that you forgot is just one one 10,000 options, so previous it was true that $K0000 \vee \ldots \vee K9999$ whereas currently we have that $\neg(K0000 \vee \ldots \vee K9999)$. (Let 0000 stand for the proposition that your pin number is 0000, etc.) Similarly, somewhat simplifying matters, the finite number of all humans only have a finite, somewhat smaller, number of names. An atomic proposition about your office keys is also a feature namely with two values only, true and false. The multiple-valued features can also be modelled as a *number* of atomic propositions; this can be done in a very uneconomic fashion as above, but also in a minimal way. We conclude that this sort of forgetting is like the previous kind.

Defaulting on obligations But there are other kinds of forgetting too. For example, say I forgot to pick you up at the airport at 4:30 PM. Forgetting an action is very different from forgetting a proposition. Forgetting an action amounts to defaulting on an obligation and the *observation* of having forgotten it is not at all related to ignorance. It points backwards in time to the moment when you were not *aware* of the obligation. Obligations can be modelled with deontic logics. We will not be concerned with this kind of forgetting.

Multi-agent versions of forgetting In a multi-agent setting additional, interactive, ways of forgetting crop up as well. Some of the above have group versions. For example, Will Smith only had to flash a whole group once, not each of its members individually. And if you have been flashed, although *you* don't know that you knew about the green ooze, Will Smith knows that you knew. So in a multi-agent setting some aspects of 'completely forgetting' can be modelled. When assuming standard notions of knowledge, that is introspective, we now run straight into trouble of another kind.

A group version for 'remembering prior knowledge' is hard to justify, because its interpretation typically involves introspection: you forgot something if *you are aware of* (in the sense of 'you know') previous knowledge and present ignorance of it. A setting wherein a group is collectively aware of its prior (common) knowledge is somewhat harder to imagine. It makes more sense to have a version of 'remembering prior knowledge' for individuals in a group, be-

[1] Emails of the authors: `hans@cs.otago.ac.nz`, `herzig@irit.fr`, `lang@irit.fr`, `marquis@cril.univ-artois.fr`. Hans van Ditmarsch is corresponding author. This document concerns work in progress. Hans van Ditmarsch acknowledges the support of the Netherlands Institute of Advanced Study where he was Lorentz Fellow in 2008.

cause they can inform and are observed by others: here you standing in front of your office door again now in company of four freshmen students, “Ohmigod, I forgot again which is my office key!”

For yet other multi-agent examples: I can notice that you forgot to pick me up at the airport, or that you no longer appear to know the way around town. The last may even be without me being aware of my ignorance. I may have forgotten whether you knew about a specific review result for our jointly editored journal issue. In other words, previously $K_{me}K_{you}\textit{accept}$ or $K_{me}K_{you}\neg\textit{accept}$ but currently $\neg K_{me}K_{you}\textit{accept}$ and $\neg K_{me}K_{you}\neg\textit{accept}$. Some meaningful propositions that can be forgotten in a multi-modal context are therefore modal.

1.1. Forgetting and progression

In theory change (belief revision) the operation of forgetting is a form of *belief contraction*. Given prior belief in p or its negation, we want to remove that from the set of believed formulas including all its dependencies. In artificial intelligence this has become a search for efficient ways to implement such a contraction. The following way to model / implement forgetting an atomic proposition has recently been proposed [5]. Given a set of formulas (‘theory’) Φ, we compute the effect of forgetting information about atom p by a binary operation

$$\textit{Fg}(\Phi, p) := \{\varphi(\top/p) \vee \varphi(\bot/p) \mid \varphi \in \Phi\}$$

Here, $\varphi(\psi/p)$ is the replacement of all (possibly zero) occurrences of p in φ by ψ. This proposal can be called the (syntactic) *progression* of Φ by the function *Fg*, relative to the forgotten information about p. It is well-known that this way to model forgetting as progression does not work for *modal* formulas. For example, if the agent already does not know whether p, surely that should remain the case after forgetting the value of p. But we now have that (write $\textit{Fg}(\varphi, p)$ for $\textit{Fg}(\{\varphi\}, p)$):

$$\begin{array}{l} \textit{Fg}(\neg Kp \wedge \neg K\neg p, p) \\ = \\ (\neg K\top \wedge \neg K\neg\top) \vee (\neg K\bot \wedge \neg K\neg\bot) \\ \Leftrightarrow \\ (\neg\top \wedge \neg\bot) \vee (\neg\bot \wedge \neg\top) \\ \Leftrightarrow \\ \bot \end{array}$$

For another example of an undesirable feature, it is also not the case that knowledge of p is transformed into ignorance about p by this procedure:

$$\textit{Fg}(Kp, p) \leftrightarrow (K\top \vee K\bot) \leftrightarrow \top.$$

In other words, this approach does not lead anywhere for modal formulas. Surely, one would like that $\neg Kp \wedge \neg K\neg p$ is true after forgetting the value of p, even when this was not true initially. For any theory Φ and atom p, the result of forgetting p should entail ignorance about p:

$$\textit{Fg}(\Phi, p) \models \neg Kp \wedge \neg K\neg p.$$

The difficulties in obtaining this result by theory revision motivated us to model forgetting as an event in a dynamic epistemic logic.

1.2. Forgetting or no-forgetting, that’s the question

We model the action of forgetting an atomic proposition p as an *event* $\textit{Fg}(p)$ (in its sense of remembering prior knowledge about p). We do this in a propositional logic expanded with an epistemic modal operator K and a dynamic modal operator $[\textit{Fg}(p)]$ (including multi-agent versions). As usual, $K\varphi$ describes that the agent knows φ. Formula $[\textit{Fg}(p)]\varphi$ means that after the agent forgets his knowledge about p, φ is true.

The obvious *precondition* for event $\textit{Fg}(p)$ is prior knowledge of the value of p: $Kp \vee K\neg p$. The obvious *postcondition* for event $\textit{Fg}(p)$ is ignorance of the value of p: $\neg Kp \wedge \neg K\neg p$. In other words

$$(Kp \vee K\neg p) \rightarrow \langle \textit{Fg}(p)\rangle(\neg Kp \wedge \neg K\neg p)$$

should be valid in the information state prior to the event of forgetting ($\langle \textit{Fg}(p)\rangle$ is the diamond version of $[\textit{Fg}(p)]$). Or, abstracting from that precondition, it should be valid that:

$$[\textit{Fg}(p)](\neg Kp \wedge \neg K\neg p).$$

Wasn’t dynamic epistemic logic supposed to satisfy the principle of ‘no forgetting’? So how on earth can one model forgetting in this setting? We can, because we cheat. ‘No forgetting’ (a.k.a. ‘perfect recall’) means that if states s and t resulting from the execution of (possibly different) events are indistinguishable, then the states before the execution of these respective events are also indistinguishable. If after the event of forgetting I am ignorant about p I cannot distinguish a p-state from a $\neg p$-state. Therefore, because of the principle of ‘no forgetting’, I should already have been unable to distinguish these states before the execution of this supposed event $\textit{Fg}(p)$... I should have been already ignorant about p before... We solve this dilemma by the standard everyday solution of forgetful people: blame others for your forgetfulness. In this case, we blame the world, i.e., the state of the world: we *simulate* forgetting by *nondeterministically changing the value of* p *in the actual or other states, in a way unobservable by the agent.* Thus resulting in his ignorance about p. Note that this solution is different from how belief revision is modelled in dynamic epistemic logic: prior belief in p that is revised with $\neg p$ and results in belief in $\neg p$ is standardly modelled by considering this a 'soft' or defeasible form of belief (i.e., not knowledge) and implemented by changing a preference relation between states [11, 3].

Once we have the above, the relation with theory progression becomes clear. Let (M, s) be an information state (pointed Kripke model with designated state) for the theory Φ. Suppose we want to know if $Fg(\Phi, p) \models \psi$. The artificial intelligence community is particularly interested in efficient ways to perform such computations. Well, if it is the case, we then should also have $M, s \models [Fg(p)]\psi$. We propose a $[Fg(p)]$-operator that can be reduced (eliminated): there is an epistemic formula χ such that $[Fg(p)]\psi \leftrightarrow \chi$. There are fast and efficient algorithms to determine the truth of an epistemic formula in a given Kripke model: $M, s \models \chi$ can be done quickly. This answers the question whether $Fg(\Phi, p) \models \psi$. There is one drawback: the reduction of $[Fg(p)]\psi$ to a formula without reference to an event that should be initially true is known as regression. But in AI we also want to do progression: compute some Φ' from $Fg(\Phi, p)$. This is harder, also in a dynamic epistemic logical context, and we have no answer to that, although a few suggestions.

Expanding our perspective This contribution focusses on a clean solution on how to model an event $Fg(p)$ satisfying $(Kp \vee K\neg p) \rightarrow \langle Fg(p)\rangle(\neg Kp \wedge \neg K\neg p)$. From there on, the modelling desiderata diverge. There are many interesting options.

Is our perspective that of a modelling observer, in which case we might require that forgetting is an information-changing event only, so that the value of p in the actual state should not change? Or is our perspective that of an agent in the system, so that we are only considering the value of local propositions, i.e., formulas of the form $K\varphi$? Whether we simulate the desiderata by factual change inducing informational change does not matter in that case.

Are we talking about one or about all agents forgetting? How about forgetting epistemic propositions?

Our solution presumes the interpretation of $K\varphi$ as 'the agent *knows* φ' and, correspondingly, even though our semantics is general all our examples are of $S5$-structures. There are obvious slightly weaker modellings of forgetting to model introspective belief (to be interpreted on $KD45$-structures.

From the perspective of the agent we also want to look backwards. Let $Fg(p)^-$ be the converse of $Fg(p)$ (e.g. in the sense of [1, 15, 8]). All the time we are saying that an agent that has forgotten about p remembers prior knowledge of p. This we can now express as the validity of

$$K(\neg Kp \wedge \neg K\neg p \wedge \langle Fg(p)^-\rangle(Kp \vee K\neg p))$$

in the information state after the event of forgetting. (Note the different perspective from before, where our perspective was the information state before the act of forgetting.) In other words: the agent is aware of its current ignorance and its previous knowledge. We will indicate some ways to address this in the further research section at the end of our contribution. There is one main drawback of this approach: there is no way to reduce expressions with converse events to purely epistemic formulas. So, the advantage of dynamic epistemic logic for regression questions in AI has not been reached there (yet).

2. Language and semantics

Language Our language is

$$p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_a\varphi \mid Fg_B(p)$$

In the single-agent context write $K\varphi$ for $K_a\varphi$ and $Fg(p)$ for $Fg_a(p)$.

Later on, in a multi-agent context, we write $Fg(p)$ for $Fg_A(p)$, and we also distinguish the converse ('remember') operator $Fg_B(p)^-$. For the forgetting of (not necessarily atomic) formulas φ we write $Fg_B(\varphi)$.

Structures Our structures are pointed Kripke models $(S, R, V), s)$ (with $R : A \rightarrow \mathfrak{P}(S \times S)$ and $V : P \rightarrow \mathfrak{P}(S)$) and multiple-pointed event models ('action models'). Our typical example structures are $S5$ to model knowledge and knowledge change and for $KD45$ to model belief and belief change.

The dynamic structures are event models, i.e., action models including assignments of atoms (a.k.a. substitutions) [10, 13]. We follow notational conventions as in [13]: if in event s the precondition is φ and the postcondition is that the valuation of atom p becomes that of ψ, we write: in s: if φ then $p := \psi$.

We visualize $S5$ models by linking states that are indistinguishable for an agent, possibly labelling the link with the agent name (not in the single-agent situation). Transitivity is assumed. In pictures of event models: a formula next to an event is its precondition, an assignment next to it a postcondition.

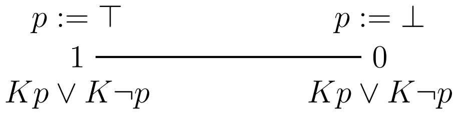
Semantics Assume an epistemic model $M = (S, R, V)$.

$$\begin{array}{lll}
M, s \models p & \text{iff} & s \in V(p) \\
M, s \models \neg\varphi & \text{iff} & M, s \not\models \varphi \\
M, s \models \varphi \wedge \psi & \text{iff} & M, s \models \varphi \text{ and } M, s \models \psi \\
M, s \models K_a\varphi & \text{iff} & \text{for all } t \in S : (s, t) \in R_a \\
 & & \text{implies } M, t \models \varphi \\
M, s \models [Fg_B(p)]\psi & \text{iff} & \text{for all } M', s' : (M, s)[\![Fg_B(p)]\!](M', s') \\
 & & \text{implies } M', s' \models \varphi
\end{array}$$

where $[\![Fg_B(p)]\!]$ is a binary relation between pointed epistemic states, as usual for the interpretation of events in dynamic epistemic logic. Of course, we model the execution of an event $Fg_B(p)$ as a restricted modal product and this will be the relation induced by that. In the next section we will define the event model $Fg_B(p)$. The set of validities in our logic is called FG.

3. Forgetting

In a single-agent setting we model forgetting as the non-deterministic event where the (anonymous) agent is uncertain which of two assignments have taken place: p becomes true, or p becomes false. Formally, this is a non-deterministic event model consisting of two events 0 and 1 that are indistinguishable for the agent, and such that $\mathsf{pre}(0) = \mathsf{pre}(1) = Kp \vee K\neg p$, $\mathsf{post}(0)(p) = \bot$, and $\mathsf{post}(1)(p) = \top$. We can visualize this event model $Fg(p)$ as follows (postconditions above, preconditions below actions):



The event model $Fg(p)$ is non-deterministic choice between two deterministic events $(Fg(p), 1)$ and $(Fg(p), 0)$. For the interpretation of such a pointed event we use the standard semantics of 'action models', for events/points $i = 0, 1$ (and we recall the equivalence $[Fg(p)]\psi \leftrightarrow ([Fg(p), 0]\psi \wedge [Fg(p), 1]\psi)$:

$$\begin{array}{lll} M, s \models [Fg(p), i]\varphi & \text{iff} & M, s \models Kp \vee K\neg p \text{ implies} \\ & & M \otimes Fg(p), (s, i) \models \varphi \end{array}$$

In the language we'd like to avoid directly referring to the pointed versions (out of some possibly mistaken sense of minimalism), and therefore introduce the pointed versions of forgetting by abbreviation (and this amounts indeed to the same):

$$\begin{array}{lll} \langle Fg(p), 0\rangle\varphi & = & \langle Fg(p)\rangle(\neg p \wedge \varphi) \\ \langle Fg(p), 1\rangle\varphi & = & \langle Fg(p)\rangle(p \wedge \varphi) \end{array}$$

To obtain a complete axiomatization for FG we can simply apply the reduction axioms for event models, as specified in [13]. This is the axiomatization **FG** in Table 1. Note that from the above abbreviation also follows that $[Fg(p)](p \to \varphi)$ is equivalent to $[Fg(p), 1]\varphi$, and that $[Fg(p)](\neg p \to \varphi)$ equals $[Fg(p), 0]\varphi$.

Proposition 3.1 *Axiomatization* **FG** *is sound and complete.*

$$\begin{array}{lll} [Fg(p)]p & \leftrightarrow & \neg(Kp \vee K\neg p) \\ [Fg(p)]q & \leftrightarrow & (Kp \vee K\neg p) \to q \;\text{ for } q \neq p \\ [Fg(p)]\neg\varphi & \leftrightarrow & (Kp \vee K\neg p) \to \\ & & (\neg[Fg(p)](\neg p \to \varphi) \wedge \neg[Fg(p)](p \to \varphi)) \\ [Fg(p)](\varphi \wedge \psi) & \leftrightarrow & [Fg(p)]\varphi \wedge [Fg(p)]\psi \\ [Fg(p)]K\varphi & \leftrightarrow & (Kp \vee K\neg p) \to K[Fg(p)]\varphi \end{array}$$

Table 1. Axiomatization FG—only reduction rules involving Fg are presented

Proof. We show that the axiomatization resulted from application of the reduction axioms in action model logic by Baltag et al. [2], by applying, case by case, the standard reduction rules for event models. This kills two birds (soundness and completeness) at one throw.

Case p.
$[Fg(p)]p$
$\Leftrightarrow$
$[Fg(p), 0]p \wedge [Fg(p), 1]p$
$\Leftrightarrow$
$(\mathsf{pre}(0) \to \mathsf{post}(0)(p)) \wedge (\mathsf{pre}(1) \to \mathsf{post}(1)(p))$
$\Leftrightarrow$
$((Kp \vee K\neg p) \to \bot) \wedge ((Kp \vee K\neg p) \to \top)$
$\Leftrightarrow$
$(Kp \vee K\neg p) \to \bot$
$\Leftrightarrow$
$\neg(Kp \vee K\neg p)$

In other words, it is not the case that p is true after every execution of $Fg(p)$.

Case q.
$[Fg(p)]q$
$\Leftrightarrow$
$[Fg(p), 0]q \wedge [Fg(p), 1]q$
$\Leftrightarrow$
$(\mathsf{pre}(0) \to \mathsf{post}(0)(q)) \wedge (\mathsf{pre}(1) \to \mathsf{post}(1)(q))$
$\Leftrightarrow$ $\mathsf{pre}(0) = \mathsf{pre}(1) = Kp \vee K\neg p, \mathsf{post}(0)(q) = \mathsf{post}(q)(1) = q$
$(Kp \vee K\neg p) \to q$

This axiom expresses that, if $Fg(p)$ is executable, the value of atoms q other than p remains the same.

Case $\neg\varphi$.
$[Fg(p)]\neg\varphi$
$\Leftrightarrow$
$[Fg(p), 0]\neg\varphi \wedge [Fg(p), 1]\neg\varphi$
$\Leftrightarrow$

$(\mathsf{pre}(0) \to \neg[\mathit{Fg}(p),0]\varphi) \wedge (\mathsf{pre}(1) \to \neg[\mathit{Fg}(p),1]\varphi)$
$\Leftrightarrow$ $\qquad$ $\mathsf{pre}(0) = \mathsf{pre}(1) = Kp \vee K\neg p$
$(Kp \vee K\neg p) \to (\neg[\mathit{Fg}(p),0]\varphi \wedge \neg[\mathit{Fg}(p),1]\varphi)$
$\Leftrightarrow$
$(Kp \vee K\neg p) \to (\neg[\mathit{Fg}(p)](\neg p \to \varphi) \wedge \neg[\mathit{Fg}(p)](p \to \varphi))$

Note that the expression on the right is *not* equivalent to $\neg[\mathit{Fg}(p)]\varphi$. It would be if the conjunction in the middle had been a disjunction.

Case $\varphi \wedge \psi$.
$[\mathit{Fg}(p)](\varphi \wedge \psi)$
$\Leftrightarrow$
$[\mathit{Fg}(p),0](\varphi \wedge \psi) \wedge [\mathit{Fg}(p),1](\varphi \wedge \psi)$
$\Leftrightarrow$
$[\mathit{Fg}(p),0]\varphi \wedge [\mathit{Fg}(p),0]\psi \wedge [\mathit{Fg}(p),1]\varphi \wedge [\mathit{Fg}(p),1]\psi$
$\Leftrightarrow$
$[\mathit{Fg}(p)]\varphi \wedge [\mathit{Fg}(p)]\psi$

Case $K\varphi$.
$[\mathit{Fg}(p)]K\varphi$
$\Leftrightarrow$
$[\mathit{Fg}(p),0]K\varphi \wedge [\mathit{Fg}(p),1]K\varphi$
$\Leftrightarrow$
$(\mathsf{pre}(0) \to K[\mathit{Fg}(p)]\varphi) \wedge (\mathsf{pre}(1) \to K[\mathit{Fg}(p)]\varphi)$
$\Leftrightarrow$
$(Kp \vee K\neg p) \to K[\mathit{Fg}(p)]\varphi$
QED

Proposition 3.2 *The formula* $[\mathit{Fg}(p)](\neg Kp \wedge \neg K\neg p)$ *is valid and derivable.*

Proof. Validity is trivial. Thus we have derivability. It is instructive to see part of the derivation. We apply the reduction rules in the axiomatization **FG**.

$[\mathit{Fg}(p)](\neg Kp \wedge \neg K\neg p)$
$\Leftrightarrow$
$[\mathit{Fg}(p)]\neg Kp \wedge [\mathit{Fg}(p)]\neg K\neg p$

Left conjunct of previous line:
$[\mathit{Fg}(p)]\neg Kp$
$\Leftrightarrow$
$(Kp \vee K\neg p) \to (\neg[\mathit{Fg}(p),0]Kp \wedge \neg[\mathit{Fg}(p),1]Kp)$
$\Leftrightarrow$
$((Kp \vee K\neg p) \to \neg[\mathit{Fg}(p),0]Kp) \wedge ((Kp \vee K\neg p) \to \neg[\mathit{Fg}(p),1]Kp)$

Again, left conjunct of previous line:
$(Kp \vee K\neg p) \to \neg[\mathit{Fg}(p),0]Kp$
$\Leftrightarrow$
$(Kp \vee K\neg p) \to \neg((Kp \vee K\neg p) \to K[\mathit{Fg}(p)]p$
$\Leftrightarrow$
$(Kp \vee K\neg p) \to \neg((Kp \vee K\neg p) \to K\neg(Kp \vee K\neg p))$
$\Leftrightarrow$
$(Kp \vee K\neg p) \to \neg\bot$
$\Leftrightarrow$
$\top$

All together we have four cases (conjuncts), of which have now done one. The four cases are similar. QED

Proposition 3.3 $[\mathit{Fg}(p)][\mathit{Fg}(p)]\varphi$ *is valid.*

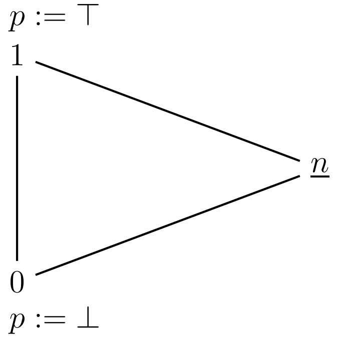
Proof. Assume the first $\mathit{Fg}(p)$ can be executed. Then the precondition $Kp \vee K\neg p$ was satisfied in the initial model. After the execution of that $\mathit{Fg}(p)$, we have $\neg(Kp \vee K\neg p)$. Therefore the second $\mathit{Fg}(p)$ cannot be executed. (So, trivially, any postcondition φ of that is then true.) QED

Unlike in real life, you cannot forget something *twice*. After you have forgotten it the first time, you have to be informed again about p and only then you can forget it again. Maybe that's quite a bit like real life after all.

Progress towards seeing this modelling of forgetting as *progression* in the AI sense would be made if we were to prove that $\psi \to [\mathit{Fg}(p)]\psi$ is valid for all ψ that do not contain occurrences of p. We think this is valid, and it may even be trivial, but we haven't given it sufficient attention yet.

4. Forgetting without changing the real world

An unfortunate side effect of our modelling of forgetting is that the actual value of p gets lost in the process. This is 'somewhat strange' if we only want to model that the agents forget the value of p, but that 'otherwise' nothing changes: the real value of p should then be unchanged. We can overcome that deficiency in the alternative modelling (Fg, n). It is very much like $\mathit{Fg}(p)$ except that there is one more alternative event in the model, indistinguishable from the other two, that represents the event 'nothing happens' except that the truth of p should be known (its precondition is $Kp \vee K\neg p$ and there are no postconditions). Also, unlike $\mathit{Fg}(p)$, the alternative (Fg, n) is pointed: this event model is deterministic, the real event is event n. This ensures that the real value of p does not change. In the figure we have not indicated the preconditions $Kp \vee K\neg p$.



The reduction rules for (Fg, n) are the same as for *Fg* except for the atomic case p and for negation, where:

$$\begin{array}{lcl} [\mathsf{Fg}(p), n]p & \leftrightarrow & (Kp \vee K\neg p) \to p \\ [\mathsf{Fg}(p), n]\neg\varphi & \leftrightarrow & (Kp \vee K\neg p) \to (\neg[\mathsf{Fg}(p), 0]\varphi \wedge \\ & & \neg[\mathsf{Fg}(p), 1]\varphi \wedge \neg[\mathsf{Fg}(p), n]\varphi) \end{array}$$

We can introduce $(\mathsf{Fg}(p), 0)$ and $(\mathsf{Fg}(p), 1)$ by abbreviation, somewhat different from before. We now have the interesting results that

Proposition 4.1 *Valid are (proof omitted):*

$$\begin{array}{ll} \psi \to [\mathsf{Fg}, n]\psi & \textit{for boolean } \psi \\ [\mathsf{Fg}, n]K\psi \leftrightarrow [Fg]K\psi & \textit{for any } \psi \end{array}$$

In other words: from the perspective of the agent, the different modellings of forgetting are indistinguishable. That makes the simpler modelling *Fg*(p) preferable over the slightly more complex $(\mathsf{Fg}(p), n)$.

5. Further research and variations

In this section we present some less developed lines of research.

5.1. Forgetting by bisimulation quantification

A not strictly modal way to model forgetting is to see forgetting p as universal bisimulation quantification over p (as in [Hollenberg] [Visser]), i.e.:

$$[Fg^{\forall}(p)]\varphi := \forall p \varphi$$

where

$M, s \models \forall p \varphi$
$\Leftrightarrow$
for all (M', s') such that $(M', s') \underline{\leftrightarrow}_{P-p} (M, s) : (M', s') \models \varphi$

The notation $(M', s') \underline{\leftrightarrow}_{P-p} (M, s)$ means that epistemic state (M', s') is bisimilar to epistemic state (M, s) with respect to the set P of all atoms *except* p. In other words the valuation of p may vary 'at random'. This includes the models constructed by *Fg*(p) and by $(\mathsf{Fg}(p), n)$ from that given M. That is, for any epistemic model M we have that

$$\begin{array}{lcl} M & \underline{\leftrightarrow}_{P-p} & M \otimes Fg(p) \\ M & \underline{\leftrightarrow}_{P-p} & M \otimes \mathsf{Fg}(p) \end{array}$$

from which follow the validities

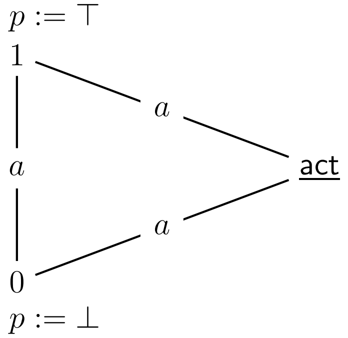
$$\begin{array}{l} [Fg^{\forall}(p)]\psi \to [Fg(p)]\psi \\ [Fg^{\forall}(p)]\psi \to [\mathsf{Fg}(p), n]\psi \end{array}$$

The axiomatizations of such bisimulation quantified logics are often complex; for $S5$ models they behave somewhat better [4]. We treat a bisimulation quantification operation non-standardly as 'some sort of dynamic modal operator' here. We justify this because it is a model changing operation. This perspective is also explored in [12].

Although theoretically an interesting alternative, the much simpler *Fg*(p) and $(\mathsf{Fg}(p), n)$ seem to be preferable for computational results. However, the bisimulation version may have other advantages we are currently unaware of.

5.2. Single agent forgetting in a multi-agent context

Suppose a single agent says 'I forgot p' in the presence of others. This can be modelled by the multi-agent event model $Fg_a(p)$, where, again, all preconditions are $K_a p \vee K_a \neg p$.



The visualization means that all agents *except* a can distinguish between the three alternatives. (So for all agents in $A - a$ the accessibility relation is the identity on the domain.) In this case, a similar two-event model (as in the single-agent approach) would not suffice (said to be irrelevant from agent a's point of view), as we also have to take the other agents into consideration: surely, we don't want them do doubt the value of p all of a sudden.

Again, there is an obvious complete axiomatization applying the reduction rules for event models, and we have the validity

$$[Fg_a(p)]C_A(\neg K_a p \wedge \neg K_a \neg p)$$

where C_A stands for 'common knowledge among group A.'

5.3. Remembering that you have forgotten

To remember in the object language that you have forgotten something requires a language allowing

$$K(\neg Kp \wedge \neg K\neg p \wedge \langle Fg(p)^{-}\rangle(Kp \vee K\neg p))$$

By instead of pointed Kripke models taking what is known as the 'forest' produced by the initial model and all possible sequences of all *Fg*(p) events (for all atoms) (see [9] and various other publications, this relates strongly to the history-based approaches by Parikh & Ramanujam [7], and later Pacuit [6], and others [14]), we get a model that allows us to refer to past events (à la [15] and [8] — Sack's approach is also properly based on histories). We can now combine this recent strand of research, with another strand of adding assignments to the language, as we already did, and additionally to that we can add theories for event models using converse actions, as done in [1] and also outlined in, e.g. [9]. This not so grand but nevertheless not yet realized scheme leads somewhere, namely to a complete axiomatization, but very likely not to the desirable result that expressions containing event operators (converse or not) can be reduced to epistemic formulas. So from an AI point of view, this is probably not a productive point of view. From a cognitive modelling point of view, it is of course interesting as we can refer to previous knowledge. (Also note that, unlike the typical counterexamples in [15], in this case the agent *knows* that prior to the current situation he/she had knowledge of p—absence of that created the problems with finding reduction axioms. So within this restricted setting of specific events, maybe more useful can be done... To be continued.

5.4. Other matters

Modeling the forgetting of features with multiple values can be done by a simple adjustment of the above proposals. This is easy. How to model the forgetting modal formulas is a different piece of cake altogether; in this case we have made no progress yet.

References

[1] G. Aucher. *Perspectives on belief and change*. PhD thesis, University of Otago & Institut de Recherche en Informatique de Toulouse, New Zealand & France, 2008.

[2] A. Baltag, L.S. 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, 1998.

[3] A. Baltag and S. Smets. Dynamic belief revision over multi-agent plausibility models. Proceedings of LOFT 2006 (7th Conference on Logic and the Foundations of Game and Decision Theory), 2006.

[4] T. French. *Bisimulation quantifiers for modal logic*. PhD thesis, University of Western Australia, 2006.

[5] J. Lang, P. Liberatore, and P. Marquis. Propositional independence: Formula-variable independence and forgetting. *J. Artif. Intell. Res. (JAIR)*, 18:391–443, 2003.

[6] E. Pacuit. Some comments on history-based structures. To appear in Journal of Applied Logic, 2007.

[7] R. Parikh and R. Ramanujam. Distributed processing and the logic of knowledge. In *Logic of Programs*, volume 193 of *Lecture Notes in Computer Science*, pages 256–268. Springer, 1985. A newer version appeared in *Journal of Logic, Language and Information*, vol. 12, 2003, pp. 453–467.

[8] Y. Sack. *Adding Temporal Logic to Dynamic Epistemic Logic*. PhD thesis, Indiana University, Bloomington, USA, 2007.

[9] J.F.A.K. van Benthem, J.D. Gerbrandy, and E. Pacuit. Merging frameworks for interaction: DEL and ETL. In D. Samet, editor, *Proceedings of TARK 2007*, pages 72–81, 2007.

[10] J.F.A.K. van Benthem, J. van Eijck, and B.P. Kooi. Logics of communication and change. *Information and Computation*, 204(11):1620–1662, 2006.

[11] H.P. van Ditmarsch. Prolegomena to dynamic logic for belief revision. *Synthese (Knowledge, Rationality & Action)*, 147:229–275, 2005.

[12] H.P. van Ditmarsch and T. French. Simulation and information. (Electronic) Proceedings of LOFT 2008, Amsterdam, 2008.

[13] H.P. van Ditmarsch and B.P. Kooi. Semantic results for ontic and epistemic change. In G. Bonanno, W. van der Hoek, and M. Wooldridge, editors, *Post-proceedings of LOFT 2006*. Amsterdam University Press, 2008. To appear in the series Texts in Logic and Games.

[14] H.P. van Ditmarsch, J. Ruan, and W. van der Hoek. Model checking dynamic epistemics in branching time. In *(Informal) Proceedings of FAMAS 2007, Durham UK*, 2007.

[15] A. Yap. Product update and looking backward. Technical report, University of Amsterdam, 2006. ILLC Research Report PP-2006-39.