WIEBE FEST 2009

A workshop in honour of our friend and colleague
Professor Wiebe van der Hoek
on the occasion of his fiftieth birthday.

16 March 2009 — University of Liverpool

Edited by
Johan van Benthem, John-Jules Ch. Meyer,
Cees Witteveen, and Michael Wooldridge

Preface

In honour of our good friend and valued colleague Wiebe van der Hoek on the occasion of his 50th birthday, we would like to announce a 1-day interdisciplinary workshop on logical approaches to reasoning about knowledge and rational action – topics that have characterised Wiebe's research throughout his career.

Logics of knowledge and logics of action have a distinguished history in philosophy, and both areas have been given new directions and new energy over the past two decades by the interest of researchers in computer science and artificial intelligence. In this workshop we will bring together researchers from philosophy, logic, and computer science, to explore the state of the art in this area.

Topics of Interest

Include, but are not restricted to:

- logics combining knowledge and action (DEL, ...)
- logics for game theory and game theory for logics
- logics for belief revision
- logics of mental state (beliefs, desires, intentions, ...)
- logics for social laws and normative systems

Organisers

- Johan van Benthem, Amsterdam & Stanford
- John-Jules Meyer, Utrecht
- Cees Witteveen, Delft
- Mike Wooldridge, Liverpool

Contents

A Complete First-Order Logic of Knowledge and Time

Francesco Belardinelli
Scuola Normale Superiore, Pisa
F.Belardinelli@sns.it

Alessio Lomuscio
Department of Computing
Imperial College London
A.Lomuscio@imperial.ac.uk

February 22, 2009

Abstract

We introduce and investigate quantified interpreted systems, a semantics to reason about knowledge and time in a first-order setting. We provide an axiomatisation, which we show to be sound and complete. We utilise the formalism to study message passing systems [16, 8] in a first-order setting, and compare the results obtained to those available for the propositional case.

1 Introduction

The area of modal logic [3, 4] has received considerable attention in artificial intelligence over the years. Research has pursued both fundamental theoretical investigations (completeness, decidability, complexity, etc), as well as the use of modal formalisms in specification and automatic system verification, as in model checking [5].

Among the most well-known formalisms are propositional modal logics for reasoning about knowledge, or propositional epistemic logics [8, 21]. The typical epistemic language extends propositional logic by adding n modalities K_i representing the knowledge of agent i in a group $A = \{1, \ldots, n\}$ of agents. For expressiveness purposes, epistemic logic has been extended in several ways. In one direction, further modalities have been added to the formalism (distributed knowledge, common knowledge, belief, etc.) for representing the knowledge shared in a group of agents. In another one, the epistemic language has been enriched with temporal operators under the assumption of a given model of time (e.g., linear or branching, discrete or continuous, etc.). In all these lines of work there is a tension between extending the expressiveness of the language reflecting the system to be modeled and retaining some useful theoretical properties of the formalism, such as decidability.

This tension is still present in the exercise conducted here, where we aim at extending a combination of epistemic and temporal logic to predicate level. We apply this result in the modeling of a class of computational structures normally referred to as message passing systems [16]. We also show that known metatheoretical properties of message passing systems [8] become validities in the predicate logic here considered.

Our starting point is a number of results by Halpern, van der Meyden, and others regarding the combination of time and knowledge at propositional level [9, 19] together with studies by, among others, Hodkinson, Reynolds, Wolter, Zakharyaschev for first-order temporal logic including both positive [14, 25, 31] and negative results [32]. In this note we also make use of our initial work in this direction [2, 1], where static (i.e., non-temporal) quantified epistemic logics were axiomatised.

Our motivation for the above comes from an interest in reasoning about reactive, autonomous distributed systems, or multi-agent systems (MAS), whose high-level properties

may usefully be modeled by epistemic formalisms suitably extended to incorporate temporal logic. While temporal epistemic logics are well understood at propositional level [8, 21], their usefulness has been demonstrated in a number of applications (security and communication protocols, robotics), and model checking tools have been developed for them [12, 23, 7], still there is a growing need in web-services, security, as well as other areas, to extend these languages to first-order (see [18, 26, 29]). Moreover, a number of formalisms, including BDI logics [24], the $KQML$ framework [6], and $\mathcal{LORA}$ [33], have put forward agent theories that include the power of first-order quantification. However, most of these contributions do not address the issue of completeness, a core concern here.

In MAS applications the power of first-order logic is welcome every time agents' knowledge is concerned with:

- Relational statement, as in *agent i knows that message μ was sent by a to b*, or formally
$$K_i\langle P\rangle Send(a, b, \mu);$$
(where $\langle P\rangle$ is the diamond for past time);
- Functional dependency and identity: *at some future point agent i will know that message μ is the encryption of message μ' with key k*, formally
$$\langle F\rangle K_i(\mu = enc(k, \mu'));$$
- An infinite domain of individuals, or a finite domain whose cardinality cannot be bounded in advance: *agent i has to read an e-mail before deleting it*,
$$\forall\mu(Delete(i, \mu) \rightarrow \langle P\rangle Read(i, \mu));$$
- Quantification on agents [17]: *the child of any process knows which process launched it*
$$\forall i K_{child(i)}\langle P\rangle Launch(i, child(i))$$

Furthermore, in the context of logics for knowledge it is known that epistemic modalities can be combined with quantifiers to express concepts such as knowledge *de re* and *de dicto* [10, 15]. For instance, an agent i might know that every computation will eventually produce an output, thus having the *de dicto* knowledge expressed by the following specification:
$$\forall comp\ K_i\ \langle F\rangle\ \exists y\ Output(comp, y)$$
but she might not know the actual output of every computation. Therefore, the following *de re* specification:
$$\forall comp\ \exists y\ K_i\ \langle F\rangle\ Output(comp, y)$$
would not be satisfied. From the examples above we conclude that quantification can significantly extend the expressiveness of epistemic languages.

While the specifications above call for a first-order language, we need to consider why one should use an undecidable language when a decidable one (propositional temporal epistemic logic in our case) does a reasonable job already. Although this is a sensible objection, we should stress that in many practical applications, such as in model checking, we are typically not so much concerned with the validity problem but with satisfaction in a given model, which is often an easier problem, particularly for some classes of formulas. Additionally, recent research, including among others [14, 27, 28, 30], has put forward useful decidable fragments of first-order modal logic, thereby opening the way for further extensions.

We approach the problem by introducing quantified interpreted systems, an extension to first-order of "standard" interpreted systems [13, 22], which are used to interpret a language for temporal epistemic logic including distributed knowledge. First, a sound and complete axiomatisation is presented. Second, message passing systems, a basic framework for reasoning about asynchronous systems [16] are analysed in the light of the novel formalism, and the results compared to the treatment in propositional logic.

2 A Quantified Temporal Epistemic Logic

In this section we extend to first-order the formalism of interpreted systems, a class of structures introduced to model the behaviour of multi-agent systems [8, 21]. In what follows we assume a finite set $A = \{i_1, \ldots, i_n\}$ of agents.

2.1 Syntax

The first-order modal language $\mathcal{L}_n$ contains individual variables $x_1, x_2, \ldots$, n-ary functors $f_1^n, f_2^n, \ldots$ and n-ary predicative letters $P_1^n, P_2^n, \ldots$, for $n \in \mathbb{N}$, the identity predicate $=$, the propositional connectives $\neg$ and $\rightarrow$, the universal quantifier $\forall$, the epistemic operators K_i, for $i \in A$, the distributed knowledge operators D_G, for non-empty $G \subseteq A$, the future operator $[F]$, and the past operator $[P]$.

Definition 1 *Terms and formulas in the language $\mathcal{L}_n$ are defined in the Backus-Naur form as follows:*

$$t ::= x \,|\, f^k(\vec{t})$$
$$\phi ::= P^k(\vec{t}) \,|\, t = t' \,|\, \neg\phi \,|\, \phi \rightarrow \psi \,|\, K_i\phi \,|\, D_G\phi \,|\, [F]\phi \,|\, [P]\phi \,|\, \forall x\phi$$

The formula $K_i\phi$ means "*agent i knows ϕ*", while $D_G\phi$ represents "*ϕ is distributed knowledge among the agents in G*", and $[F]\phi$ (respectively $[P]\phi$) stands for "*ϕ will always be true*" (respectively "*ϕ has always been true*"). The symbols $\bot$, $\wedge$, $\vee$, $\leftrightarrow$, $\exists$, $\langle F\rangle$ (sometime in the future), $\langle P\rangle$ (sometime in the past) are defined as standard. The temporal operators $[F]^+$ (every future time including the present) and $[P]^+$ (every past time including the present) can be defined as $\phi \wedge [F]\phi$ and $\phi \wedge [P]\phi$ respectively.

We refer to 0-ary functors as *individual constants* $c_1, c_2, \ldots$ A closed term v is a term where no variable appears; closed terms are either constants or terms obtained by applying functors to closed terms.

By $t[\vec{y}]$ (resp. $\phi[\vec{y}]$) we mean that $\vec{y} = y_1, \ldots, y_n$ are all the free variables in t (resp. ϕ); while $t[\vec{y}/\vec{t}]$ (resp. $\phi[\vec{y}/\vec{t}]$) denotes the term (resp. formula) obtained by substituting simultaneously some, possibly all, free occurrences of $\vec{y}$ in t (resp. ϕ) with $\vec{t} = t_1, \ldots, t_n$, renaming bounded variables if necessary.

2.2 Quantified Interpreted Systems

Interpreted systems are widely used to model the behaviour of MAS, in this subsection we extend these structures to first-order. This extension can be performed in several ways, all leading to different results. For instance, we could introduce a domain of quantification for each agent and/or for each computational state (see [2, 1] for a discussion of the static case). In this paper we consider the simplest extension, obtained by adding a single quantification domain D common to all agents and states. We present further options in the conclusions.

More formally, for each agent $i \in A$ in a multi-agent system we introduce a set L_i of local states $l_i, l_i', \ldots$, and a set Act_i of actions $\alpha_i, \alpha_i', \ldots$. We consider local states and actions for the environment e as well. The set $\mathcal{S} \subseteq L_e \times L_1 \times \ldots \times L_n$ contains all possible global states of the MAS, while $Act \subseteq Act_e \times Act_1 \times \ldots \times Act_n$ is the set of all possible joint actions. Note that some states may never be reached and some joint actions may never

be performed. We also introduce a transition function $\tau : Act \to (\mathcal{S} \to \mathcal{S})$. Intuitively, $\tau(\alpha)(s) = s'$ encodes that the agents can access the global state s' from s by performing the joint action $\alpha \in Act$. The transition function τ defines the admissible evolutions of the MAS. We say that the global state s' is *reachable in one step* from s, or $s \prec s'$, iff there is $\alpha \in Act$ such that $\tau(\alpha)(s) = s'$; while s' is *reachable* from s iff $s \prec^+ s'$, where $\prec^+$ is the transitive closure of relation $\prec$.

To represent the temporal evolution of the MAS we consider the flow of time $\mathcal{T} = \langle T, < \rangle$ defined as a weakly connected, strict partial order, i.e., T is a non-empty set and the relation $<$ on T is irreflexive, transitive and weakly connected: for n, n', n'' in T,

- $n \not< n$
- $(n < n' \wedge n' < n'') \to (n < n'')$
- $(n < n' \wedge n < n'') \to (n' < n'' \vee n'' < n' \vee n' = n'')$
- $(n' < n \wedge n'' < n) \to (n' < n'' \vee n'' < n' \vee n' = n'')$

The relation $<$ can be thought of as the precedence relation on the set T of moments in time. A run r over $\langle \mathcal{S}, Act, \tau, \mathcal{T} \rangle$, where $\mathcal{S}$, Act, τ, and $\mathcal{T}$ are defined as above, is a function from T to $\mathcal{S}$ such that $n < n'$ implies $r(n) \prec^+ r(n')$. Intuitively, a run represents a possible evolution of the MAS on the flow of time $\mathcal{T}$.

We now define the quantified interpreted systems for the language $\mathcal{L}_n$ as follows:

Definition 2 *A quantified interpreted system, or QIS, over $\langle \mathcal{S}, Act, \tau, \mathcal{T} \rangle$ is a triple $\mathcal{P} = \langle R, D, I \rangle$ such that R is a non-empty set of runs over $\langle \mathcal{S}, Act, \tau, \mathcal{T} \rangle$; D is a non-empty set of individuals; $I(f^k)$ is a k-ary function from D^k to D; for $r \in R$, $n \in T$, $I(P^k, r, n)$ is a k-ary relation on D and $I(=, r, n)$ is the equality on D. We denote by $\mathcal{QIS}$ the class of all quantified interpreted systems.*

Note that individual constants as well as functors in $\mathcal{L}_n$ are interpreted rigidly, that is, their interpretation is the same in every global state. Further, the present definition of quantified interpreted systems covers the most intuitive formalisations of time, as it includes $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$ with a notion of precedence among instants. Therefore, QIS are general enough to cover a wide range of cases, while still being interesting for applications.

Now we assign a meaning to the formulas of $\mathcal{L}_n$ in quantified interpreted systems. Following standard notation [8] a pair (r, m) is a *point* in $\mathcal{P}$. If $r(m) = \langle l_e, l_1, \ldots, l_n \rangle$ is the global state at (r, m), then $r_e(m) = l_e$ and $r_i(m) = l_i$ are the environment's and agent i's local state at (r, m) respectively. We consider also the converse relation $>$ defined as $n > m$ iff $m < n$, and the partial order $\leq$ such that $n \leq m$ iff $n < m$ or $n = m$.

Let σ be an assignment from the variables in $\mathcal{L}_n$ to the individuals in D, the valuation $I^\sigma(t)$ of a term t is defined as $\sigma(y)$ for $t = y$, and $I^\sigma(t) = I(f^k)(I^\sigma(t_1), \ldots, I^\sigma(t_k))$, for $t = f(\vec{t})$. A variant $\sigma\binom{x}{a}$ of an assignment σ assigns $a \in D$ to x and coincides with σ on all the other variables.

Definition 3 *The satisfaction relation $\models$ for $\phi \in \mathcal{L}_n$, $(r, m) \in \mathcal{P}$, and an assignment σ is defined as follows:*

$(\mathcal{P}^\sigma, r, m) \models P^k(\vec{t})$ iff $\langle I^\sigma(t_1), \ldots, I^\sigma(t_k) \rangle \in I(P^k, r, m)$
$(\mathcal{P}^\sigma, r, m) \models t = t'$ iff $I^\sigma(t) = I^\sigma(t')$
$(\mathcal{P}^\sigma, r, m) \models \neg\psi$ iff $(\mathcal{P}^\sigma, r, m) \not\models \psi$
$(\mathcal{P}^\sigma, r, m) \models \psi \to \psi'$ iff $(\mathcal{P}^\sigma, r, m) \not\models \psi$ or $(\mathcal{P}^\sigma, r, m) \models \psi'$
$(\mathcal{P}^\sigma, r, m) \models K_i\psi$ iff $r_i(m) = r'_i(m')$ implies $(\mathcal{P}^\sigma, r', m') \models \psi$
$(\mathcal{P}^\sigma, r, m) \models D_G\psi$ iff $r_i(m) = r'_i(m')$ for all $i \in G$, implies $(\mathcal{P}^\sigma, r', m') \models \psi$
$(\mathcal{P}^\sigma, r, m) \models [F]\psi$ iff $m < m'$ implies $(\mathcal{P}^\sigma, r, m') \models \psi$
$(\mathcal{P}^\sigma, r, m) \models [P]\psi$ iff $m > m'$ implies $(\mathcal{P}^\sigma, r, m') \models \psi$
$(\mathcal{P}^\sigma, r, m) \models \forall x\psi$ iff for all $a \in D$, $(\mathcal{P}^{\sigma\binom{x}{a}}, r, m) \models \psi$

The truth conditions for $\bot$, $\wedge$, $\vee$, $\leftrightarrow$, $\exists$, $\langle F \rangle$, and $\langle P \rangle$ are defined from those above. In particular, the temporal operators $[F]^+$ and $[P]^+$ respect the intended semantics:

$(\mathcal{P}^\sigma, r, m) \models [F]^+\psi$ iff $m \leq m'$ implies $(\mathcal{P}^\sigma, r, m') \models \psi$
$(\mathcal{P}^\sigma, r, m) \models [P]^+\psi$ iff $m \geq m'$ implies $(\mathcal{P}^\sigma, r, m') \models \psi$

A formula $\phi \in \mathcal{L}_n$ is said to be *true at a point* (r, m) iff it is satisfied at (r, m) by every σ; ϕ is *valid on a QIS* $\mathcal{P}$ iff it is true at every point in $\mathcal{P}$; ϕ is *valid on a class* $\mathcal{C}$ *of QIS* iff it is valid on every QIS in $\mathcal{C}$.

The present definition of QIS is based on two assumptions. Firstly, the domain D of individuals is the same for every agent i, so all agents reason about the same objects. This choice is consistent with the *external account of knowledge* usually adopted in the framework of interpreted systems: if knowledge is ascribed to agents by an external observer, i.e., the specifier of the system, it seems natural to focus on the set of individuals assumed to exist by the observer. Secondly, the domain D is assumed to be the same for every global state, i.e., no individual appears nor disappears in moving from one state to another. This also can be justified by the external account of knowledge: all individuals are supposed to be existing from the observer's viewpoint. However, either assumption can be relaxed to accommodate agent-indexed domains as well as individuals appearing and disappearing in the flow of time. We discuss further options in the conclusions. Finally, it can be the case that $A \subseteq D$: this means that the agents can reason about themselves, their properties, and relationships.

2.3 Expressiveness

Clearly, the language $\mathcal{L}_n$ is extremely expressive. We can use it to specify the temporal evolution of agents' knowledge, as well as the knowledge agents have of temporal facts about individuals. Both features are exemplified in the following specification: *agent i will know that someone sent him a message when he receives it*,

$$\forall j, \mu\, [F]\, (Rec(i, j, \mu) \rightarrow K_i\, \langle P \rangle\, Send(j, i, \mu)) \tag{1}$$

In $\mathcal{L}_n$ we can also express that *if agent i receives a message, then he will know that someone sent it to him*:

$$\forall \mu\, [F](\exists j\, Rec(i, j, \mu) \rightarrow K_i\, \exists j'\, \langle P \rangle\, Send(j', i, \mu)) \tag{2}$$

The latter specification is weaker than the former: (2) says nothing about the identity of the sender, while (1) requires that *the receiver knows the identity of the sender*. Further, we can express the fact that the existence of a sender is assumed only at the time the message is sent:

$$\forall \mu\, [F](\exists j\, Rec(i, j, \mu) \rightarrow K_i\, \langle P \rangle\, \exists j'\, Send(j', i, \mu))$$

In the section on message passing systems we provide further examples of the expressiveness of $\mathcal{L}_n$. Most importantly, we will show that this expressiveness is attained while retaining completeness.

We conclude this paragraph by considering some relevant validities on the class of QIS. Given that the domain of quantification is the same in every global state, both the Barcan formula and its converse are valid on the class of all QIS for all primitive modalities:

$$\begin{aligned}
\mathcal{QIS} &\models \forall x K_i\phi \leftrightarrow K_i \forall x \phi \\
\mathcal{QIS} &\models \forall x D_G\phi \leftrightarrow D_G \forall x \phi \\
\mathcal{QIS} &\models \forall x [F]\phi \leftrightarrow [F] \forall x \phi \\
\mathcal{QIS} &\models \forall x [P]\phi \leftrightarrow [P] \forall x \phi
\end{aligned}$$

Also, these validities are in line with the bird's eye approach usually adopted in epistemic logic. However, should we wish to do so, we can drop them by introducing quantified interpreted systems with varying domains.

For what concerns identity, the following principles hold:

$\mathcal{QIS} \models t = t' \rightarrow K_i(t = t')$ $\quad$ $\mathcal{QIS} \models t \neq t' \rightarrow K_i(t \neq t')$
$\mathcal{QIS} \models t = t' \rightarrow D_G(t = t')$ $\quad$ $\mathcal{QIS} \models t \neq t' \rightarrow D_G(t \neq t')$
$\mathcal{QIS} \models t = t' \rightarrow [F](t = t')$ $\quad$ $\mathcal{QIS} \models t \neq t' \rightarrow [F](t \neq t')$
$\mathcal{QIS} \models t = t' \rightarrow [P](t = t')$ $\quad$ $\mathcal{QIS} \models t \neq t' \rightarrow [P](t \neq t')$

These validities, which hold because of rigid designation, are consistent with the external account of knowledge. However, should we require terms whose denotations depends on the epistemic states of agents, or change accordingly to the evolution of the MAS, we can consider introducing *flexible* terms in the language [1]. In such an extended formalism none of the validities above holds whenever t and t' are flexible terms.

3 The System QKT.S5$_n$

In this section we provide a sound and complete axiomatisation of quantified interpreted systems. This result shows that, even though language $\mathcal{L}_n$ is highly expressive, QIS provide a perfectly adequate semantics for it. This also opens the possibility of developing automated verification methods for the formalism. We first prove the completeness of the first-order multi-modal system QKT.S5$_n$ with respect to Kripke models. The proof presented here is an extension of [11], where completeness of a first-order temporal language on weakly-connected partial orders was presented. Then, by means of a map from Kripke models to QIS, the completeness of QKT.S5$_n$ with respect to $\mathcal{QIS}$ follows.

The system QKT.S5$_n$ is a first-order multi-modal version of the propositional system $S5$ combined with a linear temporal logic. Although tableaux proof systems and natural deduction calculi are more suitable for automated theorem proving, Hilbert-style systems are easier to handle for the completeness proof. Hereafter we list the postulates of QKT.S5$_n$. Note that $\Rightarrow$ is the inference relation between formulas, while $\Box$ is a placeholder for any primitive modality in $\mathcal{L}_n$ (both temporal and epistemic).

Definition 4 *The system QKT.S5$_n$ on $\mathcal{L}_n$ contains the following schemes of axioms and inference rules:*

Taut	*every instance of classic propositional tautologies*
MP	$\phi \rightarrow \psi, \phi \Rightarrow \psi$
Dist	$\Box(\phi \rightarrow \psi) \rightarrow (\Box\phi \rightarrow \Box\psi)$
4	$\Box\phi \rightarrow \Box\Box\phi$
Nec	$\phi \Rightarrow \Box\phi$
T	$K_i\phi \rightarrow \phi$ $\quad$ $D_G\phi \rightarrow \phi$
5	$\neg K_i\phi \rightarrow K_i\neg K_i\phi$ $\quad$ $\neg D_G\phi \rightarrow D_G\neg D_G\phi$
D1	$D_{\{i\}}\phi \leftrightarrow K_i\phi$
D2	$D_G\phi \rightarrow D_{G'}$, *for* $G \subseteq G'$
FP	$\phi \rightarrow [F]\langle P\rangle\phi$
PF	$\phi \rightarrow [P]\langle F\rangle\phi$
WConF	$\langle P\rangle\langle F\rangle\phi \rightarrow (\langle P\rangle\phi \vee \phi \vee \langle F\rangle\phi)$
WConP	$\langle F\rangle\langle P\rangle\phi \rightarrow (\langle P\rangle\phi \vee \phi \vee \langle F\rangle\phi)$
Ex	$\forall x\phi \rightarrow \phi[x/t]$
Gen	$\phi \rightarrow \psi[x/t] \Rightarrow \phi \rightarrow \forall x\psi$, *where* x *is not free in* ϕ
Id	$t = t$
Func	$t = t' \rightarrow (t''[x/t] = t''[x/t'])$
Subst	$t = t' \rightarrow (\phi[x/t] \rightarrow \phi[x/t'])$

By the definition above the operators K_i and D_G are $S5$ type modalities, while the future $[F]$ and past $[P]$ operators are axiomatised as linear-time modalities. To this we add the classic theory of quantification, consisting of postulates Ex and Gen, which are both sound in our interpretation as we are considering a unique domain of individuals. Finally, we have the axioms for identity.

We consider the standard definitions of *proof* and *theorem*: $\vdash \phi$ means that $\phi \in \mathcal{L}_n$ is a theorem in QKT.S5$_n$. A formula $\phi \in \mathcal{L}_n$ is *derivable* in QKT.S5$_n$ from a set Δ of formulas, or $\Delta \vdash \phi$, iff $\vdash \phi_1 \wedge \ldots \wedge \phi_n \rightarrow \phi$ for some $\phi_1, \ldots, \phi_n \in \Delta$.

It can be easily checked that the axioms of QKT.S5$_n$ are valid on every QIS and the inference rules preserve validity. As a consequence, we have the following soundness result:

Theorem 5 (Soundness) *The system QKT.S5$_n$ is sound for the class $\mathcal{QIS}$ of quantified interpreted systems.*

Now we show that the axioms in QKT.S5$_n$ are not only necessary, but also sufficient to prove all validities on $\mathcal{QIS}$.

3.1 Kripke Models

Although quantified interpreted systems are useful for modeling MAS, for showing that QKT.S5$_n$ is complete with respect to $\mathcal{QIS}$ we introduce an appropriate class of Kripke models [3, 4], which are more suitable for theoretical investigations, namely, the completeness proof.

Definition 6 *A Kripke model, or K-model, for the language $\mathcal{L}_n$ is a tuple $\mathcal{M} = \langle W, \{\sim_i\}_{i\in A}, <, D, I\rangle$ such that W is a non-empty set; for $i \in A$, $\sim_i$ is an equivalence relation on W; $<$ is a weakly connected, strict partial order on W; D is a non-empty set of individuals; $I(f^k)$ is a k-ary function from D^k to D; for $w \in W$, $I(P^k, w)$ is a k-ary relation on D, and $I(=, w)$ is the equality on D. The class of all Kripke models is denoted by $\mathcal{K}$.*

Further, the satisfaction relation $\models$ for an assignment σ is inductively defined as follows:

$(\mathcal{M}^\sigma, w) \models P^k(\vec{t})$ iff $\langle I^\sigma(t_1), \ldots, I^\sigma(t_k)\rangle \in I(P^k, w)$
$(\mathcal{M}^\sigma, w) \models t = t'$ iff $I^\sigma(t) = I^\sigma(t')$
$(\mathcal{M}^\sigma, w) \models \neg\psi$ iff $(\mathcal{M}^\sigma, w) \not\models \psi$
$(\mathcal{M}^\sigma, w) \models \psi \rightarrow \psi'$ iff $(\mathcal{M}^\sigma, w) \not\models \psi$ or $(\mathcal{M}^\sigma, w) \models \psi'$
$(\mathcal{M}^\sigma, w) \models [F]\psi$ iff $w < w'$ implies $(\mathcal{M}^\sigma, w') \models \psi$
$(\mathcal{M}^\sigma, w) \models [P]\psi$ iff $w > w'$ implies $(\mathcal{M}^\sigma, w') \models \psi$
$(\mathcal{M}^\sigma, w) \models K_i\psi$ iff $w \sim_i w'$ implies $(\mathcal{M}^\sigma, w') \models \psi$
$(\mathcal{M}^\sigma, w) \models D_G\psi$ iff $(w, w') \in \bigcap_{i\in G} \sim_i$ implies $(\mathcal{M}^\sigma, w') \models \psi$
$(\mathcal{M}^\sigma, w) \models \forall x\psi$ iff for all $a \in D$, $(\mathcal{M}^{\sigma\binom{x}{a}}, w) \models \psi$

We formally compare Kripke models to quantified interpreted systems by means of a map $g : \mathcal{K} \rightarrow \mathcal{QIS}$. Let $\mathcal{M} = \langle W, \{\sim_i\}_{i\in A}, <, D, I\rangle$ be a Kripke model. For every equivalence relation $\sim_i$, for $w \in W$, let the equivalence class $[w]_{\sim_i} = \{w' \mid w \sim_i w'\}$ be a local state for agent i; while W is the set of local states for the environment. Let $\langle W, <\rangle$ be the irreflexive, transitive and weakly connected flow of time. Then define $g(\mathcal{M})$ as the triple $\langle R, D, I'\rangle$, where R contains the run r such that $r(w) = \langle w, [w]_{\sim_1}, \ldots, [w]_{\sim_n}\rangle$ for $w \in W$, D is the same as in $\mathcal{M}$, and $I'(P^k, r, w) = I(P^k, w)$. The structure $g(\mathcal{M})$ is a QIS that satisfies the following result:

Lemma 7 *For every $\phi \in \mathcal{L}_n$, $w \in W$,*

$$(\mathcal{M}^\sigma, w) \models \phi \quad \text{iff} \quad (g(\mathcal{M})^\sigma, r, w) \models \phi$$

where r is the only run in $g(\mathcal{M})$. We refer to the appendix for a proof of this lemma.

3.2 Completeness

We show that the system QKT.S5$_n$ is complete by extending to first-order the proof for the propositional system $S5_n^D$ in [9], together with the completeness proof for the first-order temporal logic discussed in [11]. The relevance of our result consists in showing that these two methods can be combined together to prove an original completeness result, as long as there is no interaction between epistemic and temporal modalities. Note that an independent completeness proof for $S5_n^D$ appeared in [20].

More formally, we show that if QKT.S5$_n$ does not prove a formula $\phi \in \mathcal{L}_n$, then the canonical model $\mathcal{M}^{QKT.S5_n}$ for QKT.S5$_n$ does not pseudo-validate ϕ. It is not guaranteed that pseudo-validity (as defined below) coincides with plain validity, but by results in [9, 11] from $\mathcal{M}^{QKT.S5_n}$ we can obtain a K-model $\mathcal{M}^+$ such that $\mathcal{M}^{QKT.S5_n}$ pseudo-validates ϕ iff $\mathcal{M}^+ \models \phi$, and completeness follows.

In order to prove the first part of the completeness result we rely on two lemmas: the *saturation lemma* and the *truth lemma*, whose statements require the following definitions: let Λ be a set of formulas in $\mathcal{L}_n$,

Λ is *consistent* iff $\Lambda \nvdash \bot$;
Λ is *maximal* iff for every $\phi \in \mathcal{L}_n$, $\phi \in \Lambda$ or $\neg\phi \in \Lambda$;
Λ is *max-cons* iff Λ is consistent and maximal;
Λ is *rich* iff $\exists x\phi \in \Lambda \Rightarrow \phi[x/c] \in \Lambda$, for some $c \in \mathcal{L}_n$;
Λ is *saturated* iff Λ is max-cons and rich.

Assume that QKT.S5$_n$ does not prove ϕ, then the set $\{\neg\phi\}$ is consistent, and by the saturation lemma below $\{\neg\phi\}$ can be extended to a saturated set:

Lemma 8 (Saturation [15]) *If Δ is a consistent set of formulas in $\mathcal{L}_n$, then it can be extended to a saturated set Π of formulas on some expansion $\mathcal{L}_n^+$ obtained by adding an infinite enumerable set of new individual constants to $\mathcal{L}_n$.*

Now we introduce the canonical model for QKT.S5$_n$. Note that $\wp^+(A)$ is the set of non-empty sets of agents.

Definition 9 (Canonical model) *The canonical model for QKT.S5$_n$ on the language $\mathcal{L}_n$, with an expansion $\mathcal{L}_n^+$, is a tuple $\mathcal{M}^{QKT.S5_n} = \langle W, \{R_j\}_{j \in A \cup \wp^+(A)}, <, D, I\rangle$ such that:*

- *W is the set of saturated sets of formulas in $\mathcal{L}_n^+$;*
- *for $i \in A$, $w, w' \in W$, wR_iw' iff $\{\phi \mid K_i\phi \in w\} \subseteq w'$;*
- *for non-empty $G \subseteq A$, wR_Gw' iff $\{\phi \mid D_G\phi \in w\} \subseteq w'$;*
- *for $w, w' \in W$, $w < w'$ iff $\{\phi \mid [F]\phi \in w\} \subseteq w'$;*
- *D is the set of equivalence classes $[v] = \{v' \mid v = v' \in w\}$, for each closed term $v \in \mathcal{L}_n^+$;*
- *$I(f^k)([v_1], \ldots, [v_k]) = [f^k(v_1, \ldots, v_k)]$;*
- *$\langle [v_1], \ldots, [v_k]\rangle \in I(P^k, w)$ iff $P^k(v_1, \ldots, v_k) \in w$.*

If $QKT.S5_n \nvdash \phi$, then by the saturation lemma there is a saturated set $w \supseteq \{\neg\phi\}$, so the set W of possible worlds is non-empty. Since T, 4 and 5 are axioms of QKT.S5$_n$, the various R_i and R_G are equivalence relations. Moreover, from $D1$ and $D2$ it follows that $R_{\{i\}}$ is equal to R_i and $R_G \subseteq \bigcap_{i \in G} R_i$. However, in general $R_G \neq \bigcap_{i \in G} R_i$ [9]. On the other hand, the relation $<$ is transitive and weakly connected by axioms 4, $WConF$, $WConP$. By FP, PF the relation $w > w'$ defined as $\{\phi \mid [P]\phi \in w\} \subseteq w'$ is the converse of $<$. However, $<$ might not be irreflexive [11].

These remarks give the *rationale* for introducing the pseudo-satisfaction relation $\models^p$, defined as $\models$ but for the distributed knowledge operator D_G (in what follows we simply write $\mathcal{M}$ for $\mathcal{M}^{QKT.S5_n}$):

$$(\mathcal{M}^\sigma, w) \models^p D_G\psi \quad \text{iff} \quad wR_Gw' \text{ implies } (\mathcal{M}^\sigma, w') \models^p \psi$$

We state the *truth lemma* for the pseudo-satisfaction relation $\models^p$ and refer to [9] for a proof.

Lemma 10 (Truth lemma) *Let* $w \in \mathcal{M}, \psi \in \mathcal{L}_n^+, \sigma(y_i) = [v_i]$,

$$(\mathcal{M}^\sigma, w) \models^p \psi[\vec{y}] \quad \text{iff} \quad \psi[\vec{y}/\vec{v}] \in w$$

We remarked that the canonical model $\mathcal{M}$ might not satisfy $\bigcap_{i\in G} R_i = R_G$. However, by applying the techniques in [9] $\mathcal{M}$ can be unwound to get a K-model $\mathcal{M}'$ in such a way that $R_G = \bigcap_{i\in G} R_i$ and the same formulas hold. We refer to the appendix for a proof of the following lemma.

Lemma 11 *For every* $\psi \in \mathcal{L}_n^+$,

$$\mathcal{M}' \models \psi \quad \text{iff} \quad \mathcal{M} \models^p \psi$$

In conclusion, if QKT.S5$_n \nvdash \phi$, then the canonical model $\mathcal{M}$ pseudo-satisfies $\neg\phi$ by lemma 10. By lemma 11 we obtain that the K-model $\mathcal{M}'$ does not validate ϕ.

Note that the relation $<'$ on W' might not be irreflexive, as $<$ on W is not such. However, we can apply the techniques in [11] to construct an irreflexive K-model $\mathcal{M}^+$ from $\mathcal{M}'$ such that:

Lemma 12 *For every* $\psi \in \mathcal{L}_n^+$,

$$\mathcal{M}^+ \models \psi \quad \text{iff} \quad \mathcal{M}' \models \psi$$

Also in this case we refer to the appendix for a proof.

By lemma 12 we conclude that the K-model $\mathcal{M}^+$ falsifies the unprovable formula ϕ. Therefore, the following completeness result holds:

Theorem 13 (Completeness) *The system QKT.S5$_n$ is complete for the class $\mathcal{K}$ of Kripke models.*

In order to prove completeness for the class $\mathcal{QIS}$ consider the quantified interpreted system $g(\mathcal{M}^+)$. In lemma 7 we showed that $\mathcal{M}^+ \models \phi$ iff $g(\mathcal{M}^+) \models \phi$, hence $g(\mathcal{M}^+)$ satisfies $\neg\phi$. As a result, we have the following implications and a further completeness result:

$$\mathcal{QIS} \models \phi \quad \Rightarrow \quad \mathcal{K} \models \phi \quad \Rightarrow \quad QKT.S5_n \vdash \phi$$

Theorem 14 (Completeness) *The system QKT.S5$_n$ is complete for the class $\mathcal{QIS}$ of quantified interpreted systems.*

By combining together the soundness and completeness theorems we can compare directly the axiomatisation QKT.S5$_n$ and QIS, so we state our main result:

Corollary 15 (Soundness and Completeness) *A formula $\phi \in \mathcal{L}_n$ is valid on the class $\mathcal{QIS}$ of quantified interpreted systems iff ϕ is provable in QKT.S5$_n$.*

4 Message Passing Systems as QIS

In this section we model message passing systems [8, 16] in the framework of QIS. A message passing system (MPS) is a MAS in which the only external actions for the agents are message exchanges, specifically sending and receiving messages. This setting is common in the study of a variety of distributed systems, well beyond the realms of MAS and AI. Indeed, any synchronous or asynchronous networked system can be seen as an MPS.

The notion of time is crucial for the analysis of the ordering of events in MPS. As remarked in [16], a message μ can be said to have been sent (received) before message μ' if μ was sent (respectively received) at an earlier time than μ'. We can of course specify this condition in terms of an external global clock. However, maintaining synchronicity in a distributed system is known to be costly. An alternative is to study asynchronous MPS (or AMPS), where only internal clocks exist and agents can work at arbitrary rates relative to each other.

In what follows we show how both (synchronous) MPS and AMPS can be thought of as particular classes of QIS satisfying a finite number of specifications expressed in the first-order modal language $\mathcal{L}_n$. Further, we analyse in detail the agents' knowledge about the ordering of events in AMPS. Our main result consists in showing that the characterisation of AMPS at propositional level given as a metatheorem (specifically, in [8], Proposition 4.4.3) can naturally be cast as a formula in $\mathcal{L}_n$, which turns out to be a validity on the class of QIS we introduce. While the basic details are given below, we refer to [8], sections 4.4.5-6, for more details on MPS.

We introduce a set Act of actions $\alpha_1, \alpha_2, \ldots$, and a set Msg of messages $\mu_1, \mu_2, \ldots$ For each agent $i \in A$, we consider a set Σ_i of initial events $init(i, \alpha)$, and a set Int_i of internal events $int(i, \alpha)$. We define the local state l_i for agent i as a *history* over Σ_i, Int_i and Msg, that is, a sequence of events whose first element is in Σ_i, and whose following elements either belong to Int_i or are events of the form $send(i, j, \mu)$, $rec(i, j, \mu)$ for $j \in A$, $\mu \in Msg$. Intuitively, $init(i, \alpha)$ represents the event where *agent i performs the initial action α*, $send(i, j, \mu)$ represents the event where *agent i sends message μ to j*, while the meaning of $rec(i, j, \mu)$ is that *agent i receives message μ from j*. Finally, $int(i, \alpha)$ means that *agent i performs the internal action α*.

A global state $s \in \mathcal{S}$ is a tuple $\langle l_e, l_1, \ldots, l_n \rangle$, where $l_1, \ldots, l_n$ are local states as above, and l_e contains all the events in $l_1, \ldots, l_n$. In what follows we assume that the natural numbers $\mathbb{N}$ as the flow of time. This choice implies that we cannot provide a complete characterisation of MPS in this formalism, as first-order temporal logic on $\mathbb{N}$ is unaxiomatisable [11]. Still, we can express a number of interesting properties of MPS in the language $\mathcal{L}_n$.

A run r over $\langle \mathcal{S}, \mathbb{N} \rangle$ is a function from the natural numbers $\mathbb{N}$ to $\mathcal{S}$ such that:

MP1 $r_i(m)$ is a history over Σ_i, Int_i and Msg;

MP2 for every event $rec(i, j, \mu)$ in $r_i(m)$ there exists a corresponding event $send(j, i, \mu)$ in $r_j(m)$.

MP3 $r_i(0)$ is a sequence of length one (the initial state $init(i, \alpha)$), and $r_i(m + 1)$ is either identical to $r_i(m)$ or results from appending an event to $r_i(m)$.

The last specification MP4 has only a simplifying purpose and does not restrict our analysis:

MP4 All events in a given agent's history are distinct. An agent can never perform the same action twice in a given run.

By MP1 the local states of each agent records her initial state, the messages she has sent or received, as well as the internal actions she has taken. MP2 guarantees that any

received message was actually sent, while MP3 specifies that at each step at most a single event occurs to any agent. Finally, MP4 is not essential, but it simplifies proofs as we do not have to distinguish different occurrences of the same action by, for example, time-stamping actions. We will use this constraint throughout the present section without explicitly mentioning it.

We now define message passing QIS (MPQIS) as a particular class of quantified interpreted systems $\mathcal{P} = \langle R, D, I \rangle$, where R is a non-empty set of runs satisfying the constraints MP1-4 above, D contains the agents in A, the actions in Act, the messages in Msg, and the events $e_1, e_2, \ldots$, and I is an interpretation for $\mathcal{L}_n$. We assume that our language has terms and predicative letters for representing the objects in the domain D and the relations among them. In particular, $e_1, e_2, \ldots$ are metaterms ranging over events; for instance, $\forall e \phi[e]$ is a shorthand for

$$\forall i, j, \mu \; \phi[send(i, j, \mu)] \wedge \phi[rec(i, j, \mu)] \wedge \phi[init(i)] \wedge \phi[int(i, \alpha)]$$

where $\phi[t]$ means that the term t occurs in the formula ϕ.

We use the same notation for the objects in the model and the syntactic elements, the distinction will be clear by the context.

For the specification of MPS it is useful to introduce a predicative constant H for *happens* such that $(\mathcal{P}^\sigma, r, m) \models H(e, i)$ iff the event e occurs to agent i at time m in run r, i.e., $r_i(m)$ is the result of appending e to $r_i(m-1)$. We write $H(e)$ as a shorthand for $\exists i H(e, i)$. By definition of the environment's local state, $(\mathcal{P}^\sigma, r, m) \models H(e)$ iff e occurs at time m in run r. Also, we introduce the predicate $H'ed(e, i)$ for *happened* as $\langle P \rangle^+ H(e, i)$, and $H'ed(e) := \exists i H'ed(e, i)$. Finally, $Sent(i, j, \mu)$, $Recd(i, j, \mu)$, $Init(i, \alpha)$, and $Int(i, \alpha)$ are shorthands for $H'ed(send(i, j, \mu))$, $H'ed(rec(i, j, \mu))$, $H'ed(init(i, \alpha))$, and $H'ed(int(i, \alpha))$ respectively.

Let us now explore the range of specifications that can be expressed in the formalism. A property often required is *channel reliability*. We express this by stating that every sent message is eventually received. According to the definition of message passing QIS, it is possible that a message is lost during a run of the system. We can force channel reliability by requiring the following specification on MPQIS:

$$\forall i, j, \mu (Sent(i, j, \mu) \rightarrow \langle F \rangle^+ Recd(j, i, \mu))$$

Another relevant property of MPQIS concerns *authentication*: if agent i has received a message μ from agent j, then i knows that μ had actually been sent by j. This specification can be expressed as:

$$\forall j, \mu (Recd(i, j, \mu) \rightarrow K_i Sent(j, i, \mu))$$

Further, we may require that agents have *perfect recall*, that is, they know everything that has happened to them:

$$\forall e (H'ed(e, i) \rightarrow K_i H'ed(e, i))$$

It is easy to show that by definition MPQIS satisfy authentication and perfect recall but not channel reliability.

We anticipated that the formalism of QIS is powerful enough for expressing the specifications MP1-4 in $\mathcal{L}_n$. Moreover, we can reason about the knowledge agents have of the ordering of events in asynchronous MPS. To show this, we define $Prec(e, e', i)$ as a shorthand for:

$$H'ed(e', i) \wedge H'ed(e, i) \wedge [P]^+ (H'ed(e', i) \rightarrow H'ed(e, i))$$

It follows that $(\mathcal{P}^\sigma, r, m) \models Prec(e, e', i)$ iff events e and e' both occur to agent i by round m of run r, and e occurs no later than e' in r. Also, the ordering $Prec(e, e')$ is defined as:

$$H'ed(e') \wedge H'ed(e) \wedge [P]^+ (H'ed(e') \rightarrow H'ed(e))$$

Note that in the propositional language of [8] $Prec(e, e')$ is assumed as a primitive proposition.

We can express that the events in a state $r(m)$ are partially ordered by specifying that $Prec(e, e')$ is a reflexive and transitive relation on the set of past events:

$$\forall e \ (H'ed(e) \rightarrow Prec(e, e)) \tag{3}$$

$$\forall e, e', e'' \ (Prec(e, e') \wedge Prec(e', e'') \rightarrow Prec(e, e'')) \tag{4}$$

Moreover, $Prec(e, e', i)$ can be defined as an anti-symmetric, linear, discrete order on the events in $r_i(m)$, where with each non-final point is associated an immediate successor, that is, it is also anti-symmetric and total:

$$\forall e, e' \ (Prec(e, e', i) \wedge Prec(e', e, i) \rightarrow (e = e')) \tag{5}$$

$$\forall e, e' \ (H'ed(e, i) \wedge H'ed(e', i) \rightarrow Prec(e, e', i) \vee Prec(e', e, i)) \tag{6}$$

and each non-final point has an immediate successor:

$$\begin{aligned} \forall e, e'(Prec(e, e', i) \rightarrow \exists e'' \ (Prec(e, e'', i) \wedge \\ \wedge \neg \exists e'''(Prec(e, e''', i) \wedge Prec(e''', e'', i)))) \end{aligned} \tag{7}$$

We define $LinDisc(Prec(e, e', i))$ as the conjunction of (3)-(7) above, expressing that the relation $Prec(e, e', i)$ is a linear, discrete order where every non terminal event has a successor. Also, we define the first event as the minimal one with respect to $Prec(e, e', i)$, that is,

$$Fst(e, i) ::= \forall e'(H'ed(e', i) \rightarrow Prec(e, e', i))$$

the first event is provably unique as the order on histories is total. We formally define the specifications MP1-4 as follows:

MP1' $LinDisc(Prec(e, e', i)) \wedge$
$\wedge \exists e(Fst(e, i) \wedge \exists \alpha(e = init(i, \alpha))) \wedge$
$\wedge \forall e(H'ed(e, i) \wedge \neg Fst(e, i) \rightarrow \exists j, \alpha, \mu(e = int(i, \alpha) \vee$
$\vee e = send(i, j, \mu) \vee e = rec(i, j, \mu)))$

MP2' $\forall i, j, \mu(Recd(i, j, \mu) \rightarrow Sent(j, i, \mu))$

MP3' $\langle P \rangle^+([P]\bot \wedge \exists e(H'ed(e, i) \wedge \exists \alpha(e = init(\alpha, i)) \wedge$
$\wedge \forall e'(H'ed(e', i) \rightarrow e' = e))) \wedge$
$\wedge \forall e(H'ed(e, i) \rightarrow (\langle P \rangle H'ed(e, i) \vee$
$\vee (H(e, i) \wedge \forall e'(H(e', i) \rightarrow e' = e))))$

MP4' $H(e, i) \rightarrow ([P]\neg H(e, i) \wedge [F]\neg H(e, i))$

By MP1' the events in the local of agent i are a linear, discrete order, whose first element is an initial event, and whose following events are either send or receive events or internal events. According to MP2' each local state trivially satisfies MP2. By MP3' there is a moment (the starting point) when the only event in an agent's local state is the initial event, and for every event already happened, either it happened at some point strictly in the past, or it is the single event which happened in the last round. Finally, by MP4' each event happens only once in a given run, thus satisfying MP4. MP1'-4' are the basic specifications for MPQIS. We underline that these specifications are defined by means of only the predicative constant H.

As we pointed out above, synchronicity is a costly assumption in terms of computational resources in MPS. This remark prompts us to consider asynchronous MPS, where agents have no common clock. To make this informal definition precise, we follow once more [8]. First, we say that a set V of histories is *prefix closed* if whenever $h \in V$, every non-empty prefix of h is in V as well. Then, we consider the following constraint for AMPQIS:

MP5 The set R of runs in an AMPQIS includes *all* runs satisfying MP1-4 such that the local states of agent i belong to V_i, for some prefix closed set V_i of histories.

This constraint implies that at round m of a run r, each agent i considers possible that any other agent j has performed only a proper subset $r'_j(m)$ of the actions listed in $r_j(m)$.

We can now prove the main result of this section: Proposition 4.4.3 in [8] can be restated as a validity on the class of AMPQIS. We do not provide the full statement here, but we note that this metatheoretical result can be restated as a formula in the first-order modal language $\mathcal{L}_n$. We introduce a relation of *potential causality* between events, as first discussed in [16]. This relation is intended to capture the intuition that event e might have caused event e'. Fix a subset G of A, the relation $\mapsto_G$ holds between events e, e' at a point (r, m) iff both e and e' occur by round m in the run r, and

1. for some $i, j \in G$, e' is a *receive* event and e is the corresponding *send* event, or
2. for some $i \in G$, events e, e' are both in $r_i(m)$ and either $e = e'$ or e comes earlier than e' in $r_i(m)$, or
3. for some e'', we have that $e \mapsto_G e''$ and $e'' \mapsto_G e'$ hold at (r, m).

Note that $\mapsto_G$ is a partial order on events, it is also anti-symmetric by MP4. We can say that two events e, e' are *concurrent* iff $e \not\mapsto_G e'$ and $e' \not\mapsto_G e$. Intuitively, the relation $\mapsto_G$ holds between events e and e' iff it is possible for event e to causally affect event e'. Two events are concurrent if neither can affect the other. We say that $(\mathcal{P}^\sigma, r, m) \models e \mapsto_G e'$ if $e \mapsto_G e'$ holds at (r, m).

Now we prove that the potential causality relation $\mapsto_G$ is the closest we can come in AMPS to an ordering of events, that is, even if the agents in G could combine all their knowledge of the order $Prec(e, e')$ on events, they could not deduce any more about this ordering than is implied by the relation $\mapsto_G$. This is due to the fact that the delivery of messages can be arbitrarily delayed in AMPS, and the agents might be unaware of this because of asynchronicity. We refer to the appendix for a detailed proof.

Lemma 16 *The following validity holds in the class of AMPQIS satisfying the specifications MP1-5 above:*

$$AMPQIS \quad \models \quad \forall e, e'((e \mapsto_G e') \leftrightarrow D_G Prec(e, e'))$$

By virtue of the analysis above we remark that the quantified language we have introduced has the power to express complex specifications, which identify metaproperties about the semantical class under discussion. In particular, by using language $\mathcal{L}_n$ we are able to formalise various constraints on MPS such as reliability, authentication and perfect recall. The traditional propositional specifications MP1-4 for MPS can be given formal counterparts MP1'-4' in $\mathcal{L}_n$, which can be shown valid on the corresponding semantical classes thereby signaling the general correctness of the approach.

5 Conclusions and Future Work

In this paper we analysed a quantified variant of interpreted systems and showed completeness for the axiomatisation QKT.S5$_n$ involving temporal and epistemic modalities on

the first-order language $\mathcal{L}_n$. Retaining completeness seems noteworthy given the known difficulties of these formalisms.

Further, we used this formalism to reason about message passing systems, a mainstream framework to reason about asynchronous systems. In particular, we compared the results obtained at first-order with what was already known at propositional level, and observed that some properties in the latter setting become formal validities in the former.

Still, further work seems to be needed in this line of research. First, it seems interesting to relax the assumption on the domain of quantification, and admit a different domain $D_i(s)$ for each agent i and for each global state s. In such a framework we should check how to modify the completeness proof for QKT.S5$_n$ to accommodate varying domains.

Moreover, we aim at extending the temporal fragment of our language with the *next* $\bigcirc$ and *until* U operators. Completeness results are available for various *monodic* fragments of such a language [31], and for the fragment with $\bigcirc$ over the rational numbers [19]. It is yet to be checked whether these results extend to first-order languages with epistemic operators as well. Also, we would like to analyse relevant classes of QIS, such as *synchronous* QIS and QIS with *perfect recall*. We have sound and complete axiomatisations for these structures at propositional level [8], but it is not clear whether these results extend to first-order.

References

[1] F. Belardinelli and A. Lomuscio. A quantified epistemic logic with flexible terms. Submitted to AAAI07.

[2] F. Belardinelli and A. Lomuscio. A complete quantified epistemic logic for reasoning about message passing systems. In *Proceedings of the 8th International Workshop on Computational Logic in Multi-Agent Systems (CLIMA VIII)*, pages 258–273, 2008.

[3] P. Blackburn, J. van Benthem, and F. Wolter. *Handbook of Modal Logic, volume 53 of Cambridge Tracts in Theoretical Computer Science*. Elsevier, 2007.

[4] A. Chagorov and M Zakharyaschev. *Modal Logic*, volume 35 of *Oxford Logic Guides*. Clarendon Press, Oxford, 1997.

[5] E. M. Clarke, O. Grumberg, and D. A. Peled. *Model Checking*. The MIT Press, Cambridge, Massachusetts, 1999.

[6] P. Cohen and H. Levesque. Communicative actions for artificial agents. In *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS'95)*, 1995.

[7] P. Dembiński, A. Janowska, P. Janowski, W. Penczek, A. Pólrola, M. Szreter, B. Woźna, and A. Zbrzezny. VerICS: weryfikator dla automatów czasowych i specyfikacji zapisanych w języku Estelle. In *Mat. X Konf. Systemy Czasu Rzeczywistego (SCR'03)*, pages 17–26. Instytut Informatyki Politechniki Śląskiej, 2003. In Polish.

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

[9] R. Fagin, J. Y. Halpern, and M. Y. Vardi. What can machines know? On the properties of knowledge in distributed systems. *Journal of the ACM*, 39(2):328–376, 1992.

[10] M. Fitting and R. L. Mendelsohn. *First-order Modal Logic*. Kluwer Academic Publishers, Dordrecht, 1999.

[11] D. M. Gabbay, I. M. Hodkinson, and M. A. Reynolds. *Temporal Logic: Mathematical Foundations and Computational Aspects, Volume 1: Mathematical Foundations*. Oxford University Press, 1993.

[12] P. Gammie and R. van der Meyden. MCK: Model checking the logic of knowledge. In *Proceedings of 16th International Conference on Computer Aided Verification (CAV'04)*, volume 3114 of *LNCS*, pages 479–483. Springer-Verlag, 2004.

[13] J. Y. Halpern and R. Fagin. Modelling knowledge and action in distributed systems. *Distributed Computing*, 3(4):159–179, 1989.

[14] Ian M. Hodkinson, Frank Wolter, and Michael Zakharyaschev. Decidable fragment of first-order temporal logics. *Ann. Pure Appl. Logic*, 106(1-3):85–134, 2000.

[15] G. E. Hughes and M. J. Cresswell. *A New Introduction to Modal Logic*. Routledge, New York, 1996.

[16] Leslie Lamport. Time, clocks, and the ordering of events in a distributed system. *Commun. ACM*, 21(7):558–565, 1978.

[17] A. Lomuscio and M. Colombetti. QLB: a quantified logic for belief. In J. P. Müller, M. J. Wooldridge, and N. R. Jennings, editors, *Intelligent Agents III — Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages (ATAL-96)*, volume 1193 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag, Heidelberg, 1996.

[18] M. Dams M. Cohen. A complete axiomatisation of knowledge and cryptography. In *Logic in Computer Science (LICS)*, 2007.

[19] R. van der Meyden. Axioms for knowledge and time in distributed systems with perfect recall. In *Proceedings, Ninth Annual IEEE Symposium on Logic in Computer Science*, pages 448–457, Paris, France, 1994. IEEE Computer Society Press.

[20] J. J. C. Meyer and W. van der Hoek. Making some issues of implicit knowledge explicit. *International Journal of Foundations of Computer Science*, 3(2):193–223, 1992.

[21] J.-J. Ch. Meyer and W. van der Hoek. *Epistemic Logic for AI and Computer Science*, volume 41 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, 1995.

[22] R. Parikh and R. Ramanujam. Distributed processes and the logic of knowledge. In R. Parikh, editor, *Logic of Programs*, volume 193 of *Lecture Notes in Computer Science*, pages 256–268. Springer, 1985.

[23] F. Raimondi and A. Lomuscio. Automatic verification of multi-agent systems by model checking via OBDDs. *Journal of Applied Logic*, 2005. To appear in Special issue on Logic-based agent verification.

[24] A. Rao and M. Georgeff. Deliberation and its role in the formation of intentions. In B. D. D'Ambrosio, P. Smets, and P. Bonissone, editors, *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence*, pages 300–307, San Mateo, CA, USA, 1991. Morgan Kaufmann Publishers.

[25] M. Reynolds. Axiomatising first-order temporal logic: until and since over linear time. *Studia Logica*, 57(2/3):279–302, 1996.

[26] M. Solanki, A. Cau, and H. Zedan. Asdl: a wide spectrum language for designing web services. In L. Carr, D. De Roure, A. Iyengar, C. A. Goble, and M. Dahlin, editors, *WWW*, pages 687–696. ACM, 2006.

[27] H. Sturm, F. Wolter, and M. Zakharyaschev. Monodic epistemic predicate logic. In M. Ojeda-Aciego, I. P. de Guzmán, G. Brewka, and L. Moniz Pereira, editors, *JELIA*, volume 1919 of *Lecture Notes in Computer Science*, pages 329–344. Springer, 2000.

[28] H. Sturm, F. Wolter, and M. Zakharyaschev. Common knowledge and quantification. *Economic Theory*, 19:157–186, 2002.

[29] F. Viganó. A framework for model checking institutions. In A. Lomuscio S. Edelkamp, editor, *MoChart IV*, volume 4428 of *LNCS*. Springer, 2007.

[30] F. Wolter and M. Zakharyaschev. Decidable fragments of first-order modal logics. *J. Symb. Log.*, 66(3):1415–1438, 2001.

[31] F. Wolter and M. Zakharyaschev. Axiomatizing the monodic fragment of first-order temporal logic. *Ann. Pure Appl. Logic*, 118(1-2):133–145, 2002.

[32] Frank Wolter. First order common knowledge logics. *Studia Logica*, 65(2):249–271, 2000.

[33] M. Wooldridge. *Reasoning about Rational Agents*. MIT Press, 2000.

6 Appendix

Lemma 7 *For every $\phi \in \mathcal{L}_n$, $w \in W$,*

$$(\mathcal{M}^\sigma, w) \models \phi \quad \text{iff} \quad (g(\mathcal{M})^\sigma, r, w) \models \phi$$

Proof. The proof of this lemma is by induction on the length of the formula ϕ. The base of induction for $\phi = P^k(\vec{t})$ or $\phi = (t = t')$ follows by definition of the interpretation I' in $g(\mathcal{M})$. The inductive cases for the propositional connectives are straightforward.

For $\phi = K_i\psi$, $(\mathcal{M}^\sigma, w) \models \phi$ iff for all $w' \sim_i w$, $(\mathcal{M}^\sigma, w') \models \psi$, iff for $r_i(w') = r_i(w)$, $(g(\mathcal{M})^\sigma, r, w) \models \phi$, by definition of r and induction hypothesis, iff $(g(\mathcal{M})^\sigma, r, w) \models \phi$.

The inductive cases for the other modal operators can be shown similarly.

Lemma 11 *For every $\psi \in \mathcal{L}_n^+$,*

$$\mathcal{M}' \models \psi \quad \text{iff} \quad \mathcal{M} \models^p \psi$$

Proof. We first show that if the canonical model $\mathcal{M}$ pseudo-validates $\psi \in \mathcal{L}_n$, then there is a tree-like structure $\mathcal{M}^*$ which pseudo-validates ϕ as well. Then, from $\mathcal{M}^*$ we can obtain a K-model $\mathcal{M}'$ satisfying lemma 11.

In order to define $\mathcal{M}^*$ we need few more definitions. Let w, w' be worlds in W, a *path from w to w'* is a sequence $\langle w_1, l_1, w_2, l_2, \ldots, l_{k-1}, w_k\rangle$ such that (1) $w = w_1$ and $w' = w_k$; (2) $w_1, \ldots, w_k \in W$; (3) each l_j is either an agent or a set of agents; (4) $\langle w_j, w_{j+1}\rangle \in R_{l_j}$.

The *reduction* of a path $\langle w_1, i_1, w_2, i_2, \ldots, i_{k-1}, w_k\rangle$ is obtained by replacing each maximal consecutive subsequence $\langle w_q, i_q, w_{q+1}, i_{q+1}, \ldots, i_{r-1}, w_r\rangle$ where $i_q = i_{q+1} = \ldots = i_{r-1}$ by $\langle w_q, i_q, w_r\rangle$. A path is said to be *reduced* is it is equal to its reduction.

Given the canonical model $\mathcal{M} = \langle W, R, <, D, I\rangle$, we define a structure $\mathcal{M}^* = \langle W^*, R^*, <^*, D, I^*\rangle$ and a surjective function $h : W^* \to W$ such that (i) $\mathcal{M}^*$ is a tree, that is, for $w, w' \in W^*$ there is at most one reduced path from w to w'; (ii) wR_i^*w' implies $h(w)R_ih(w')$; (iii) wR_G^*w' implies $h(w)R_Gh(w')$; (iv) $w <^* w'$ implies $h(w) < h(w')$; (v) $\langle a_1, \ldots, a_k\rangle \in I^*(P^k, w)$ iff $\langle a_1, \ldots, a_k\rangle \in I(P^k, h(w))$.

We define W^* by induction. Let W_1^* be W, and define W_{k+1}^* as the set of worlds $v_{w,l,w'}$ such that $w \in W_k^*$, $w' \in W$ and l is an agent or group of agents. Let $W^* = \bigcup_{k\in\mathbb{N}} W_k^*$, then define $h : W^* \to W$ by letting $h(w) = w$, for $w \in W_1^*$ and $h(v_{w,l,w'}) = w'$, for $w \in W_k^*$. Further, R_l^* is the reflexive, symmetric and transitive closure of the relation defined for $w, w' \in W^*$ if $w' = v_{w,l,w''}$, for some $w'' \in W$, and $h(w)R_lh(w')$; while $<^*$ is the relation defined for $w, w' \in W^*$ if $h(w) < h(w')$. Finally, $I^*(P^k, w) = I(P^k, h(w))$. By results in [9] $\mathcal{M}^*$ and h satisfy (i)-(v) above. In particular, we can show the following:

Proposition 17 *For $w \in W^*$, $\psi \in \mathcal{L}_n^+$,*

$$(\mathcal{M}^{*\sigma}, w) \models^p \psi \quad \text{iff} \quad (\mathcal{M}^\sigma, h(w)) \models^p \psi$$

Finally, we make use of the structure $\mathcal{M}^*$ to define a K-model $\mathcal{M}'$ such that lemma 11 holds. Define $\mathcal{M}' = \langle W', R', <', D', I'\rangle$ as follows:

- $W' = W^*$, $<' = <^*$, $D' = D^*$ and $I' = I^*$;
- R_i' is the transitive closure of $R_i^* \cup \bigcup_{i\in G} R_G^*$.

Since the various R_i^* and R_G^* are reflexive, transitive and symmetric, R_i' is an equivalence relation. We state the following result about $\mathcal{M}'$ and refer to [9] for further details.

Proposition 18 *For $w \in W'$, $\psi \in \mathcal{L}_n^+$,*

$$(\mathcal{M}'^{\sigma}, w) \models \psi \quad \text{iff} \quad (\mathcal{M}^{*\sigma}, w) \models^p \psi$$

In conclusion, The canonical model $\mathcal{M}$ pseudo-validates $\psi \in \mathcal{L}_n$ if and only if $\mathcal{M}^*$ pseudo-validates ψ by proposition 17, iff by proposition 18 the K-model $\mathcal{M}'$ validates ψ.

Lemma 12 *For every $\psi \in \mathcal{L}_n^+$,*

$$\mathcal{M}^+ \models \psi \quad \text{iff} \quad \mathcal{M}' \models \psi$$

Proof. Let $W^{ir} = \{w \in W' \mid w \not<' w\}$ be the set of irreflexive worlds in $\mathcal{M}'$ and define the equivalence relation $\approx$ on $W^r = \{w \in W' \mid w <' w\}$ as $w_1 \approx w_2$ iff $w_1 <' w_2$ and $w_2 <' w_1$. For every $\approx$-equivalence class a, define a map $a()$ from the reals $\mathbb{R}$ onto a such that for every $w \in a, p \in \mathbb{R}$ there are $s, t \in \mathbb{R}$ and

- $s < p < t$;
- $a(s) = w = a(t)$.

This can be done as every $\approx$-equivalence class contains at most $2^{\aleph_0}$ saturated sets of formulas.

Further, for $w \in W^{ir}$ we set $\{w\}(0) = w$. Now we define the K-model $\mathcal{M}^+$, where $W^+ = \{(\{w\}, 0) \mid w \in W^{ir}\} \cup \{(a, p) \mid a$ is a $\approx$-equivalence class, $p \in \mathbb{R}\}$ is the set of possible worlds. The order $<^+$ on W^+ is such that $(a, p) <^+ (b, s)$ iff

- $a \neq b$ and there are $w_a \in a, w_b \in b$ and $w_a <' w_b$; or
- $a = b$ and $p < s$.

The relation $<^+$ is a weakly connected, strict partial order on W^+, in particular $<^+$ is irreflexive. Also, the relation R_i^+ on W^+ such that $(a, p)R_i^+(b, s)$ iff $a(p)R_i'b(s)$ is an equivalence relation as R_i' is such. Finally, the domain D^+ is equal to D', and I^+ is such that $\langle u_1, \ldots, u_k\rangle \in I^+(P^k, (a, p))$ iff $\langle u_1, \ldots, u_k\rangle \in I'(P^k, a(p))$.

It is straightforward to check that $(\mathcal{M}^{+\sigma}, (a, p)) \models \psi$ iff $(\mathcal{M}'^{\sigma}, a(p)) \models \psi$, so the lemma follows.

Playing Cards with Wiebe

[Solving Knowledge Puzzles with "Exactly One" $\mathbf{S5}^n$]

Boris Konev, Clare Dixon, and Michael Fisher

Department of Computer Science, University of Liverpool, UK

1 Introduction

Temporal and modal logics have been widely used in the specification and verification of systems which require dynamic aspects or aspects that deal with knowledge, belief etc [6, 11, 9]. A particular area that Prof. van der Hoek has been closely involved in is that of the representation of knowledge and reasoning using *epistemic* logics [12]. We here particularly target his work, with Ditmarsh and Kooi, on knowledge representation and actions such as public announcements [15, 14]. Specifically, we are interested in the epistemic logic used to represent the knowledge of players in a card game, as described in [13, 14]. In this simple game there are three different cards; one a heart, one a spade and one a club. In the most basic scenario, one card is dealt to one player, a further card is placed face down on the table and the final card is returned (face down) to the card holder.

How should such situations be represented? A very popular approach is to use a logic of *knowledge*, i.e. an *epistemic* logic, in order to represent the knowledge the player has [7]. Thus, in such a logic, we can describe the knowledge a player has about the cards. If we move on to a scenario involving multiple players, then the logical basis naturally extends to *multi-dimensional* logics of knowledge [10] where multiple agents each have an associated notion of knowledge. We can then reason not only about the agent's knowledge of the cards, but also about the agent's knowledge of other agents, the agent's knowledge of other agents' cards, the agent's knowledge of other agents' knowledge about the cards, and so on.

The description of the above scenario given in [14] is very interesting and, although a full logical specification of the card scenario is not given, one can easily imagine what it would look like. A key aspect that is often implicit within the description of the card-game scenario in [14] are the "exactly one" aspects. Thus, a card is in *exactly one* suit, a player holds *exactly one* card, and each card is in *exactly one* place, etc. While these aspects are implicit within the English description, they are explicit within the models shown in [14]. Yet, in order to formalise these aspects within an epistemic logic, typically an **S5** modal logic, quite a few formulae are required.

It is here that we can involve our own work. Over recent years we have been investigating, mechanising, and applying, *temporal* logics with additional constraints of the "exactly one" type considered above [3, 4, 5]. In our work, each logic is parametrised by a set of propositions (or, predicates in the first-order case) where exactly one of these propositions is satisfied at any temporal state. We have shown that, if problems can be described in such a logical framework, then not

only is the description more succinct, but the decision procedure for the logic is simpler (reducing certain aspects of the decision procedure from *exponential* to *polynomial*).

In this paper we will apply the same idea to epistemic logic, defining an "exactly one" variant of **S5** modal logic, and showing improved model size (and hence complexity). In particular, we will provide a tableau-style algorithm for reasoning about such logics which allow "exactly one" (or constrained) sets as input, tackle the card game example from [14], specify it in our constrained epistemic logic, show the algorithm in action to prove certain properties described in [14].

We begin, in Section 2, by providing the syntax and semantics for a standard epistemic logic [11, 7]. Section 3 describes the card playing example from [14] and specifies it using the knowledge logic **S5**. In Section 4 we introduce the constrained logic **SX5**$_n$ showing the complexity lower bound and in Section 5 we provide a tableau algorithm for this logic. We apply the tableau to the card games introduced earlier in Section 6 and provide conclusions in Section 7.

2 Logic of Knowledge

In this section, we give the syntax and semantics of a multi-modal logic of knowledge **S5**$_n$.

2.1 Syntax

The formulae of **S5**$_n$ are constructed using the following connectives and proposition symbols, assuming a set of agents $Ag = \{1, \ldots n\}$:

- a set, PROP, of proposition symbols, p, q, r, ...;
- the constants **false** and **true**;
- the propositional connectives $\neg$, $\vee$, $\wedge$, and $\Rightarrow$; and
- a set of modal connectives K_i (where $i \in Ag$).

The set of (well-formed) formulae of **S5**$_n$ is defined by the following rules:

- any element of PROP is a formula;
- **false** and **true** are formulae;
- if A and B are formulae then so are $\neg A$, $A \vee B$, $A \wedge B$, $A \Rightarrow B$, $K_i A$ where $i \in Ag$.

We define some particular classes of formulae that will be useful later.

Definition 1 *A* literal *is either r, or $\neg r$ where r is a proposition.*

2.2 Semantics

A *model structure*, M, for **S5**$_n$ is a structure $M = \langle S, R_1, \ldots, R_n, \pi \rangle$, where:

- S is a set of states;

- $R_i \subseteq S \times S$, for all $i \in Ag$ is the agent accessibility relation where R_i is an equivalence relation;
- $\pi : S \times \text{PROP} \rightarrow \{\textbf{true}, \textbf{false}\}$ is a valuation

As usual, we define the semantics of the language via the satisfaction relation '$\models$'. This relation holds between pairs of the form $\langle M, s\rangle$ (where M is a model structure and $s \in S$), and **S5**$_n$-formulae. The rules defining the satisfaction relation are given below.

$$
\begin{array}{rll}
\langle M, s\rangle \models \textbf{true} & & \\
\langle M, s\rangle \not\models \textbf{false} & & \\
\langle M, s\rangle \models q & \text{iff} & \pi(s, q) = \textbf{true} \text{ (where } q \in \text{PROP)} \\
\langle M, s\rangle \models \neg\phi & \text{iff} & \langle M, s\rangle \not\models \phi \\
\langle M, s\rangle \models \phi \vee \psi & \text{iff} & \langle M, s\rangle \models \phi \text{ or } \langle M, s\rangle \models \psi \\
\langle M, s\rangle \models \phi \wedge \psi & \text{iff} & \langle M, s\rangle \models \phi \text{ and } \langle M, s\rangle \models \psi \\
\langle M, s\rangle \models \phi \Rightarrow \psi & \text{iff} & \langle M, s\rangle \not\models \phi \text{ or } \langle M, s\rangle \models \psi \\
\langle M, s\rangle \models K_i\phi & \text{iff} & \forall s' \in S' \text{ if } (s, s') \in R_i \text{ then } \langle M, s'\rangle \models \phi
\end{array}
$$

If there is a model structure M and state s such that $\langle M, s\rangle \models \varphi$ then φ is said to be *satisfiable*. If $\langle M, s\rangle \models \varphi$ for all states s and all states s then φ is said to be valid. As the set of modal relations (for each agent i) are equivalence relations they satisfy all the following modal axioms.

$$
\begin{array}{rlrcl}
K: & \vdash & K_i(\phi \Rightarrow \psi) & \Rightarrow & (K_i\phi \Rightarrow K_i\psi) \\
B: & \vdash & \phi & \Rightarrow & K_i\neg K_i\neg\phi \\
T: & \vdash & K_i\phi & \Rightarrow & \phi \\
D: & \vdash & K_i\phi & \Rightarrow & \neg K_i\neg\phi \\
4: & \vdash & K_i\phi & \Rightarrow & K_iK_i\phi \\
5: & \vdash & \neg K_i\neg\phi & \Rightarrow & K_i\neg K_i\neg\phi
\end{array}
$$

Thus, **S5**$_n$ is a multi-modal logic comprising n sub-logics, each of which is itself an **S5** modal logic.

3 Wiebe Playing Cards

We now take the basic single-player card example from [14] and specify it in **S5**$_n$.

Here, an agent (called Wiebe) can hold one of three cards. Each of these cards is in a different suit: hearts, spades, or clubs. The cards are dealt so that Wiebe holds one, one is on the table, and the final one is in a holder (*aka* deck). Following [14] we use simple propositions to represent the position of the cards. So, if $spades_w$ is true, then Wiebe holds a spade, if $clubs_t$ is true, then the clubs card is on the table, if $hearts_h$ is true, then the hearts card is in the holder, etc. Similarly, $K_w spades_w$ means that Wiebe *knows* he holds a spade. And so on.

Now we specify the above problem as follows.

- Wiebe's card is spades or hearts or clubs:

$$(spades_w \vee clubs_w \vee hearts_w) \tag{1}$$

- but Wiebe cannot hold both spades and clubs, both spades and hearts, or both clubs and spades:

$$\neg(spades_w \wedge clubs_w) \wedge \neg(spades_w \wedge hearts_w) \wedge \neg(clubs_w \wedge hearts_w) \tag{2}$$

- The card in the holder is spades or hearts or clubs:

$$(spades_h \vee clubs_h \vee hearts_h) \tag{3}$$

- but the card holder cannot contain both spades and clubs, both spades and hearts, or both clubs and spades:

$$\neg(spades_h \wedge clubs_h) \wedge \neg(spades_h \wedge hearts_h) \wedge \neg(clubs_h \wedge hearts_h) \tag{4}$$

- The card on the table is spades or hearts or clubs:

$$(spades_t \vee clubs_t \vee hearts_t) \tag{5}$$

- but the card on the table cannot be both spades and clubs, both spades and hearts, or both clubs and spades:

$$\neg(spades_t \wedge clubs_t) \wedge \neg(spades_t \wedge hearts_t) \wedge \neg(clubs_t \wedge hearts_t) \tag{6}$$

- The spades card must be either held by Wiebe or be in the holder or be on the table:

$$(spades_w \vee spades_h \vee spades_t) \tag{7}$$

- but cannot be in more than one place:

$$\neg(spades_w \wedge spades_h) \wedge \neg(spades_w \wedge spades_t) \wedge \neg(spades_h \wedge spades_t) \tag{8}$$

- Similarly with the hearts card:

$$\begin{gathered}(hearts_w \vee hearts_h \vee hearts_t) \\ \wedge \\ \neg(hearts_w \wedge hearts_h) \wedge \neg(hearts_w \wedge hearts_t) \wedge \neg(hearts_h \wedge hearts_t)\end{gathered} \tag{9}$$

- And with the clubs card:

$$\begin{gathered}(clubs_w \vee clubs_h \vee clubs_t) \\ \wedge \\ \neg(clubs_w \wedge clubs_h) \wedge \neg(clubs_w \wedge clubs_t) \wedge \neg(clubs_h \wedge clubs_t)\end{gathered} \tag{10}$$

- Wiebe knows all the above statements. For example, formula (3K) is

$$K_w(spades_h \vee clubs_h \vee hearts_h)$$

If, at some point, Wiebe looks at his card and sees it is clubs (but still has not seen the identity of the card in the holder or on the table) then additionally both $clubs_w$ and $K_w clubs_w$.

4 SX5$_n$— "Exactly One" sets in Epistemic Logic

The logic we consider is called "**SX5**$_n$", and its syntax and semantics essentially follow that given above.

The main novelty in **SX5**$_n$ is that it is parametrised by "exactly-one"-sets $\mathcal{P}_1$, $\mathcal{P}_2$,..., and the formulae of TLX($\mathcal{P}_1, \mathcal{P}_2, \ldots$) are constructed under the restrictions that *exactly* one proposition from every set $\mathcal{P}_i$ is true in any state. Furthermore, we assume that there exists a set of propositions in addition to those defined by the parameters, and that these propositions are unconstrained as normal. Thus, **SX5**$_n$() is essentially a standard S5 logic of knowledge while **SX5**$_n$($\mathcal{P}$,$\mathcal{Q}$,$\mathcal{R}$) is a knowledge logic containing at *least* the propositions $\mathcal{P} \cup \mathcal{Q} \cup \mathcal{R}$, where $\mathcal{P} = \{p_1, p_2, \ldots, p_l\}$, $\mathcal{Q} = \{q_1, q_2, \ldots, q_m\}$, and $\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$ and also satisfying the constraint that, for each state, exactly one of $\mathcal{P}$, exactly one of $\mathcal{Q}$, and exactly one of $\mathcal{R}$ holds.

4.1 Complexity

We begin by showing that **SX5**$_n$ has an NP-Hard lower bound.

Lemma 1 *The satisfiability problem for **SX5**$_1$ formulae, even when no variable is unconstrained, is NP-hard.*

Proof We reduce the Boolean satisfiability problem to satisfiability of **SX5**$_1$ formulae. Let ϕ be a Boolean formula over variables x_1,..., x_n. Let $\psi = s \wedge \phi'$, where s is a new proposition and ϕ' is obtained from ϕ by replacing every occurrence of a proposition x_i with the expression K$\neg\overline{x}_i$, where $\overline{x}_i$ is a new proposition, and let $\mathcal{X} = \{s, \overline{x}_1, \ldots, \overline{x}_n\}$ be the 'exactly one' constraint. Notice that the size of ψ is linear in size of ϕ.

For example, if $\phi = x_1 \wedge (\neg x_1 \vee \neg x_2)$ then $\psi = s \wedge \mathsf{K}\neg\overline{x}_1 \wedge (\neg\mathsf{K}\neg\overline{x}_1 \vee \neg\mathsf{K}\neg\overline{x}_2)$ and $\mathcal{X} = \{s, \overline{x}_1, \overline{x}_2\}$. We show that the Boolean formula ϕ is satisfiable if, and only if, the **SX5**$_1$ formula ψ with $\mathcal{X}$ is satisfiable.

Clearly, if ψ is satisfiable, ϕ is satisfiable. Indeed, suppose ψ is true in a Kripke frame $\mathcal{M}$. Then there must exist a world $a \in \mathcal{M}$ such that $(\mathcal{M}, a) \models \psi$. We define an assignment $\mathcal{A}$ as follows: $\mathcal{A}(x_i) =$ **true** if, and only if, $(\mathcal{M}, a) \models \mathsf{K}\neg\overline{x}_i$. Since ϕ' is obtained from ϕ by renaming every occurrence of x_i with K$\neg\overline{x}_i$ we have $\mathcal{A} \models \phi$.

Conversely, we show how to construct a model $\mathcal{M}$ for ψ given $\mathcal{A}$, a satisfying assignment for ϕ. The worlds of the Kripke frame are sets of literals as follows:

$$D = \{\{s, \neg\overline{x}_1, \ldots, \neg\overline{x}_n\}\} \bigcup \{\{\overline{x}_i, \neg s, \cup_{j \neq i} \neg\overline{x}_j\} \mid \mathcal{A}(x_i) = \textbf{false}\}$$

In the example above the only satisfying assignment for ϕ is $\mathcal{A} = \{x_1, \neg x_2\}$ and so $D = \{\{\neg s, \neg\overline{x}_1, \neg\overline{x}_2\}, \{\overline{x}_2, \neg s, \neg\overline{x}_1\}\}$.

The relation R is the full relation on D and $(\mathcal{M}, a) \models p$, for any variable $p \in \{s, \overline{x}_1, \ldots, \overline{x}_n\}$, if, and only if, $p \in a$ (every a is a set of literals). Notice that $(\mathcal{M}, \{s, \neg\overline{x}_1, \ldots, \neg\overline{x}_n\}) \models \mathsf{K}\neg\overline{x}_i$ if, and only if, $\{\overline{x}_i, \neg s, \cup_{j \neq i} \neg x_j\} \notin D$ if, and only if, (by construction of D) $\mathcal{A}(x_i) =$ **true**. Since ϕ' is obtained from ϕ by renaming every occurrence of x_i with K$\neg\overline{x}_i$ we have $(\mathcal{M}, \{s, \neg\overline{x}_1, \ldots, \neg\overline{x}_n\}) \models \psi$. □

Observation 1 *Let ϕ be an **SX5**$_n$ formula such that every proposition in ϕ is constrained. Then satisfiability of ϕ can be decided in polynomial space.*

Proof Since every proposition in ϕ is constrained, there are polynomially many consistent sets of propositions. Thus, the size of a Kripke frame for ϕ is polynomial. □

5 Tableau for SX5$_n$

Next we introduce a tableau algorithm for **SX5**$_n$. Consider an **SX5**$_n$ formula φ to be shown satisfiable. The algorithm constructs sets of *extended assignments* of propositions and modal subformulae i.e. a mapping to true or false, that satisfy both the exactly one sets and φ. However, rather than using the usual alpha and beta rules (see for example the modal tableau in [11, 16]) these are constructed using a DPLL-based expansion [1]. Next the algorithm attempts to satisfy formulae of the form $\neg \mathrm{K}_i\psi$ made true in such an extended assignment by constructing R_i successors which are themselves extended assignments which must satisfy particular subformulae (and the exactly one sets). We begin with some definitions.

Definition 2 *If φ is an **SX5**$_n$ formula, then $sub(\varphi)$ is the set of all subformulae of φ:*

$$sub(\varphi) = \begin{cases} \{\varphi\} & \text{if } \varphi \in \textsc{Prop} \text{ or } \varphi = \textbf{true} \text{ or } \varphi = \textbf{false} \\ \{\neg\psi\} \cup sub(\psi) & \text{if } \varphi = \neg\psi \\ \{\psi * \chi\} \cup sub(\psi) \cup sub(\chi) & \text{if } \varphi = \psi * \chi \text{ where } * \text{ is } \vee, \wedge \text{ or } \Rightarrow \\ \{K_i\psi\} \cup sub(\psi) & \text{if } \varphi = K_i\psi \end{cases}$$

A formula $\psi \in sub(\varphi)$ is a modal subformula *of φ if, and only if, ψ is of the form $K_i\psi'$ for some ψ'.*

Definition 3 *Let φ be an **SX5**$_n$ formula, $\textsc{Prop}(\varphi)$ be the set of all propositions occurring in φ, and $\textsc{Mod}(\varphi)$ be the set of all modal subformulae of φ. We assume, w.l.o.g., that $\mathcal{P}_i \subseteq \textsc{Prop}(\varphi)$ for $i : 1 \leq i \leq n$. An* extended assignment ν for φ *is a mapping from $\textsc{Prop}(\varphi) \cup \textsc{Mod}(\varphi)$ to $\{$**true**, **false**$\}$.*

Every extended assignment ν can be represented by a set of formulae

$$\Delta_\nu = \bigcup_{\substack{p \in \textsc{Prop}(\varphi) \\ \nu(p) = \textbf{true}}} \{p\} \cup \bigcup_{\substack{p \in \textsc{Prop}(\varphi) \\ \nu(p) = \textbf{false}}} \{\neg p\} \cup \bigcup_{\substack{K_i\psi \in \textsc{Mod}(\varphi) \\ \nu(K_i\psi) = \textbf{true}}} \{K_i\psi\} \cup \bigcup_{\substack{K_i\psi \in \textsc{Mod}(\varphi) \\ \nu(K_i\psi) = \textbf{false}}} \{\neg K_i\psi\}$$

*Let ψ be an **SX5**$_n$ formula such that $\textsc{Prop}(\psi) \subseteq \textsc{Prop}(\varphi)$ and $\textsc{Mod}(\psi) \subseteq \textsc{Mod}(\varphi)$. An extended assignment ν for φ is* compatible *with ψ if, and only if, the following conditions hold.*

- *For every set $\mathcal{P}_i$, there exists exactly one proposition $p \in \mathcal{P}_i$ such that $\nu(p) =$ **true** (and so $\nu(q) =$ **false** for all $q \in \mathcal{P}_i, q \neq p$).*
- *The result of replacing every occurrence of a proposition $p \in \textsc{Prop}(\psi)$ in ψ with $\nu(p)$ and every occurrence of a modal subformula $K_i\psi' \in \textsc{Mod}(\psi)$, such that $K_i\psi'$ is not in the scope of another modal operator in ψ, with $\nu(K_i\psi')$ evaluates to **true**.*
- *If $\nu(K_j\chi) =$ **true**, for some modal subformula χ of ψ, then ν is compatible with χ.*

We denote $\mathcal{N}(\varphi)$ the set of all extended assignments of φ.

Example 2 *Let $\mathcal{P}_1 = \{p, q\}$ and $\varphi = \neg K_1(p \wedge K_2 \neg p)$. Suppose ψ is φ itself. Consider the extended assignment ν_1 represented by the set $\Delta_{\nu_1} = \{p, \neg q, \neg K_1(p \wedge K_2 \neg p), K_2 \neg p\}$. Then the first two conditions of compatibility with ψ hold true. Notice, however, that ν_1 is not compatible with ψ since $\nu_1(K_2 \neg p) =$ **true** but $\neg p$ evaluates to **false** under ν_1. The extended assignment ν_2 represented by the set $\Delta_{\nu_2} = \{p, \neg q, \neg K_1(p \wedge K_2 \neg p), \neg K_2 \neg p\}$ is compatible with ψ. Also the extended assignment ν_3 represented by the set $\Delta_{\nu_3} = \{\neg p, q, \neg K_1(p \wedge K_2 \neg p), K_2 \neg p\}$ is compatible with ψ.*

Lemma 3 *Let φ be an **SX5**$_n$ formula and ψ its subformula. Then the set of all extended assignments for φ compatible with ψ can be computed in $O\left(|\mathcal{P}_1| \times \ldots \times |\mathcal{P}_n| \times 2^{|A|} \times 2^k\right)$ time, where $|\mathcal{P}_i|$ is the size of the set $\mathcal{P}_i$ of constrained propositions, $|A|$ is the size of the set A of non-constrained propositions, and k is the number of K_i operators in φ.*

Proof The set of all extended assignments compatible with ψ can be constructed by the DPLL algorithm, where we first split on elements of $\mathcal{P}_i$ (that requires $O(|\mathcal{P}_1| \times \ldots \times |\mathcal{P}_n|)$ time) and then on elements of A and $\textsc{Mod}(\varphi)$. □

We now move on to the model-like structures that will be generated by the tableau algorithm. Note that $State = \{s, s', \ldots\}$ is the set of all states.

Definition 4 *Let φ be an **SX5**$_n$ formula. A* structure*, H, is a tuple $H = (S, R_1, \ldots, R_n, L)$, where:*

- $S \subseteq State$ *is a set of states;*
- $R_i \subseteq S \times S$ *represents an accessibility relation over S for agent $i \in Ag$;*
- $L : S \to \mathcal{N}(\varphi)$ *labels each state with an extended assignment for φ.*

The tableau algorithm first expands the structure and then contracts it. We try to construct a structure from which a model may possibly be extracted, and then delete states in this structure that are labelled with formulae such as $\neg K_i p$ which are not satisfied in the structure. Expansion uses the formulae in the labels of each state to build R_i successors.

Given the **SX5**$_n$ formula φ to be shown unsatisfiable, perform the following steps.

1. *Initialisation*.

   First, set
   $$S = R_1 = \cdots = R_n = L = \emptyset.$$
   Construct $\mathcal{F}$, the set of all extended assignments for φ compatible with φ. For each $\nu_i \in \mathcal{F}$ create a new state s_i and let $L(s_i) = \nu_i$ and $S = S \cup \{s_i\}$.

2. *Creating R_i successors*.

   For any state s labelled by an extended assignment ν, i.e. $L(s) = \nu$ for each formula of the form $\neg K_i \psi \in \Delta_{L(s)}$ create a formula
   $$\psi' = \neg\psi \wedge \bigwedge_{K_i\chi \in \Delta_{L(s)}} \chi \wedge \bigwedge_{K_i\chi \in \Delta_{L(s)}} K_i\chi \wedge \bigwedge_{\neg K_i\chi \in \Delta_{L(s)}} \neg K_i\chi$$
   For each ψ' above construct $\mathcal{F}$, the set of all extended assignments for φ compatible with ψ' and for each member $\nu \in \mathcal{F}$ if there exists a state $s'' \in S$ such that $\nu = L(s'')$ then add (s, s'') to R_i, otherwise add a new state s' to S, labelled by $L(s') = \nu$, and add (s, s') to R_i.

3. *Contraction*.

   Continue deleting any state s where there exists a formula $\psi \in \Delta_{L(s)}$ such that ψ is of the form $\neg K_i\chi$ and there is no state $s' \in S$ such that $(s, s') \in R_i$ and $L(s')$ is compatible with $\neg\chi$;

   until no further deletions are possible.

If φ is a formula then we say the tableau algorithm is *successful* if, and only if, the structure returned contains a state s such that φ is compatible with $L(s)$. We claim that a formula φ is **S5**$_n$ satisfiable if, and only if, the tableau algorithm performed on φ is successful.

Theorem 4 *Let $\mathcal{P}_1$,..., $\mathcal{P}_n$ be sets of constrained propositions, and φ be an **SX5**$_n(\mathcal{P}_1, \ldots, \mathcal{P}_n)$ formula such that $\bigcup_{i=1}^{n} \mathcal{P}_i \subseteq \textsf{PROP}(\varphi)$. Then*

- *φ is satisfiable if, and only if, the tableau algorithm applied to φ returns a structure $(S, \eta, R_1, \ldots, R_n, L)$ in which there exists a state $s \in S$ such that φ is compatible with $L(s)$.*
- *The tableau algorithm runs in time polynomial in $\left(k \times |\mathcal{P}_1| \times \ldots \times |\mathcal{P}_n| \times 2^{|A|+k}\right)$, where $|\mathcal{P}_i|$ is the size of the set $\mathcal{P}_i$ of constrained propositions, $|A|$ is the size of the set A of non-constrained propositions, and k is the number of K_i operators in φ.*

Proof The correctness and completeness of the tableau algorithm can be proved by adapting the correctness and completeness proof given in [16]. The only difference between the two algorithms is that the algorithm in [16] applies propositional tableau expansion rules to formulae and the one given above uses DPLL-based expansion.

For the second part of the theorem, notice that the number of nodes in any structure does not exceed $\left(|\mathcal{P}_1| \times \ldots \times |\mathcal{P}_n| \times 2^{|A|+k}\right)$. When creating R_i successors, we consider at most k formulae of the form $\neg K_i\psi \in \Delta_\nu$, and, by Lemma 3, the set of all extended assignments for φ compatible with ψ' can be computed in $O\left(|\mathcal{P}_1| \times \ldots \times |\mathcal{P}_n| \times 2^{|A|+k}\right)$ time. Building the structure and applying the contraction rule can be implemented in time polynomial in the structure size. □

6 Modelling Wiebe's Card Game using SX5$_n$

Here we demonstrate the tableau algorithm applied to the card playing scenario described previously.

6.1 A Single Agent Game

Next we consider the card game described above. We can model this with six exactly one sets **SX5**$_n(\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \mathcal{P}_4, \mathcal{P}_5, \mathcal{P}_6)$ where

- $\mathcal{P}_1 = \{spades_w, clubs_w, hearts_w\}$ — Wiebe has exactly one card.
- $\mathcal{P}_2 = \{spades_h, clubs_h, hearts_h\}$ — exactly one card is in the holder.
- $\mathcal{P}_3 = \{spades_t, clubs_t, hearts_t\}$ — exactly one card is on the table.
- $\mathcal{P}_4 = \{spades_w, spades_h, spades_t\}$ — the spades card is in exactly one place.
- $\mathcal{P}_5 = \{clubs_w, clubs_h, clubs_t\}$ — the clubs card is in exactly one place.

- $\mathcal{P}_6 = \{hearts_w, hearts_h, hearts_t\}$ — the hearts card is in exactly one place.

We assume that initially Wiebe holds the clubs card and he knows this ($clubs_w$ and $K_w clubs$). Let us prove that Wiebe knows that either hearts is in the holder or spades is in the holder (i.e. $K_w(hearts_h \vee spades_h)$).

We try to apply the tableau algorithm to $(clubs_w \wedge K_w clubs_w) \Rightarrow K_w(hearts_h \vee spades_h)$. Negating to give $\varphi = \neg((clubs_w \wedge K_w clubs_w) \Rightarrow K_w(hearts_h \vee spades_h))$ we obtain two extended assignments ν_0 and ν_1 for φ compatible with φ. We construct two sets s_0 and s_1 such that $L(s_0) = \nu_0$ and $L(s_1) = \nu_1$. The following are the set of formulae $\Delta_{L(s_0)}$ and $\Delta_{L(s_1)}$ associated with these. To save space we assume that if propositions are not listed in the sets they are false.

$$\Delta_{L(s_0)} = \{clubs_w, K_w clubs_w, \neg K_w(hearts_h \vee spades_h), spades_h, hearts_t\}$$
$$\Delta_{L(s_1)} = \{clubs_w, K_w clubs_w, \neg K_w(hearts_h \vee spades_h), hearts_h, spades_t\}$$

Next we try to create an R_w successor to satisfy $\neg K_w(hearts_h \vee spades_h)$. That is we try to construct the set of extended assignments for φ compatible with ψ where

$$\psi = \neg(hearts_h \vee spades_h) \wedge clubs_w \wedge K_w clubs_w \wedge \neg K_w(hearts_h \vee spades_h)$$

The set of extended assignments for φ compatible with ψ is empty. Essentially any assignment ν compatible with ψ must have $\nu(hearts_h) =$ **false** and $\nu(spades_h) =$ **false** and $\nu(clubs_w) =$ **true**. However, given the first two of these, the exactly one set $\mathcal{P}_2$ forces $\nu(clubs_h) =$ **true** which contradicts with $\nu(clubs_w) =$ **true** and the exactly one set $\mathcal{P}_5$.

So we can construct no R_w successors to either s_0 or s_1. Both $\Delta_{L(s_0)}$ and $\Delta_{L(s_1)}$ contain $\neg K_w(hearts_h \vee spades_h)$ and there is no state s_j such that $(s0, s_j) \in R_w$ or $(s1, s_j) \in R_w$ such that $L(s_j)$ is compatible with $\neg(hearts_h \vee spades_h)$. Hence s_0 and s_1 are deleted and the tableau algorithm is unsuccessful so φ is unsatisfiable and the original formula is valid.

6.2 A Multi-Agent Game

Next we extend the previous example allowing more than one agent. As before we have the agent Wiebe and introduce a new agent called Marta. Wiebe and Marta each hold one of three cards: hearts, spades, or clubs. The cards are dealt so that Wiebe holds one, Marta holds one and one is on the table. We assume that Wiebe holds the clubs card, Marta holds the hearts card and they both have both looked at their cards so Wiebe knows he holds clubs and Marta knows she holds hearts. Modelling this scenario we replace the propositions $hearts_h$, $spades_h$, and $clubs_h$ by $hearts_m$, $spades_m$, and $clubs_m$ denoting that Marta holds the hearts, spades or clubs card respectively. We have the same exactly one sets $\mathcal{P}_1 - \mathcal{P}_6$ as previously but with the propositions subscripted by h replaced by those subscripted by m. We try to prove that Wiebe knows that Marta considers that it is possible that Wiebe holds the clubs card, i.e. $K_w \neg K_m \neg clubs_w$.

Thus we attempt to prove $(K_w clubs_w \wedge clubs_w \wedge K_m hearts_m \wedge hearts_m) \Rightarrow K_w \neg K_m \neg clubs_w$. As before we negate and apply the tableau algorithm to

$$\varphi = \neg((K_w clubs_w \wedge clubs_w \wedge K_m hearts_m \wedge hearts_m) \Rightarrow K_w \neg K_m \neg clubs_w)$$

We first construct the set of extended assignments of φ compatible with φ. We obtain one set ν_0 so we construct a state s_0 such that $L(s_0) = \nu_0$. The set of formulae representing this extended assignment is:

$$\Delta_{L(s_0)} = \{K_w clubs_w, clubs_w, K_m hearts_m, hearts_m, \neg K_w \neg K_m \neg clubs_w, \neg K_m \neg clubs_w, spades_t\}$$

As $\Delta_{L(s_0)}$ contains $\neg K_w \neg K_m \neg clubs_w$ we must try to construct an extended assignment, ν, for φ which is compatible with ψ where

$$\psi = \neg\neg K_m \neg clubs_w \wedge clubs_w \wedge K_w clubs_w \wedge \neg K_w \neg K_m \neg clubs_w$$

There are no such assignments. From the first conjunct any extended assignment, ν, must make $\nu(K_m \neg clubs_w) =$ **true** and therefore must make $\nu(clubs_w) =$ **false**. However from the second conjunct the assignment must make $\nu(clubs_w) =$ **true** which gives a contradiction.

As $\Delta_{L(s_0)}$ also contains $\neg K_m \neg clubs_w$ we would next continue by trying to construct the extended assignments for φ which are compatible with ψ' where

$$\psi' = \neg\neg clubs_w \wedge hearts_m \wedge K_m hearts_m \wedge \neg K_m \neg clubs_w.$$

However, when we come to the deletion phase as $\Delta_{L(s_0)}$ contains $\neg K_w \neg K_m \neg clubs_w$ and there is no state s_j such that $(s_0, s_j) \in R_w$ such that $L(s_j)$ is compatible with $\neg\neg K_m \neg clubs_w$. and we would delete s_0. Hence the tableau is unsuccessful so φ must be unsatisfiable and the original formula valid.

Remark In the above two example to make sure φ contains all the propositions from the constrained set, we can conjoin ϕ with a tautology $(p \vee \neg p)$ for every proposition p such that $p \in \mathcal{P}_i$, for some i, but p does not occur to φ. We omit such tautological conjuncts for presentation purposes.

7 Conclusions and Related Work

We have taken recent ideas relating to temporal logics which allow the input of sets of proposition where exactly one from each set must hold and have adapted them to the framework of multi-modal logics of knowledge. We have motivated the need for such constraints by considering a particular card game discussed in [13, 14]. We have provided a tableau based algorithm to prove **SX5**$_n$ formulae which replaces the usual alpha and beta rules with a DPLL-based expansion. We show the lower bound for such logics is NP-Hard and provide a complexity analyisis for the developed tableau algorithm. This shows that the tableau is useful when applied to problems with a large number of constrained propositions and a comparatively low number of unconstrained propositions and modal operators in the formula to be proved.

We can also see, at least for one particular example, how much more succinct the new logic is than the standard. If we look at the difference between the set of formulae in Section 3 and the set required in Section 6, we see significant improvement in succinctness. For other, similar, problems we can expect equally impressive results.

For future work in this area we will consider two aspects.

1. The extension of **SX5**$_n$ with temporal aspects, providing an "exactly one" analogue of standard temporal logic of knowledge [7]. We believe this will combine the benefits of the **SX5**$_n$ logic developed here, with the "exactly one" temporal logic considered in [4] and will therefore provide a more succinct and efficient framework in which to describe and verify scenarios concerning dynamic knowledge.

2. Just as Prof. van der Hoek, et al, tackled the evolution of epistemic scenarios with *public announcements* through the medium of *Dynamic Epistemic Logic* [13, 14], so we aim to provide a more succinct version (using (1) above) capturing public announcements in an "exactly one" temporal logic of knowledge.

References

[1] M. Davis, G. Logemann, and D. Loveland. A Machine Program for Theorem-Proving. *Communications of the ACM*, 5(7):394–397, 1962.

[2] A. Degtyarev, M. Fisher, and B. Konev. A Simplified Clausal Resolution Procedure for Propositional Linear-Time Temporal Logic. In *Automated Reasoning with Analytic Tableaux and Related Methods*, volume 2381 of *Lecture Notes in Computer Science*, pages 85–99. Springer-Verlag, July 30–Aug. 1 2002.

[3] C. Dixon, M. Fisher, and B. Konev. Temporal Logic with Capacity Constraints. In *Proc. 6th International Symposium on Frontiers of Combining Systems*, volume 4720 of *Lecture Notes in Computer Science*, pages 163–177. Springer, 2007.

[4] C. Dixon, M. Fisher, and B. Konev. Tractable Temporal Reasoning. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, pages 318–323. AAAI Press, 2007.

[5] C. Dixon, M. Fisher, B. Konev, and A. Lisitsa. Practical First-Order Temporal Reasoning. In *Proc. 15th International Symposium on Temporal Representation and Reasoning (TIME)*, pages 156–163. IEEE Computer Society, 2008.

[6] E. A. Emerson. Temporal and Modal Logic. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science*, pages 996–1072. Elsevier, 1990.

[7] R. Fagin, J. Halpern, Y. Moses, and M. Vardi. *Reasoning About Knowledge*. MIT Press, 1996.

[8] M. Fisher, C. Dixon, and M. Peim. Clausal Temporal Resolution. *ACM Transactions on Computational Logic*, 2(1):12–56, Jan. 2001.

[9] M. Fisher, D. Gabbay, and L. Vila, editors. *Handbook of Temporal Reasoning in Artificial Intelligence*, volume 1 of *Advances in Artificial Intelligence*. Elsevier, 2005.

[10] D. Gabbay, A. Kurucz, F. Wolter, and M. Zakharyaschev. *Many-Dimensional Modal Logics: Theory and Applications*. Number 148 in Studies in Logic and the Foundations of Mathematics. Elsevier, 2003.

[11] J. Y. Halpern and Y. Moses. A Guide to Completeness and Complexity for Modal Logics of Knowledge and Belief. *Artificial Intelligence*, 54:319–379, 1992.

[12] J. Meyer and W. van der Hoek. *Epistemic Logic for Computer Science and Artificial Intelligence*, volume 41 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, 1995.

[13] H. van Ditmarsch, W. van der Hoek, and B. Kooi. Playing cards with Hintikka. An Introduction to Dynamic Epistemic Logic. *Phi-News; the newsletter for philosophical logic and its applications*, 6:6–32, October 2004.

[14] H. van Ditmarsch, W. van der Hoek, and B. Kooi. Playing Cards with Hintikka — An Introduction to Dynamic Epistemic Logic. *Australasian Journal of Logic*, 3:108–134, 2005. `http://www.philosophy.unimelb.edu.au/ajl/2005/2005_8.pdf`.

[15] H. P. van Ditmarsch, W. van der Hoek, and B. P. Kooi. Public Announcements and Belief Expansion. In *Proc. 5th International Conference on Advances in Modal Logic (AiML)*, pages 335–346. King's College Publications, 2005.

[16] M. Wooldridge, C. Dixon, and M. Fisher. A Tableau-Based Proof Method for Temporal Logics of Knowledge and Belief. *Journal of Applied Non-Classical Logics*, 8(3):225–258, 1998.

Rational Play and Rational Beliefs under Uncertainty

Nils Bulling and Wojciech Jamroga
Clausthal University of Technology, Germany
{bulling,wjamroga}@in.tu-clausthal.de

Abstract

Alternating-time temporal logic (ATL) is one of the most influential logics for reasoning about agents' abilities. Constructive Strategic Logic (CSL) is a variant of ATL for imperfect information games that allows to express strategic and epistemic properties of coalitions under uncertainty. In this paper, we propose a logic that extends CSL with a notion of plausibility that can be used for reasoning about the outcome of rational behavior (in the game-theoretical sense). Moreover, we show how a particular notion of beliefs can be defined on top of plausibility. The resulting logic, CSLP, turns out to be very expressive.

We show that beliefs satisfy axioms **KD45** in the logic. We also demonstrate how solution concepts for imperfect information games can be characterized and used in CSLP and that the model checking complexity increases only slightly when plausibility and rational beliefs are added.

Keywords: temporal logic, imperfect information games, knowledge and beliefs.

*This is a slightly revised version of a paper which will appear in the proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems **AAMAS'09**, to be held on May 10–15, 2009 in Budapest, Hungary.*

1 Introduction

Alternating-time temporal logic (ATL) [1, 2] is one of the most influential logics for reasoning about abilities of agents with perfect information. The key constructs are *cooperation modalities* $\langle\langle A\rangle\rangle$ where A is a group of agents. The reading of $\langle\langle A\rangle\rangle\gamma$ is that agents A have a collective strategy to enforce γ. In [9] a variant of ATL for imperfect information scenarios has been proposed. The logic, called *Constructive Strategic Logic* (CSL), unified several attempts to incorporate epistemic concepts into ATL, and solved various problems of these previous attempts. However, it included only strategic and epistemic modalities; in particular, doxastic and rationality concepts were absent.

On the other hand, another extension of ATL, called ATL *with Plausibility* (ATLP), has been proposed in [10, 4] for reasoning about *rational* or *plausible* behavior.[1] That logic allowed to describe and/or impose rationality assumptions on (a subset of) agents, and to reason about the outcome of the play if irrational behavior was disregarded. For example, one might assume that agents are not completely dumb and do not play dominated strategies (in the game-theoretical sense). Such assumptions allow to restrict the vast number of possibilities each agent has to consider.

In this paper we present *Constructive Strategic Logic with Plausibility* (CSLP), a combination of CSL and ATLP where the new language goes far beyond the pure union of both logics. Firstly, the plausibility concept allows us to neatly define the relationship between epistemic and doxastic concepts, in a similar way to [3]. As the basic modality we introduce *weak constructive rational beliefs*: $\mathbb{CW}_A$ (common beliefs), $\mathbb{DW}_A$ (distributed beliefs), and $\mathbb{EW}_A$ (mutual beliefs). The term *constructive* is used in the same sense as in [10, 4], where it referred to an "operational" kind of knowledge that, in order to "know how to play", requires the agents to be able to *identify* and *execute* an appropriate strategy. Like for CSL, the semantics of CSLP is non-standard: formulae are interpreted in *sets of states*. For example, the intuitive reading of $M, Q \models \langle\langle A\rangle\rangle\gamma$ is that agents A have a collective strategy which enforces γ from *each* state in Q. Thanks to the plausibility concept provided by ATLP we can define *knowledge* and *rational beliefs* on top of weak beliefs. We point out that our notion of rational belief is rather specific, and show interesting properties of knowledge, rational belief, and plausibility. In particular, it is shown that knowledge and belief are **KD45** modalities.

We show that CSLP is very expressive, and we demonstrate how solution concepts for imperfect information games can be characterized and used in CSLP. It also turns out that, despite the logic's expressiveness, the model checking complexity does not increase when compared to ATLP, and increases only slightly compared to CSL when plausibility and rational beliefs are added.

1.1 Related Work

Our idea to build beliefs on top of plausibility has been inspired by [14, 8]. In [3], we extended CTLK [13], a straightforward combination of the branching-time logic CTL [6] and standard epistemic logic [7], by a notion of plausibility which in turn was used to define a particular notion of beliefs. Plausibility assumptions were defined in terms of paths in the underlying system. Then, agent's beliefs were given by his knowledge if only plausible paths were considered.

Another source of inspiration is [17, 16], where the semantics of ability was influenced by particular notions of rationality. We generalized these ideas in [10, 4]. Semantically, a subset of strategies (behaviors) was identified as *rational* in the model; a typical formula was $\mathbf{Pl}_B\langle\langle A\rangle\rangle\gamma$ with the following reading: Agents A can

[1] In this paper we use the terms *rational* and *plausible* interchangeably.

enforce γ if agents in B act rationally. We showed how one can use the logic to characterize solution concepts (Nash equilibria, Pareto optimal profiles etc.), and reason about the outcome of rational play.

The current paper is an attempt to integrate the notions of time, knowledge, belief, strategic ability, rationality, and uncertainty in a single logical framework.

2 CSL with Plausibility

We start with an informal presentation of the idea. Then, we describe the formal syntax and semantics, and we discuss the new operators in more detail.

2.1 Agents, Beliefs, and Rational Play

In the following, let $A \subseteq \mathbb{A}\mathrm{gt}$ be a team of agents where $\mathbb{A}\mathrm{gt}$ denotes the set of all agents. Formulae are interpreted given a model M and a set of states Q. The reading of $M, Q \models \langle\langle A \rangle\rangle \gamma$ is that agents A have a collective strategy which enforces γ from *all* states in Q. $\mathbf{Pl}_A\, \varphi$ assumes that agents in A play plausibly according to some rationality criterion which can be set (resp. refined) by operators (**set-pl** ω) (resp. (**refn-pl** ω)). The set of such *rational agents* is denoted by $\mathbb{R}\mathrm{gt}$. Plausibility terms ω refer to sets of strategy profiles that implement the rationality criteria. Finally, the logic includes operators for *constructive weakly rational belief* (*constructive weak belief*/cwb in short): $\mathbb{CW}_A\varphi$ (agents A have common cwb in φ); $\mathbb{EW}_A\varphi$ (agents A have mutual cwb in φ); and $\mathbb{DW}_A\varphi$ (agents A have distributed cwb in φ). Semantically, the cwb operators yield "epistemic positions" of team A that serve as reference for the semantic evaluation of strategic formulae.

Consider formula $\mathbb{EW}_A\mathbf{Pl}_{\mathbb{A}\mathrm{gt}\setminus A}\langle\langle A \rangle\rangle \Box \mathsf{safe}$ (*coalition A has a constructive mutual weak belief that they can keep the system safe forever if the opponents behave rationally*) in model M and set of states Q. Firstly, Q is extended with all states indistinguishable from some state in Q for any agent from A. Let us call the extended set Q'. Now, A have cwb in $\mathbf{Pl}_{\mathbb{A}\mathrm{gt}\setminus A}\langle\langle A \rangle\rangle \Box \mathsf{safe}$ iff they have a strategy that maintains safe from all states in Q' assuming that implausible behavior for the agents in $\mathbb{A}\mathrm{gt} \setminus A$ is disregarded.

Later, we will define strongly rational beliefs (resp. knowledge) as a special case of cwb's in which all agents are (resp. no agent is) assumed to play plausibly.

2.2 Syntax

The language of *Constructive Strategic Logic with Plausibility* (CSLP) includes atomic propositions, Boolean connectives, strategic formulae, operators for *constructive weakly rational beliefs*, and operators that handle *plausibility updates*. As we will see, standard/constructive strongly rational beliefs and knowledge can be defined on top of these.

Definition 1 ($\mathcal{L}_{\mathbf{CSLP}}$) *Let* $\mathbb{A}\mathrm{gt}$ *be a set of agents,* Π *a set of propositions, and* Ω *a set of primitive plausibility terms. The logic* $\mathcal{L}_{CSLP}(\mathbb{A}\mathrm{gt}, \Pi, \Omega)$ *is generated by the following grammar:*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \langle\langle A\rangle\rangle \bigcirc \varphi \mid \langle\langle A\rangle\rangle \Box \varphi \mid \langle\langle A\rangle\rangle \varphi \,\mathcal{U}\, \varphi \mid \mathbb{CW}_A\varphi \mid \mathbb{EW}_A\varphi \mid \mathbb{DW}_A\varphi \mid \mathbf{Pl}_A\,\varphi \mid (\textbf{\textit{set-pl}}\ \omega)\varphi \mid (\textbf{\textit{refn-pl}}\ \omega)\varphi.$$

The temporal operators $\bigcirc, \Box, \mathcal{U}$ stand for "next", "always", and "until", respectively. We use the standard definitions of derived Boolean connectives $\vee$, $\rightarrow, \leftrightarrow$, plus the following derived modalities: $\Diamond\varphi \equiv \top\,\mathcal{U}\,\varphi$ (sometime), $\mathsf{Now}\varphi \equiv \varphi\,\mathcal{U}\,\varphi$ (now), $\mathbb{W}_a\varphi \equiv \mathbb{CW}_{\{a\}}\varphi$ (individual cwb), $\mathsf{CW}_A\varphi \equiv \mathbb{CW}_A\langle\langle\emptyset\rangle\rangle\mathsf{Now}\varphi$, $\mathsf{EW}_A\varphi \equiv \mathbb{EW}_A\langle\langle\emptyset\rangle\rangle\mathsf{Now}\varphi$, $\mathsf{DW}_A\varphi \equiv \mathbb{DW}_A\langle\langle\emptyset\rangle\rangle\mathsf{Now}\varphi$ (standard weak belief, wb), $\mathsf{W}_a\varphi \equiv \mathsf{CW}_{\{a\}}\varphi$ (individual wb), $\mathbf{Pl} \equiv \mathbf{Pl}_{\mathbb{A}\mathrm{gt}}$ (reasoning under the assumption that all agents behave plausibly), and $\mathbf{Ph} \equiv \mathbf{Pl}_{\emptyset}$ (reasoning about outcome of all "physically" possible behaviors). Finally, we define operators for constructive and standard strongly rational belief (csb) as:

$\mathbb{Bel}_a \equiv \mathbb{W}_a\mathbf{Pl}$, $\quad \mathbb{CBel}_A \equiv \mathbb{CW}_A\mathbf{Pl}$,
$\mathbb{EBel}_A \equiv \mathbb{EW}_A\mathbf{Pl}$, $\quad \mathbb{DBel}_A \equiv \mathbb{DW}_A\mathbf{Pl}$,
$\mathsf{Bel}_a \equiv \mathbf{Ph}\,\mathsf{W}_a\mathbf{Pl}$, $\quad \mathsf{CBel}_A \equiv \mathbf{Ph}\,\mathsf{CW}_A\mathbf{Pl}$,
$\mathsf{EBel}_A \equiv \mathbf{Ph}\,\mathsf{EW}_A\mathbf{Pl}$, $\quad \mathsf{DBel}_A \equiv \mathbf{Ph}\,\mathsf{DW}_A\mathbf{Pl}$,

and the constructive and standard knowledge operators as:

$\mathbb{K}_a \equiv \mathbf{Ph}\,\mathbb{W}_a$, $\quad \mathbb{C}_A \equiv \mathbf{Ph}\,\mathbb{CW}_A$, $\quad \mathbb{E}_A \equiv \mathbf{Ph}\,\mathbb{EW}_A$,
$\mathbb{D}_A \equiv \mathbf{Ph}\,\mathbb{DW}_A$, $\quad \mathsf{K}_a \equiv \mathbf{Ph}\,\mathsf{W}_a$, $\quad \mathsf{C}_A \equiv \mathbf{Ph}\,\mathsf{CW}_A$,
$\mathsf{E}_A \equiv \mathbf{Ph}\,\mathsf{EW}_A$, $\quad \mathsf{D}_A \equiv \mathbf{Ph}\,\mathsf{DW}_A$.

We will show in Section 2.4 that these definitions capture the respective notions of knowledge and belief appropriately.

2.3 Semantics

Firstly, we introduce the basic models of time, action, and knowledge.

Definition 2 (CEGS) *A* concurrent epistemic game structure (CEGS) *is a tuple* $M = \langle \mathbb{A}\mathrm{gt}, St, \Pi, \pi, Act, d, o, \sim_1, \ldots, \sim_k\rangle$*, with: a nonempty finite set of all agents* $\mathbb{A}\mathrm{gt} = \{1, ..., k\}$*, a nonempty set of states* St*, a set of atomic propositions* Π*, a valuation of propositions* $\pi : St \rightarrow 2^{\Pi}$*, and a nonempty finite set of atomic actions* Act*.* $\sim_1, ..., \sim_k \subseteq St \times St$ *are epistemic equivalence relations;* $q \sim_a q'$ *means that, while the system is in state* q*, agent* a *cannot determine whether it is in* q *or* q'*. Function* $d : \mathbb{A}\mathrm{gt} \times St \rightarrow 2^{Act}$ *defines nonempty sets of actions available to agents at each state, with* $d(a, q) = d(a, q')$ *for* $q \sim_a q'$*. Finally,* o *is a (deterministic) transition function that assigns the outcome state* $q' = o(q, \alpha_1, \ldots, \alpha_k)$ *to state* q *and a tuple of actions* $\langle\alpha_1, \ldots, \alpha_k\rangle$*,* $\alpha_i \in d(i, q)$*, that can be executed by* $\mathbb{A}\mathrm{gt}$ *in* q*.*

Remark 1 *Relations* $\sim_A^E$*,* $\sim_A^C$ *and* $\sim_A^D$*, used to model group epistemics, are derived from the individual relations of agents from* A*. First,* $\sim_A^E$ *is the union of*

relations $\sim_a$, $a \in A$. Next, $\sim_A^C$ is defined as the transitive closure of $\sim_A^E$. Finally, $\sim_A^D$ is the intersection of all the $\sim_a$, $a \in A$.

A *strategy* s_a of agent a is a conditional plan that specifies what a is going to do for every possible situation: $s_a : St \to Act$ such that $s_a(q) \in d(a, q)$. A *collective strategy* s_A for a group of agents A is a tuple of strategies, one per agent from A. Strategy s_a is *uniform* iff $q \sim_a q'$ implies $s_a(q) = s_a(q')$; a collective strategy is uniform iff it consists of only uniform individual strategies. We denote the set of uniform strategies of agent a by Σ_a; the set of uniform collective strategies of team A is given by $\Sigma_A = \times_{a \in A} \Sigma_a$, and the set of all uniform *strategy profiles* by $\Sigma = \Sigma_{\mathbb{A}\mathrm{gt}}$.

Definition 3 (CEGSP, plausibility model) *A* concurrent epistemic game structure with plausibility (CEGSP) *is given by* $M = \langle \mathbb{A}\mathrm{gt}, St, \Pi, \pi, Act, d, o, \sim_1, \ldots, \sim_k, \Upsilon, \mathbb{R}\mathrm{gt}, \Omega, [\![\cdot]\!] \rangle$, *where* $\langle \mathbb{A}\mathrm{gt}, St, \Pi, \pi, Act, d, o, \sim_1, \ldots, \sim_k \rangle$ *is a* CEGS, $\Upsilon \subseteq \Sigma$ *is a set of plausible strategy profiles (called* plausibility set*),* $\mathbb{R}\mathrm{gt} \subseteq \mathbb{A}\mathrm{gt}$ *is a set of* rational agents *(i.e., the agents to whom the plausibility assumption will apply),* Ω *is a set of plausibility terms, and* $[\![\cdot]\!] : \Omega \times 2^{St} \to \Sigma$ *is a* plausibility mapping *that provides denotation of the terms.*[2] *We refer to* $(\Upsilon, \mathbb{R}\mathrm{gt})$ *as the* plausibility model *of* M*. When necessary, we write* X_M *to denote the element* X *of model* M*.*

Note that imposing strategic restrictions on a *subset* $\mathbb{R}\mathrm{gt}$ of agents can be desirable due to several reasons. It might, for example, be the case that only information about the proponents' play is available; hence, assuming plausible behavior of the opponents is neither sensible nor justified. Or, even simpler, a group of (simple minded) agents might be known not to behave rationally.

Consider now formula $\langle\langle A \rangle\rangle \gamma$: The team A looks for a strategy that brings about γ, but the members of the team who are also in $\mathbb{R}\mathrm{gt}$ can only choose plausible strategies. The same applies to A's opponents that are contained in $\mathbb{R}\mathrm{gt}$.

Definition 4 (Plausibility of strategies) *Let* $s_A|_B$ *be the* $(A \cap B)$*'s substrategy of* s_A*, and* $\Upsilon|_B = \{s_B \in \Sigma_B \mid \exists s \in \Upsilon \; s|_B = s_B\}$*. We say that* s_A *is* plausible *iff* $\mathbb{R}\mathrm{gt}$*'s substrategy in* s_A *is part of some strategy profile in* Υ*, i.e., if* $s_A|_{A \cap \mathbb{R}\mathrm{gt}} \in \Upsilon|_{A \cap \mathbb{R}\mathrm{gt}}$*.*

By Σ^* *we denote the set of all plausible strategy profiles in the model. That is,* $\Sigma^* = \{s \in \Sigma \mid s|_{\mathbb{R}\mathrm{gt}} \in \Upsilon|_{\mathbb{R}\mathrm{gt}}\}$*. Note that* s_A *is plausible iff* $s_A \in \Sigma^*|_A$*.*

A *computation* or *path* $\lambda = q_0 q_1 \cdots \in St^\omega$ is an infinite sequence of states such that there is a transition between each q_i, q_{i+1}. By $\lambda[i] = q_i$ we denote the i-th state of λ. Λ denotes all paths in the model, and $\Lambda(q)$ the set of all paths starting in q.

[2] In this section, the denotation of such terms is fixed; in Section 4 we present a more flexible version.

Definition 5 (Plausible outcome paths) *The* plausible outcome*, $out(q, s_A)$, of strategy s_A from state q is defined as the set of paths (starting from q) which can occur when only plausible strategy profiles can be played and agents in A follow s_A; that is, $out(q, s_A) = \{\lambda \in \Lambda(q) \mid \exists t \in \Sigma^* \; t|_A = s_A$ and $out(q, t) = \{\lambda\}\}$*

Now we define the notion of *formula φ being satisfied* by a (non-empty) set of states Q in model M, written $M, Q \models \varphi$. We will also write $M, q \models \varphi$ as a shorthand for $M, \{q\} \models \varphi$. Note that it is the latter notion of satisfaction (in single states) that we are ultimately interested in – but it is defined in terms of the (more general) satisfaction in sets of states. Let $\mathrm{img}(q, \mathcal{R})$ be the image of state q with respect to binary relation $\mathcal{R}$, i.e., the set of all states q' such that $q\mathcal{R}q'$. Moreover, we use $out(Q, s_A)$ as a shorthand for $\bigcup_{q \in Q} out(q, s_A)$, and $\mathrm{img}(Q, \mathcal{R})$ as a shorthand for $\bigcup_{q \in Q} \mathrm{img}(q, \mathcal{R})$. The semantics is given through the following clauses.

$M, Q \models p$ iff $p \in \pi(q)$ for every $q \in Q$;

$M, Q \models \neg\varphi$ iff $M, Q \not\models \varphi$;

$M, Q \models \varphi \wedge \psi$ iff $M, Q \models \varphi$ and $M, Q \models \psi$;

$M, Q \models \langle\langle A \rangle\rangle \bigcirc \varphi$ iff there exists $s_A \in \Sigma^*|_A$ such that, for every $\lambda \in out(Q, s_A)$, we have that $M, \{\lambda[1]\} \models \varphi$;

$M, Q \models \langle\langle A \rangle\rangle \Box \varphi$ iff there exists $s_A \in \Sigma^*|_A$ such that, for every $\lambda \in out(Q, s_A)$ and $i \geq 0$, we have $M, \{\lambda[i]\} \models \varphi$;

$M, Q \models \langle\langle A \rangle\rangle \varphi \mathcal{U} \psi$ iff there exists $s_A \in \Sigma^*|_A$ such that, for every $\lambda \in out(Q, s_A)$, there is an $i \geq 0$ for which $M, \{\lambda[i]\} \models \psi$ and $M, \{\lambda[j]\} \models \varphi$ for every $0 \leq j < i$.

$M, Q \models \hat{\mathcal{K}} \mathbb{W}_A \varphi$ iff $M, \mathrm{img}(Q, \sim_A^{\mathcal{K}}) \models \varphi$ (where $\hat{\mathcal{K}} = \mathbb{C}, \mathbb{E}, \mathbb{D}$ and $\mathcal{K} = C, E, D$, respectively).

$M, Q \models \mathbf{Pl}_A \varphi$ iff $M', Q \models \varphi$, where the new model M' is equal to M but the new set $\mathbb{R}\mathrm{gt}_{M'}$ of rational agents in M' is set to A.

$M, Q \models$ (**set-pl** $\omega)\varphi$ iff $M', Q \models \varphi$ where M' is equal to M with $\Upsilon_{M'}$ set to $[\![\omega]\!]_M^Q$.

$M, Q \models$ (**refn-pl** $\omega)\varphi$ iff $M', Q \models \varphi$ where M' is equal to M with $\Upsilon_{M'}$ set to $\Upsilon_M \cap [\![\omega]\!]_M^Q$.

Like in CSL, we use two notions of validity, *weak* and *strong*, depending on whether formulae are evaluated with respect to single states or sets of states.

Definition 6 (Validity) *We say that φ is* valid *if $M, q \models \varphi$ for all* CEGSP*'s M with plausibility model $(\Sigma, \emptyset)$ (i.e. all strategies are assumed to be plausible and no agent plays plausibly yet) and all states $q \in St_M$.*

In addition to that, we say that φ is strongly valid *if $M, Q \models \varphi$ for* all CEGSP*'s M and all sets of states $Q \subseteq St_M$.*

Note that strong validity is interpreted in *all* models and not only in those with plausibility model $(\Sigma, \emptyset)$. This stronger notion is necessary for interchangeability of (sub)formulae. The following results are straightforward.

Proposition 2 *Strong validity implies validity.*

Proposition 3 *If $\varphi_1 \leftrightarrow \varphi_2$ is strongly valid, and ψ' is obtained from ψ through replacing an occurrence of φ_1 by φ_2, then $M, Q \models \psi$ iff $M, Q \models \psi'$.*

We also say that φ is *satisfiable* if $M, q \models \varphi$ for some CEGSP with plausibility model $(\Sigma, \emptyset)$.

2.4 Interpretation of Derived Operators

In this section we motivate the logic's epistemic and doxastic operators. We especially show that the syntactic definitions for the derived knowledge and belief operators have an intuitive semantics.

2.4.1 Knowledge

The concept behind knowledge is very simple: It is about everything which is "physically" possible, i.e., *all* behaviors are taken into account (not only the plausible ones). In particular this means that, once a knowledge operator occurs, the set of rational agents in the plausibility model becomes void, indicating that *no* agent is assumed to play rationally.

2.4.2 Weakly and Strongly Rational Beliefs

Constructive weak beliefs (cwb) ("common belief", "distributed belief", and "mutual belief") are primitive operators in our logic. All other belief/knowledge operators are derived from cwb and plausibility. In this section, we mainly discuss individual knowledge and beliefs, but the analysis extends to collective attitudes in a straightforward way.

Let us for example consider the individual cwb operator $\mathbb{W}_a\varphi$, with the following reading: Agent a has *constructive weak belief* in φ iff φ holds in all states that a considers possible, where all agents behave according to the currently specified plausibility model (Υ, A). That is, agents in A are assumed to play as specified in Υ. It is important to note that *weakly rational* beliefs restrict *only* the behavior of the agents specified in the current plausibility model (i.e. A). This is the difference between weak and strong beliefs – the latter assume plausible behavior of *all* the agents. This is why we call such beliefs *strongly rational*, as it restricts the behavior of the system in a more rigorous way due to stronger rationality assumptions.

Using rationality assumptions to define beliefs makes them rather specific. They differ from most "standard" concepts of belief in two main respects. Firstly, our notion of beliefs is focused on *behavior* and *abilities* of agents. When no action is considered, all epistemic and doxastic notions coincide.

Proposition 4 *Let φ be a propositional formula. Then, $\mathbb{W}_a\varphi \leftrightarrow \mathbb{B}\mathsf{el}_a\varphi \leftrightarrow \mathbb{K}_a\varphi$ is strongly valid.*

Secondly, rational beliefs are about *restricting the expected behavior* due to rationality assumptions: Irrational behaviors are simply disregarded. To strengthen this important point consider the following statements:

(i) *Ann (a) knows how Bill (b) can commit suicide* (which can be formalized as $\mathbb{K}_a\langle\langle b\rangle\rangle\Diamond\mathsf{suicide}$);

(ii) *Ann constructively believes that Bill can commit suicide* (which we tentatively formalize as $\mathbb{B}\mathsf{el}_a\langle\langle b\rangle\rangle\Diamond\mathsf{suicide}$).

In the usual treatment of beliefs, statement (i) should imply statement (ii), but this does not apply to *rational* beliefs. That is because, typically, beliefs and knowledge are both about "hard facts". Thus, if a knows some fact to be true, she should also include it in her belief base. On the other hand, our reading of $\mathbb{B}\mathsf{el}_a\langle\langle b\rangle\rangle\Diamond\mathsf{suicide}$ is given as follows: If all agents are constrained to act rationally then Ann knows a strategy for Bill by which he can commit suicide. However, it is natural to assume that no rational entity would commit suicide.[3] Hence, Bill's ability to commit suicide is out of question if we assume him to act rationally. Such an irrational behavior is just unthinkable and thus disregarded by Ann! While she knows how Bob can commit suicide in general, she has no *plausible* recipe for Bob to do that.

A similar analysis can be conducted for standard (i.e., non-constructive) beliefs. Consider the following variants of (i) and (ii):

(i') *Ann knows that Bill has some way of committing suicide* ($\mathsf{K}_a\langle\langle b\rangle\rangle\Diamond\mathsf{suicide}$);

(ii') *Ann believes, taking only rational behavior of all agents into account (in particular of Bill), that Bill has the ability to commit suicide* ($\mathsf{Bel}_a\langle\langle b\rangle\rangle\Diamond\mathsf{suicide}$).

Like before, (i') does not imply (ii'). While Ann knows that Bill "physically" has some way of killing himself, by assuming him to be rational she disregards the possibility. Bob's assumed rationality constrains his choices in Ann's view. This shows that in our logic knowing φ does not imply rational beliefs in φ. We will justify the intuition on a more concrete example.

Example 1 *There are two agents 1 (Ann) and 2 (Bill). Agent 2 has the ability to jump from a building and commit suicide. However, agent 1 disregards this possibility and considers it rational that 2 will not jump. The corresponding* CEGSP *is shown in Figure 1 where all different states are distinguishable from each other; the set of plausible strategy profiles consists of the single profile s in which both agents play action nop, i.e., they do nothing (in particular, we want to impose that Bill does not jump). Hence, we have $M, q_0 \models \mathsf{K}_1\langle\langle 2\rangle\rangle\bigcirc$ suicide but $M, q_0 \not\models \mathsf{Bel}_1\langle\langle 2\rangle\rangle\bigcirc$ suicide.*

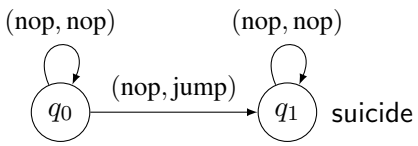


Figure 1: Simple CEGSP.

The following result, in line with [3], is immediate:

Theorem 5 *In general, standard (resp. constructive) knowledge does not imply standard (resp. constructive) rational belief. That is, formulae* $\mathbb{K}_a\varphi \wedge \neg\mathbb{B}\mathsf{el}_a\varphi$, $\mathbb{K}_a\varphi \wedge \neg\mathbb{W}_a\varphi$, $\mathsf{K}_a\varphi \wedge \neg\mathsf{Bel}_a\varphi$, $\mathsf{K}_a\varphi \wedge \neg\mathsf{W}_a\varphi$ *are satisfiable.*

2.4.3 Non-Constructive Knowledge and Beliefs

In this section, we have a closer look at the standard (non-constructive) epistemic and doxastic operators. We mainly focus on strong beliefs; the cases for knowledge and weak beliefs are given analogously.

The non-constructive versions of distributed, common, and everybody belief are based on a specific construction involving the "until" operator. For example, the non-constructive belief of agent a in φ, $\mathsf{Bel}_a\varphi$, is defined as a's *constructive* belief in the ability of the empty coalition to enforce φ until φ. In [9] it was already shown that this definition captures the right notion; we recall the intuition here.

The cooperation modality $\langle\langle\emptyset\rangle\rangle$ ensures that the state formula φ is evaluated *independently* in each indistinguishable state in Q (thus getting rid of its constructive flavour). However, a cooperation modality must be followed directly by a path formula, and φ is a state formula. The trick is to use $\varphi\,\mathcal{U}\,\varphi$ instead, which ensures that φ is true in the initial state of the path. Thus, a believes in φ iff $\mathbf{Pl}\,\varphi$ is independently true in every indistinguishable state. The following proposition (analogous to [9, Theorem 46]) states that all non-constructive operators match their intended intuitions.

Proposition 6 *Let* M *be a* CEGSP, $q \in St_M$, *and* φ *be a* CSLP *formula. Then the following holds, where* $\mathcal{K} = C, E, D$, *respectively:*

1. $M, Q \models \mathcal{K}\mathsf{W}_A\varphi$ *iff* $\Upsilon_M \neq \emptyset$ *and* $M, q \models \varphi$ *for all* $q \in \mathrm{img}(Q, \sim_A^{\mathcal{K}})$;
2. $M, Q \models \mathcal{K}\mathsf{Bel}_A\varphi$ *iff* $M, q \models \mathbf{Pl}\,\varphi$ *for all* $q \in \mathrm{img}(Q, \sim_A^{\mathcal{K}})$;
3. $M, Q \models \mathcal{K}_A\varphi$ *iff* $M, q \models \mathbf{Ph}\,\varphi$ *for all* $q \in \mathrm{img}(Q, \sim_A^{\mathcal{K}})$.

3 Properties of CSLP

In this section, we examine the relationship between plausibility, knowledge and beliefs, and discuss the standard axioms about epistemic and doxastic concepts.

[3] This assumption is given in the plausibility model; it can be any assumption the designer would like to impose on the agents.

3.1 Plausibility, Knowledge and Beliefs

Firstly, we observe that knowledge is commutative with $\mathbf{Ph}$ and belief with $\mathbf{Pl}$, which is a technically important property.

Proposition 7 *Let φ be a* CSLP *formula. Then, we have that $\mathbf{Ph}\,\mathbb{K}_a\varphi \leftrightarrow \mathbb{K}_a\mathbf{Ph}\,\varphi$ and $\mathbf{Pl}\,\mathbb{B}\mathsf{el}_a\varphi \leftrightarrow \mathbb{B}\mathsf{el}_a\mathbf{Pl}$ are strongly valid.*

From the definition of knowledge an belief it follows that a sequence of such operators collapses to the final operator in the sequence.

Proposition 8 *Let $a \in \mathbb{A}\mathrm{gt}$, φ be a* CSLP *formula, and X, Y be sequences of belief/knowledge operators; i.e. $X, Y \in \{\mathbb{B}\mathsf{el}_a, \mathbb{K}_a\}^*$. Then the following formulae are strongly valid:*

$$\textit{(i)}\ X\mathbb{B}\mathsf{el}_a\varphi \leftrightarrow Y\mathbb{B}\mathsf{el}_a\varphi \quad \textit{(ii)}\ X\mathbb{K}_a\varphi \leftrightarrow Y\mathbb{K}_a\varphi$$

In particular, we have that the following formulae are strongly valid: (1) $\mathbb{K}_a\mathbb{B}\mathsf{el}_a\varphi \leftrightarrow \mathbb{B}\mathsf{el}_a\varphi$: Agent a knows that he believes φ iff he believes φ; and (2) $\mathbb{B}\mathsf{el}_a\mathbb{K}_a\varphi \leftrightarrow \mathbb{K}_a\varphi$: Agent a believes that he knows φ iff he knows φ.

Proposition 9 *Let the premises be as in Proposition 8. Then, the following formulae are* not *valid: (i) $X\mathbb{B}\mathsf{el}_a\varphi \leftrightarrow Y\mathbb{K}_a\varphi$; (ii) $\mathbb{B}\mathsf{el}_a\varphi \rightarrow \mathbb{B}\mathsf{el}_a\mathbb{K}_a\varphi$; (iii) $\mathbb{B}\mathsf{el}_a\varphi \rightarrow \mathbb{K}_a\varphi$.*

Proposition 9 says in particular that (ii) an agent who has rational belief in φ does not necessarily believe that he also knows φ; and (iii) an agent who believes in φ does not necessarily know φ. Indeed, both formulae should not hold in a logics of knowledge and belief.

Our definitions of epistemic and doxastic operators from Section 2.2 strongly suggest that the underlying concepts are related. Let us consider formula $\mathbb{K}_a\mathbf{Pl}_B\varphi$: Agent a has constructive knowledge in φ if agents in B behave rationally. This sounds quite similar to beliefs which is formally shown below.

Proposition 10 *$\mathbf{Pl}_A\mathbb{K}_a\mathbf{Pl}_A\varphi \leftrightarrow \mathbf{Pl}_A\mathbb{W}_a\varphi$ is strongly valid. We also have that $\mathbb{K}_a\varphi \leftrightarrow \mathbb{W}_a\varphi$ is valid (but not strongly valid).*

Finally, we conclude that rational beliefs and knowledge can also be defined in terms of each other.

Theorem 11 *$\mathbb{B}\mathsf{el}_a\varphi \leftrightarrow \mathbb{K}_a\mathbf{Pl}\,\varphi$ and $\mathbb{K}_a\varphi \leftrightarrow \mathbb{B}\mathsf{el}_a\mathbf{Ph}\,\varphi$ are strongly valid.*

That is, believing in φ is knowing that φ plausibly holds, and knowing that φ is believing that φ is the case in all physically possible plays.

3.2 Axiomatic Properties

In this section we review the well-known **KDT45** axioms. For modality O these axioms are given as follows:

$(\mathbf{K}_O)$ $O(\varphi \to \psi) \to (O\varphi \to O\psi)$ $(\mathbf{D}_O)$ $O\varphi \to \neg O\neg\varphi$
$(\mathbf{T}_O)$ $O\varphi \to \varphi$ $(\mathbf{4}_O)$ $O\varphi \to OO\varphi$
$(\mathbf{5}_O)$ $\neg O\varphi \to O\neg O\varphi$

We say, for instance, that O is an **K4** modality if axioms $\mathbf{K}_O$ and $\mathbf{4}_O$ are *strongly valid*. The following result is obtained in a way analogous to [9, Theorem 37].

Theorem 12 (Weak beliefs: KD45) W_a *(standard weak beliefs) and* $\mathbb{W}_a$ *(constructive weak beliefs) are **KD45** modalities. Axiom **T** is not valid for both notions of weak beliefs.*

Remark 13 *Despite the similarities to [3], axiom **D** was not strongly valid for beliefs in* CTLKP *because the belief operator directly referred to plausible paths. Hence, if the set of paths was empty some formulae were trivially true (*$\mathsf{Bel}\varphi$*) and others are trivially false (*$\neg\mathsf{Bel}\varphi$*). In* CSLP *the notions of belief and plausibility are more modular.*

As knowledge and strong beliefs are special kinds of weak beliefs, both operators have to satisfy the same axioms as the weak belief operator. It just remains to check whether axiom **T** holds for knowledge or strong beliefs. However, for the same reason as in pure CSL this axiom does usually not hold; we refer to [9] for a rigorous discussion of this issue – including ways how axiom **T** can be restored for knowledge. The problem that **T** is not true for knowledge (what is usually assumed to be a sensible requirement) is due to the definition of negation in the non-standard semantics defined wrt sets of states.

Theorem 14 (Strong beliefs: KD45) *Standard strong beliefs* Bel_a *and constructive strong beliefs* $\mathbb{B}\mathsf{el}_a$ *are **KD45** modalities. Axiom **T** is not valid for both notions of beliefs.*

Theorem 15 (Knowledge: KD45) *Standard knowledge* K_a *and constructive knowledge* $\mathbb{K}_a$ *are **KD45** modalities. Axiom **T** is not valid for both notions of knowledge.*

Note that if we consider a formula φ which does not contain any constructive operators then the following holds.

Theorem 16 *Let $\mathcal{L}$ consist of all* CSLP *formulae that contain no constructive operators. Then:*

1. K_a *is a **KD45** modality in $\mathcal{L}$. Axiom $\boldsymbol{T}_{\mathsf{K}_a}$ is valid (but not strongly valid), and* $\mathbf{Ph}\,(\mathsf{K}_a\varphi \to \varphi)$ *is strongly valid in $\mathcal{L}$.*
2. Bel_a *is a **KD45** modality and* $\mathbf{Pl}\,(\mathsf{Bel}_a\varphi \to \varphi)$ *is strongly valid in $\mathcal{L}$.*

We observe that the validities $\mathbf{Ph}\,(\mathsf{K}_a\varphi \to \varphi)$ and $\mathbf{Pl}\,(\mathsf{Bel}_a\varphi \to \varphi)$ are very similar to the truth axiom **T**.

3.3 Relationship to Existing Logics

In this section, we compare CSLP with several relevant logics and show their formal relationships. To this end, we define the notion of *embedding*. *Logic* L_1 *embeds logic* L_2 iff there is a translation tr of L_2 formulae into formulae of L_1, and a transformation TR of L_2 models into models of L_1, such that $M, q \models_{L_2} \varphi$ iff $TR(M), q \models_{L_1} tr(\varphi)$ for every pointed model M, q and formula φ of L_2.

The following theorem is straightforward from the definition of the logic.

Theorem 17 CSLP *embeds* ATL*,* ATLP*, and* CSL.

It is easy to see that W_a is even a **KDT45** modality for a sublanguage of CSLP and that this sublanguage can embed standard epistemic propositional logic.

Proposition 18 CSLP *embeds standard epistemic propositional logic.*

The following result is not that obvious but follows from Proposition 18 and [4, Proposition 5].

Proposition 19 CSLP *embeds* CTLKP *in the class of epistemic Kripke structures.*

Remark 20 *In [9] and [4] it was shown that* CSL *and* ATLP *embed several other logics, e.g.,* ATEL *[15],* ATLI *[11], and* GLP *[17]. Due to Theorem 17 all these logics are also embeddable in* CSLP.

4 Flexible Specifications

In [10, 4] we showed that ATLP can be used to reason about temporal properties of rational play. In particular it was shown that the logic allows to characterize game theoretic solution concepts of perfect information games [12]. These characterizations were then used to describe agents rational behavior and impose the resulting rationality constraints on them. Here we show that CSLP can be used for the same purpose in the more general case of *imperfect information games* (IIG). A natural question is how solution concepts for both game-types differ?

Actually, they do not differ much. For instance, a Nash equilibrium is a strategy profile from which no agent can deviate to obtain a better payoff, for both the perfect and imperfect information case. However, only *uniform strategies* are considered for IIG. Moreover, we require the agent to *know/identify* a strategy successful in *all* states indistinguishable for him.

Before we present how solution concepts can be described in Section 4.2 we need to pave the way for it: CSLP is not yet expressive enough to *describe* strategies in the object language, only predefined plausibility terms are available.

4.1 Nesting Formulae in CSLP

In this section we present $\mathcal{L}^1_{\text{CSLP}}$ which extends $\mathcal{L}_{\text{CSLP}}$ so that plausibility terms are constructed from $\mathcal{L}_{\text{CSLP}}$ formulae.[4] In the following we proceed in an analogous way to [4]. The *extended plausibility terms* of $\mathcal{L}^1_{\text{CSLP}}$ have a structure similar to $\sigma_1.D(\sigma_1)$. Such a term *selects* all strategy profiles s_1 (referred to by the *strategic variable* σ_1) that satisfy a property D which depends on a given model, set of states, and σ_1. Let us be more precise about the structure of such properties. We allow them to be quantified $\mathcal{L}_{\text{CSLP}}$ formulae, e.g., $D(\sigma_1, \ldots, \sigma_n) = \forall\sigma_2\exists\sigma_3 \ldots \forall\sigma_n\varphi(\sigma_1, \ldots, \sigma_n)$, where the quantification takes places over strategy profiles which can be used inside φ in the same way as basic plausibility terms would be used. The variable σ_1 takes on a specific role; it *collects* the "good" strategy profiles.

Before we formally define the language we need one more notation. Solution concepts often require to combine strategies or focus on substrategies. For example, given a term ω_{NE} (describing Nash equilibria) and a term ω_{PO} (describing Pareto optimal strategies) we can use $\langle\omega_{\text{NE}}, \omega_{\text{PO}}\rangle$ to refer to all profiles in which agent 1 plays his part of an Nash equilibrium and agent 2 plays a Pareto optimal strategy. Likewise, $\omega_{\text{NE}}[1]$ refers to the strategy profiles in which 1's substrategy is a part of some Nash equilibrium.

Formally, given a non-empty set X we say that y is a *strategic combination of* X if it is generated by the following grammar: $y ::= x \mid \langle y, \ldots, y\rangle \mid y[A]$ where $x \in X$, $\langle y, \ldots, y\rangle$ is a vector of length $|\mathbb{A}\text{gt}|$, and $A \subseteq \mathbb{A}\text{gt}$. The set of *strategic combinations* over X is defined by $\mathcal{T}(X)$. It is easy to see that operator $\mathcal{T}$ is idempotent ($\mathcal{T}(X) = \mathcal{T}(\mathcal{T}(X))$). Below, we define the language $\mathcal{L}^1_{\text{CSLP}}$.

Definition 7 ($\mathcal{L}^1_{\text{CSLP}}$) *Let* $\mathbb{A}\text{gt}$ *be a set of agents,* Π *a set of propositions, and* $\mathcal{V}ars$ *a set of* strategic variables *(with typical element* σ*). The logic* $\mathcal{L}^1_{\mathcal{CSLP}}(\mathbb{A}\text{gt}, \Pi, \mathcal{V}ars)$ *is defined as* $\mathcal{L}_{\mathcal{CSLP}}(\mathbb{A}\text{gt}, \Pi, \mathcal{T}(\Omega_1))$ *where* Ω_1 *is given by*

$$\{\sigma_1.(Q_2\sigma_2)\ldots(Q_n\sigma_n)\varphi \mid n \in \mathbb{N}, \forall i\ (1 \leq i \leq n \Rightarrow \sigma_i \in \mathcal{V}ars,\\ Q_i \in \{\forall, \exists\}, \varphi \in \mathcal{L}_{\mathcal{CSLP}}(\mathbb{A}\text{gt}, \Pi, \mathcal{T}(\{\sigma_1, \ldots, \sigma_n\})))\ \}.$$

The semantics of $\mathcal{L}^1_{\text{CSLP}}$ formulae is analogously defined as for the base language but instead of the basic plausibility mapping $[\![\cdot]\!]$, the *extended plausibility mapping* $\widehat{[\![\cdot]\!]}_M$ is used, defined as follows:

1. If $\omega \in \Omega$ then $\widehat{[\![\omega]\!]}^Q_M = [\![\omega]\!]^Q_M$;
2. If $\omega = \omega'[A]$ then $\widehat{[\![\omega]\!]}^Q_M = \{s \in \Sigma \mid \exists s' \in \widehat{[\![\omega']\!]}^Q_M\ s|_A = s'|_A\}$;
3. If $\omega = \langle\omega_1, \ldots \omega_k\rangle$ then $\widehat{[\![\omega]\!]}^Q_M = \{s \in \Sigma \mid \exists t_1 \in \widehat{[\![\omega_1]\!]}^Q_M, \ldots, \exists t_k \in \widehat{[\![\omega_k]\!]}^Q_M \forall i = 1, ..., k\ s|_{a_i} = t_i|_{a_i})\}$;

[4] In order to give a brief presentation we do not allow "basic" plausibility terms anymore.

4. If $\omega = \sigma_1.(Q_2\sigma_2)\ldots(Q_n\sigma_n)\varphi$ then $\widehat{[\![\omega]\!]}^Q_M = \{s_1 \in \Sigma \mid Q_2 s_2 \in \Sigma, \ldots, Q_n s_n \in \Sigma \quad (M^{s_1,\ldots,s_n}, q \models \varphi)\}$, where $M^{s_1,\ldots,s_n}$ is equal to M except that we fix $\Upsilon_{M^{s_1,\ldots,s_n}} = \Sigma$, $\Omega_{M^{s_1,\ldots,s_n}} = \Omega_M \cup \{\sigma_1, \ldots, \sigma_n\}$, $[\![\sigma_i]\!]^Q_{M^{s_1,\ldots,s_n}} = \{s_i\}$, and $[\![\omega]\!]^Q_{M^{s_1,\ldots,s_n}} = [\![\omega]\!]^Q_M$ for all $\omega \neq \sigma_i$, $1 \leq i \leq n$, and $Q \subseteq St_M$. That is, the denotation of σ_i in $M^{s_1,\ldots,s_n}$ is set to strategy profile s_i.

An example $\mathcal{L}^1_{\text{CSLP}}$ formula is
(**set-pl** $\sigma.\langle\!\langle\emptyset\rangle\!\rangle\Box$(**Ph** $\langle\!\langle\mathbb{A}\text{gt}\rangle\!\rangle\bigcirc$ alive $\rightarrow$ (**set-pl** σ)**Pl** $\langle\!\langle\emptyset\rangle\!\rangle\bigcirc$ alive))
$\neg\text{Bel}_a\langle\!\langle b\rangle\!\rangle\Diamond$suicide: Assuming that rational agents avoid death whenever they can, it is not rational of Ann to believe that Bob can commit suicide.

Remark 21 *The nestings can be increased step by step which results in a hierarchy of logics, $\mathcal{L}^k_{CSLP}$ ($k = 1, 2, \ldots$) as in [4].*

4.2 Solution Concepts under Uncertainty

In this section we characterize solution concepts for imperfect information games in $\mathcal{L}^1_{\text{CSLP}}$. Before we do that, however, we need some way to *evaluate* different strategies. In game theory real values (payoffs) or preference relations are used to define the outcome of a given strategy. Here, we follow the approach from [4] which equips agents with *winning criteria* $\vec{\eta} = \langle\eta_1, \ldots, \eta_k\rangle$ (one per agent) where $k = |\mathbb{A}\text{gt}|$. Each criterion η_a of agent a is a temporal formula. Intuitively, a given strategy profile is successful for an agent a iff the winning criterion is fulfilled on *all* resulting paths starting from *any* indistinguishable state given the strategy profile. This requirement is motivated by the fact that an agent does not know whether the system is in q or q' provided that q and q' are indistinguishable for him. So, he should play a strategy which is "good" in both states to ensure success.

Definition 8 (From CEGSP To NF Game) *Let M be a* CEGSP*, $q \in St_M$, and $\vec{\eta}$ be a vector of winning criteria.*

We define $\mathcal{N}(M, \overrightarrow{\eta}, q)$, the normal form game associated with M, $\overrightarrow{\eta}$, and q*, as the normal form game $\langle\mathbb{A}\text{gt}, \mathcal{S}_1, \ldots, \mathcal{S}_k, \mu\rangle$, where the set $\mathcal{S}_a$ of a's strategies is given by Σ_a (a's uniform strategies) for each $a \in \mathbb{A}\text{gt}$, and the payoff function is defined as follows:*

$$\mu_a(a_1, \ldots, a_k) = \begin{cases} 1 & \textit{if } M, \lambda \models \eta_a \\ & \textit{for all } \lambda \in out(img(q, \sim_a), \langle a_1, \ldots, a_k\rangle), \\ 0 & \textit{else} \end{cases}$$

To give a clear meaning to solution concepts in a CEGSP, we relate them to the associated normal form game. The first solution concept we will define is a *best-response strategy* for IIG. Given a strategy profile $s_{-a} := (s_1, \ldots, s_{a-1}, s_{a+1}, \ldots, s_k)$ where $k = |\mathbb{A}\text{gt}|$ a strategy s_a is said to be a *best response* to s_{-a} if there is no better strategy for agent a given s_{-a}. Now, s is a *best response profile* wrt a if $s|_a$ is a best response against $s|_{\mathbb{A}\text{gt}\setminus\{a\}}$. According to [4]

σ is a best response profile for perfect information games wrt a and $\vec{\gamma}$ in M, q if $M, q \models (\textbf{set-pl}\ \sigma\mathbb{A}\mathrm{gt}\backslash\{a\}])\mathbf{Pl}\left(\langle\!\langle a\rangle\!\rangle\eta_a \rightarrow (\textbf{set-pl}\ \sigma)\langle\!\langle\emptyset\rangle\!\rangle\eta_a\right)$. It is read as follows: If agent a has any strategy to enforce η_a against $\sigma[\mathbb{A}\mathrm{gt}\setminus\{a\}]$ then his strategy given in σ should enforce η_a as well.

What do we have to modify to make it suitable for imperfect information games? Firstly, we have to ensure that the strategy σ is uniform, and indeed only uniform strategies are taken into account in the semantics of CSLP. Secondly, since the agent might not be aware of the real state of the system the described strategy should have its desired characteristics in every indistinguishable state. The agent should be able to *identify* the strategy; the key motivation behind CSL. For this purpose CSLP provides the constructive belief operators; recall that $\mathbb{W}_a\langle\!\langle a\rangle\!\rangle$ means that a has a single strategy successful in all indistinguishable states. To ensure this second point we just have to couple strategic operators with constructive operators. So we obtain the following description of a best response strategy for IIG:

$$BR_a^{\vec{\eta}}(\sigma) \equiv (\textbf{set-pl}\ \sigma[\mathbb{A}\mathrm{gt}\backslash\{a\}])\mathbf{Pl}\,(\mathbb{W}_a\langle\!\langle a\rangle\!\rangle\eta_a \rightarrow (\textbf{set-pl}\ \sigma)\mathbb{W}_a\langle\!\langle\emptyset\rangle\!\rangle\eta_a).$$

Other solution concepts characterized in [4] can be adapted to IIG's following the same scheme, e.g.:

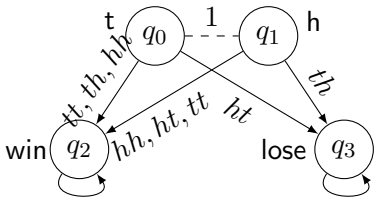
Nash equilibrium (NE): $NE^{\vec{\eta}}(\sigma) \equiv \bigwedge_{i\in\mathbb{A}\mathrm{gt}}\ BR_i^{\vec{\eta}}(\sigma)$;
Subgame perfect NE: $SPN^{\vec{\eta}}(\sigma) \equiv \mathbb{E}\mathbb{W}_{\mathbb{A}\mathrm{gt}}\langle\!\langle\emptyset\rangle\!\rangle\Box\ NE^{\vec{\eta}}(\sigma)$;
Pareto optimal strategy (PO):

$$PO^{\vec{\eta}}(\sigma) \equiv \forall\sigma'\ \mathbf{Pl}\Big(\bigwedge_{a\in\mathbb{A}\mathrm{gt}}((\textbf{set-pl}\ \sigma')\mathbb{W}_a\langle\!\langle\emptyset\rangle\!\rangle\eta_a \rightarrow (\textbf{set-pl}\ \sigma)\mathbb{W}_a\langle\!\langle\emptyset\rangle\!\rangle\eta_a) \vee \bigvee_{a\in\mathbb{A}\mathrm{gt}}((\textbf{set-pl}\ \sigma)\mathbb{W}_a\langle\!\langle\emptyset\rangle\!\rangle\eta_a \wedge \neg(\textbf{set-pl}\ \sigma')\mathbb{W}_a\langle\!\langle\emptyset\rangle\!\rangle\eta_a\Big).$$

The following result shows that these concepts match the underlying intuitions.

Theorem 22 *Let M be a* CEGSP, *$q \in St_M$, $\vec{\eta}$ a vector of winning criteria, and $\mathcal{N} := \mathcal{N}(M, \overrightarrow{\eta}, q)$. Then:*

1. *The set of* NE *strategies in $\mathcal{N}$ is given by $[\![\sigma.\widehat{NE^{\vec{\eta}}(\sigma)}]\!]_M^{\{q\}}$*
2. *The set of* PO *strategies in $\mathcal{N}$ is given by $[\![\sigma.\widehat{PO^{\vec{\eta}}(\sigma)}]\!]_M^{\{q\}}$*
3. *Let Q' collect the states that any agent from A considers possible, i.e., $\mathrm{img}(\{q\}, \sim_{\mathbb{A}\mathrm{gt}}^E)$), plus all states reachable from them by (a sequence of) temporal transitions.*
   Then, $[\![\sigma.\widehat{SPN^{\vec{\eta}}(\sigma)}]\!]_M^{\{q\}}$ is equal to $\bigcap_{q'\in Q'}[\![\sigma.\widehat{NE^{\vec{\eta}}(\sigma)}]\!]_M^{\{q'\}}$.



$1\backslash 2$	h	t
h	$1,1$	$0,0$
t	$0,1$	$1,1$

Figure 2: Simple CEGSP.

Example 2 *Consider the* CEGSP *given in Figure 2. There are two agents,* 1 *and* 2*, and a coin which initially shows tail (*q_0*) or head (*q_1*); agent* 1 *cannot distinguish between them. Now, both agents win if* 1 *guesses the right side of the coin or if both agents agree on one side (regardless of whether it is the right one). For instance, the tuple* th *denotes that* 1 *says tail and* 2 *head. Moreover, we assume that both agents have the winning criterion* $\bigcirc$win*. The associated NF game wrt* q_0 *is also given in Figure 2. Now we have that* $[\![\sigma.\widehat{NE^{\vec{\eta}}(\sigma)}]\!]_M^{\{q\}} = \{hh, tt\}$*: Only if both agents agree on the same side, winning is guaranteed.*

5 Model Checking Rational Play under Imperfect Information

In this section we discuss the model checking complexity of $\mathcal{L}_{\text{CSLP}}$ and $\mathcal{L}^1_{\text{CSLP}}$. Given a formula φ, a model M, and a set of states $Q \subseteq St_M$ the associated *model checking* problem is to determine whether $M, Q \models \varphi$ holds or not. In the following we use l to refer to the length of φ and m to denote the number of transitions in M. We only consider a restricted class of models in which the check for plausibility of a strategy profile can be done in polynomial time (wrt l and m) by a non-deterministic Turing machine. In order to conduct a sensible analysis such an assumption is necessary. To this end, we adapt an important notion from [4].

Definition 9 (Well-Behaved CEGSP) *A* CEGSP M *is called* well-behaved *if, and only if, (1)* $\Upsilon_M = \Sigma$*: all the strategy profiles are plausible in* M*; and (2) there is an algorithm which determines whether* $s \in [\![\omega]\!]_M^Q$ *for every set* $Q \subseteq St_M$*, strategy profile* $s \in \Sigma$*, and plausibility term* $\omega \in \Omega$ *in nondeterministic polynomial time wrt the length of* ω *and the number of transitions in* M*.*

We begin by reviewing the existing results for CSL and ATLP separately. The complexity results for CSLP follow in a natural way. In [9] it was shown that CSL model checking is $\mathbf{\Delta_2^P}$-complete,[5] the hard cases being formulae $\langle\!\langle A \rangle\!\rangle \Box \varphi$ and $\langle\!\langle A \rangle\!\rangle \varphi_1 \,\mathcal{U}\, \varphi_2$. The formulae require existence of a single uniform strategy which is successful in *all* states of Q. In the algorithm from [9], the strategy is guessed by

[5] $\mathbf{\Delta_2^P} = \mathbf{P^{NP}}$ is the class of problems that can be solved in polynomial time by a deterministic Turing machine that makes adaptive calls to an $\mathbf{NP}$ oracle.

the oracle and then verified in polynomial time (see further). Nested cooperation modalities are model-checked recursively (bottom-up) which puts the algorithm indeed in $\mathbf{\Delta_2^P}$.

We also recall from [4] that ATLP model checking is $\mathbf{\Delta_3^P} = \mathbf{P^{NP^{NP}}}$-complete. The algorithm for checking the hard cases ($\langle\langle A\rangle\rangle\Box\varphi$ and $\langle\langle A\rangle\rangle\varphi_1\,\mathcal{U}\,\varphi_2$) is similar: Firstly, a *plausible* strategy of A is guessed (first $\mathbf{NP}$-oracle call) and verified against all *plausible* strategies of the opponents (second $\mathbf{NP}$-oracle call, the "worst" response of the opponents is guessed). Note that, as soon as the relevant strategy (or strategy profile) s is fixed, the remaining verification can be done in deterministic polynomial time: it is enough to "trim" the model by deleting all transitions which cannot occur when the agents follow s, and model check a CTL formula in the trimmed model (which can be done in polynomial time [5]).

For $\mathcal{L}_{\text{CSLP}}$, we essentially use the ATLP model checking algorithm from [4] with an additional check for uniformity of strategies. This does not influence the complexity. We obtain the following result (we refer to [4, 9] for details).

Theorem 23 *Model checking $\mathcal{L}_{CSLP}$ in the class of well-behaved* CEGSP*'s is $\mathbf{\Delta_3^P}$-complete with respect to l and m.*

Proof (sketch). The hardness follows from the fact that $\mathcal{L}_{\text{ATLP}}$ is $\mathbf{\Delta_3^P}$-complete and can be embedded in $\mathcal{L}_{\text{CSLP}}$ (cf. Proposition 17). For the inclusion in $\mathbf{\Delta_3^P}$, we sketch the algorithm for $M, Q \models \langle\langle A\rangle\rangle\Box\varphi$: (1) Model-check φ recursively for each $q \in St$, and label the states for which $M, q \models \varphi$ with a new proposition p; (2) Guess a "good" plausible uniform strategy s_A; (3) Guess a "bad" uniform plausible strategy profile t such that $t|_A = s_A$; and (4) Return true if $Q \subseteq mcheck_{\text{CTL}}(M', \mathsf{A}\bigcirc\mathsf{p})$ and false otherwise, where M' is the trimmed model of M wrt profile t. ■

In the previous section we showed how CSLP can be used to characterize incomplete information solution concepts. However, for this reason we had to extend the language. An obvious question arises: How much does the complexity increase? The answer is quite appealing: The increase in complexity depends on how much extra-expressiveness we actually use; and in any case, we get some expressiveness for free! This can be shown analogously to [4]; here, we just give a brief summary. The model checking complexity can be completely characterized in the number of quantifier *alternations* used in the extended plausibility terms. If we have no quantifiers at all, the resulting sublanguage is no more costly to verify than the base version. Note that the quantifier-free sublanguage of $\mathcal{L}^1_{\text{CSLP}}$ is already sufficient to "plug in" important solution concepts (e.g., Nash equilibria). For each additional quantifier alternation (starting with a universal quantifier) the complexity is pushed one level up in the polynomial hierarchy. For a more detailed discussion, cf. [4].

Theorem 24 *Let $\mathcal{L} \subseteq \mathcal{L}^1_{CSLP}$ such that each sequence of quantifiers starting with an universal one in any extended plausibility term has at most i quantifier alternations. Then, model checking formulae of $\mathcal{L}$ in the class of well-behaved* CEGSP*'s is in $\mathbf{\Delta^P_{3+i}}$ with respect to l and m.*

Proof (sketch). The extension of the base algorithm discussed above is done in an analogous way to [4]. For each quantifier alternation one has to guess a new strategy. But the first existential quantified strategic variables can be guessed together with the proponents and opponents strategies; thus, no more oracle levels need to be added. ■

6 Conclusions

In this paper, we propose a logic which relates epistemic and doxastic concepts in a specific way; more importantly, it allows to reason about the outcome of rational play in imperfect information games. In the logic, called CTLKP, beliefs are defined on top of the primitive notions of plausibility and indistinguishability. We analyze the relationship between beliefs, knowledge, and rationality, and prove in particular that rational beliefs form a **KD45** modality. CSLP embeds both CTLKP and CSL; thus, the combination of knowledge, rationality, and strategic action turns out to be strictly more expressive than each of the subsets.

Moreover, we show how some important solution concepts can be characterized and used for reasoning about imperfect information scenarios. Finally, we prove that the model checking problem for the basic variant of CSLP is $\mathbf{\Delta_3^P}$-complete. That is, the complexity of model checking is only slightly higher than for CSL, and no worse than for ATLP.

Acknowledgements. Wojciech Jamroga acknowledges support of the Polish development project no. O R000024 04.

References

[1] R. Alur, T. A. Henzinger, and O. Kupferman. Alternating-time Temporal Logic. In *Proceedings of the 38th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 100–109. IEEE Computer Society Press, 1997.

[2] R. Alur, T. A. Henzinger, and O. Kupferman. Alternating-time Temporal Logic. *Journal of the ACM*, 49:672–713, 2002.

[3] N. Bulling and W. Jamroga. Agents, beliefs and plausible behaviour in a temporal setting. In *Proceedings of AAMAS'07*, pages 570–577, 2007.

[4] Nils Bulling, Wojciech Jamroga, and Jürgen Dix. Reasoning about temporal properties of rational play. *Annals of Mathematics and Artificial Intelligence*, 2009. To appear.

[5] E.M. Clarke, E.A. Emerson, and A.P. Sistla. Automatic verification of finite-state concurrent systems using temporal logic specifications. *ACM Transactions on Programming Languages and Systems*, 8(2):244–263, 1986.

[6] E. A. Emerson. Temporal and modal logic. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science*, volume B, pages 995–1072. Elsevier Science Publishers, 1990.

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

[8] N. Friedman and J.Y. Halpern. A knowledge-based framework for belief change, Part I: Foundations. In *Proceedings of TARK*, pages 44–64, 1994.

[9] W. Jamroga and T. Ågotnes. Constructive knowledge: What agents can achieve under incomplete information. *Journal of Applied Non-Classical Logics*, 17(4):423–475, 2007.

[10] W. Jamroga and N. Bulling. A framework for reasoning about rational agents. In *Proceedings of AAMAS'07*, pages 592–594, 2007.

[11] W. Jamroga, W. van der Hoek, and M. Wooldridge. Intentions and strategies in game-like scenarios. In Carlos Bento, Amílcar Cardoso, and Gaël Dias, editors, *Progress in Artificial Intelligence: Proceedings of EPIA 2005*, volume 3808 of *Lecture Notes in Artificial Intelligence*, pages 512–523. Springer Verlag, 2005.

[12] M. Osborne and A. Rubinstein. *A Course in Game Theory*. MIT Press, 1994.

[13] W. Penczek and A. Lomuscio. Verifying epistemic properties of multi-agent systems via bounded model checking. In *Proceedings of AAMAS'03*, pages 209–216, New York, NY, USA, 2003. ACM Press.

[14] K. Su, A. Sattar, G. Governatori, and Q. Chen. A computationally grounded logic of knowledge, belief and certainty. In *Proceedings of AAMAS'05*, pages 149–156. ACM Press, 2005.

[15] W. van der Hoek and M. Wooldridge. Cooperation, knowledge and time: Alternating-time Temporal Epistemic Logic and its applications. *Studia Logica*, 75(1):125–157, 2003.

[16] S. van Otterloo and G. Jonker. On Epistemic Temporal Strategic Logic. *Electronic Notes in Theoretical Computer Science*, XX:35–45, 2004. Proceedings of LCMAS'04.

[17] S. van Otterloo, W. van der Hoek, and M. Wooldridge. Preferences in game logics. In *Proceedings of AAMAS-04*, pages 152–159, 2004.

Characterization of Dominance Relations in Finite Coalitional Games

Felix Brandt Paul Harrenstein
Ludwig-Maximilians-Universität München, Germany
{brandtf, harrenst}@tcs.ifi.lmu.de

Abstract

McGarvey [15] has shown that any irreflexive and anti-symmetric relation can be obtained as a relation induced by majority rule. We address the analogous issue for dominance relations of finite cooperative games with non-transferable utility (coalitional NTU games). We find *any irreflexive relation over a finite set* can be obtained as the dominance relation of some finite coalitional NTU game. We also show that any such dominance relation is induced by a *non-cooperative* game via β-effectivity. Dominance relations obtainable via α-effectivity, however, have to comply with a more restrictive condition, which we refer to as the *edge-mapping property*.

1 Introduction

Many important concepts in the mathematical social sciences are defined in terms of a binary *dominance relation* on a set of outcomes or alternatives. These concepts can thus be applied to any model of social interaction for which such a concept of dominance can be meaningfully defined. For example, the *core*—the set of undominated outcomes—defines for different interpretations of the dominance relation Gillies' core in cooperative game theory, Nash's solution in the bargaining problem, or, more generally, the idea of Pareto optimality [see 3, p. 539]. Other conspicuous notions that are similarly defined in terms of dominance are *von Neumann-Morgenstern stable sets* for cooperative games as well as the *Condorcet winner* along with *Condorcet consistent choice rules*, such as the *Banks set*, the *uncovered set* and Dutta's *minimal covering set* in social choice theory [see, e.g., 13].[1]

[1] There are also numerous concepts that take into account more structure of the social situation at hand. Thus, Fishburn [8] distinguishes $C1$ social choice functions, which merely involve the dominance relation, from $C2$ and $C3$ functions, for which this is not the case. In cooperative game theory, the dominance relation alone does not suffice to determine *the bargaining set*, *the kernel* or *the nucleolus*. The *Shapley value* is defined on an entirely different basis.

For each social or game-theoretic model the notion of dominance is defined differently. In social choice theory, dominance is defined with respect to a profile of individual preferences over a set of social alternatives. Although other definitions are also possible, typically, an alternative a is then said to dominate another alternative b if the number of individuals preferring a to b exceeds the number of individuals preferring b to a. In coalitional games, on the other hand, the concept of dominance is generally defined in terms of *coalitional effectivity* and *individual preferences*. Effectivity can be defined in a number of ways, but intuitively reflects the powers of a coalition in terms of the outcomes it can enforce to come about. An outcome a is then said to dominate another outcome b if some coalition is effective for a and, moreover, all members of that coalition prefer outcome a to outcome b. Thus, cooperative majority voting can be seen as the special case in which the majorities are the only coalitions that are effective for any outcome [see also 21].

In either case, the dominance relation need not generally be transitive and may even contain cycles. Accordingly, the common concept of *maximality* is no longer tenable with respect to the dominance relation and new concepts have to be developed to take over its function of singling out elements that are in some sense primary. Von Neumann and Morgenstern considered this phenomenon as one of the most fundamental problems the mathematical social sciences have to cope with [see 22, Chapter 1]. On this account, each of the concepts mentioned above, be their roots in social choice theory or in cooperative game theory, has to deal with what is essentially the same problem: to come to grips with a possibly intransitive dominance relation. Each of them incorporates another intuition and approaches the issue from a different angle.

The dominance relations themselves, however, have different structural properties in both disciplines. As it is defined for social choice on the basis of the majority rule, the dominance relation is *asymmetric*, i.e., both irreflexive and anti-symmetric. In coalitional games the dominance relation is also irreflexive, but not generally anti-symmetric. The structural properties of a dominance-based solution concept, such as existence and uniqueness, may vary, depending on properties of the underlying dominance relation. Therefore, to judge the merits of a particular dominance-based solution concept as a substitute for maximality, one need to know the structural properties of the dominance relation. In this vein, [15] has shown that, in the setting of majority voting, it is precisely the asymmetric relations on a finite set of alternatives that can be obtained as the dominance relation for some profile of linear preferences over those alternatives.

We take up the analogous issue for *finite coalitional games with non-transferable utility* or *finite coalitional NTU games* and give complete characterizations of the structural properties of the dominance relations for three classes of such games.

The outcomes of a coalitional game are commonly assumed to be a convex and compact subset of Euclidean space. Our results, by contrast, pertain to NTU games, which assume a *finite* number of outcomes [see, e.g., 7, 10, 1, 12]. There is a variety of contexts in which this restriction to a finite number of outcomes is

Dominance Relation	**Properties**	**Result**
Finite NTU games	*irreflexivity*	Theorem 1
Finite NTU games through β-effectivity	*irreflexivity*	Theorem 2
Finite NTU games through α-effectivity	*irreflexivity and EMP*	Theorem 3

Table 1: Characterizing properties of the various types of dominance relation. The *edge-mapping property (EMP)* is defined in Section 4.

natural, desirable or merely convenient, e.g., in bilateral bargaining [11].

A continuum of outcomes could be motivated by the possibility of coalitions playing correlated *mixed* strategies to achieve their ends. Still, in many settings mixed strategies have been argued to be suspect or unnatural. E.g., in matters of life and death players and coalitions may not be willing to have their behavior depend on some randomization device. In other contexts, mixed strategies are simply not available [See, e.g., 14, Section 4.10 for an early discussion]. A finite number of outcomes is also a common and simplifying assumption in the the context of cooperative majority voting games, which, interestingly, is exactly where social choice and cooperative game theory intersect [see, e.g., 19, 6].

Of course, there is a range of environments in which it is natural to assume an infinite number of alternatives. Such settings, however, fall outside the scope of this paper.

Another noteworthy property of the coalitional model studied in this paper is the way the utility a coalition can guarantee its members are related to actual outcomes. In particular, our finitistic model does not assume *comprehensiveness* of the coalitional effectivity functions, in the sense that if a coalition can guarantee each of its members particular utility, it can also guarantee each its members any lesser utility. Here, we take a more general approach. Each finite NTU game we assume to be subject to a so-called *comprehension condition*, which, as a function of the finite set of outcomes, determines the range of utility vectors the various coalitions can be feasible for. Thus, comprehensiveness can be accounted for by imposing a very liberal comprehension condition. On the other extreme, the comprehension condition can be *tight*, meaning that coalitions can only be effective for utility vectors that are actually instantiated by one of the outcomes. How natural the assumptions of comprehensiveness and tightness are largely depends on the setting that is being considered [see 5, for an interesting discussion of comprehensiveness]. Our results hold for every comprehension condition and are thus independent of any specific choice in this respect.

Our first result pertains to the dominance relations of general finite NTU games. We find that every *irreflexive* relation on a finite set A of alternatives can be obtained as the dominance relation of some finite coalitional NTU game. Coalitional NTU games can also be obtained from non-cooperative games, in particular normal form games, in a variety of ways. Traditionally, the notions of α- and β-effectivity are

employed to obtain the characteristic function of a coalitional NTU game [2, 4]. It turns out that the structural properties of the dominance relations of finite coalitional NTU games obtained by means of β-effectivity are no different from those of the general case. The dominance relation induced by a finite NTU game obtained through β-effectivity may be any irreflexive relation. The formal properties of dominance relations obtained through α-effectivity, however, are subject to narrower constraints. We find that they are characterized by irreflexivity and the *edge-mapping property (EMP)*, a structural property defined in this paper. Table 1 summarizes our results.

The significance of these results is mainly of theoretical nature. They show which structural properties of the dominance relation one can rely on when proving something about a dominance-based concept in finite NTU games. On the other hand, they also determine the extent of freedom one has in constructing counterexamples. In this context, it is also worth mentioning that McGarvey even mentioned the construction of voting paradoxes in his seminal 1953 paper.

There is also an interesting conceptual connection between this paper and the literature on *non-cooperative foundations of cooperative solutions*, also commonly referred to as the *Nash progam*. The ambition of this line of research is to provide non-cooperative models, e.g., bargaining environments [16, 17], in which the cooperative and non-cooperative solutions coincide [e.g., 18, 9]. There are also interesting connections with the theory of implementation [e.g., 20, 5]. Assuming comprehensiveness, Bergin and Duggan [5] completely characterized the coalitional games that can be obtained from strategic environments both through α- and β-effectivity. The objective of this paper, however, is different, as it aims to fully characterize the structural properties of the cooperative games obtained in this manner, rather than than the games themselves.[2]

2 Finite Coalitional NTU Games

The intuition underlying the models of coalitional game theory is that the players can make binding commitments, form coalitions and thus correlate their actions. Here we consider the general case in which there is not always the possibility to make side-payments, i.e., we do not hypothesize the existence of a transferable commodity with which all players' preferences are positively associated. We do, however, assume the set of possible outcomes to be finite.

Formally, our framework involves a population $N = \{1, \ldots, |N|\}$ of individuals or players and a finite set $A = \{a_1, \ldots, a_{|A|}\}$ of outcomes or alternatives. A *coalition* C is a non-empty subset of N and we have $-C$ denote the complement $N \setminus C$

[2] Although in the cooperative model of Bergin and Duggan [5] utility is taken to be non-transferable, it is also crucially different from ours in that comprehensiveness is assumed throughout. Also the games they construct to prove their results involve an infinite number of strategies for the players. Consequently, without modification, Bergin and Duggan's Theorems 1 and 2 are not applicable to our finite model.

of C in N. The players entertain preferences over A, which we assume to be represented by a $|N| \times |A|$ *utility matrix* $U = (u_{ij})_{i\in N, j\in A}$, where u_{ij} denotes the utility of the jth outcome to the ith player. Thus, each row $(u_{i1}, \ldots, u_{i|A|})$ could be construed as representing player i's utility function over A. Accordingly we also write $u_i(a_j)$ for u_{ij} and u_i for the entire row. Similarly, for each coalition C we have $u_C(a)$ denote $(u_i(a))_{i\in C}$. On the other hand, each column $(u_{1a}, \ldots, u_{|N|a})$ is a utility vector in $\mathbb{R}^N$, which we also denote by $u(a)$. Given a utility matrix U, we have $H(U)$ denote the set $\{u(a)\colon a \in A\}$ of *feasible utility vectors* in U, omitting the reference to U when it is fixed in the context.

More generally, for each coalition $C \subseteq N$ and for each $x = (x_i)_{i\in N}$ in $\mathbb{R}^N$ we have x_C denote the vector $(x_i)_{i\in C}$. For $X \subseteq \mathbb{R}^N$ we also write $X_C = \{x_C\colon x \in X\}$. With a slight abuse of notation, we use $x_{\{i_1,\ldots,i_k\}}$ for $(x_{i_1}, \ldots, x_{i_k})$, assuming the order of the players to be fixed. If C and D are disjoint coalitions and $x_C \in \mathbb{R}^C$ and $y_D \in \mathbb{R}^D$, let (x_C, y_D) denote the utility vector $(z_i)_{i\in C\cup D} \in \mathbb{R}^{C\cup D}$ with $z_i = x_i$, if $i \in C$, and $z_i = y_i$, if $i \in D$. We also write $x_C \geq y_C$ in case $x_i \geq y_i$ for all $i \in C$ and $x_C > y_C$ if $x_i > y_i$ for all $i \in C$.

By a *comprehension condition* we understand a function χ that associates each subset $X \subseteq \mathbb{R}^n$ with a superset $\chi(X)$ of X. In this paper, we assume comprehension conditions to be *downward*, i.e., for all $X \subseteq \mathbb{R}^n$, $X \subseteq \chi(X) \subseteq \bigcup_{x\in X}\{y \in \mathbb{R}^n\colon x \geq y\}$. The largest comprehension condition, i.e., the one with $\chi(X) = \bigcup_{x\in X}\{y \in \mathbb{R}^n\colon x \geq y\}$ for all X, we call *(full) comprehensiveness*, whereas the smallest, i.e., the one for which $\chi(X) = X$ for all $X \subseteq \mathbb{R}^N$, we refer to as *tight*. It is worth observing that the assumption above excludes the convex hull as a comprehension condition.

We define a *characteristic function* V on $X \subseteq \mathbb{R}^N$ under χ as a function which maps each coalition $C \subseteq N$ to a non-empty subset $V(C)$ of utility vectors $\chi(Y)$ such that $Y \subseteq X_C$. Intuitively, a characteristic function associate with each coalition a set of utility vectors each coalition can guarantee its members. What this guarantee amounts to, is left implicit.

Definition 1 (Finite NTU games)**.** A *finite coalitional game with non-transferable utility* or *finite NTU game* under a comprehension condition χ is a tuple (N, H, V) consisting of a population N, a set $H \subseteq \mathbb{R}^N$ of feasible utility vectors given for some $|N| \times |A|$ utility matrix U defining the preferences of the players in N over a finite set A of outcomes, and a characteristic function V on H under χ.

A finite coalitional NTU game (N, H, V) is *comprehensive* if for each coalition C, $y_C \in \mathbb{R}^C$ and $x_C \in V(C)$, $x_C \geq y_C$ implies $y_C \in V(C)$. Clearly comprehensiveness can only be satisfied if the associated comprehension condition is full comprehensiveness. Similarly, we say (N, H, V) is *tight* if the associated comprehension condition is tight and, consequently, $V(C) \subseteq H_C$, for all coalitions C. A coalitional NTU game is said to be *ordinary* whenever $H \subseteq V(N)$, i.e., if the grand coalition N of all players is effective for every feasible outcome [see, e.g., 3]. It is *monotonic* in case $C \subseteq D$ and $x_C \in V(C)$ imply that there is some $y \in V(D)$ such that $y_C \geq x_C$, i.e., if a coalition can guarantee its members at least as much as each of its subcoalitions. A stronger condition is that of *superaddivity*, which a

characteristic function V satisfies if for all disjoint coalitions C and D, $x_C \in V(C)$ and $y_D \in V(D)$ imply $(x_C, y_D) \in V(C \cup D)$. Superadditivity implies monotonicity but not vice versa. Finally, a coalitional NTU game is *binary* if $H \subseteq \{0,1\}^N$.

For (N, H, V) a finite NTU game, the set H of feasible utility vectors corresponds to a finite set of actual outcomes. Each utility vector in H, thus, represents a distribution of utility that can actually come about. On the other hand, the utility vectors in $V(C)$ for which a coalition C is effective and which need not all of them be included in H_C, could be interpreted as representing the bargaining position of C. The comprehension condition determines how the bargaining position is related to the outcomes the coalition can achieve. More particularly, a coalition's bargaining position may be based on the *sets* of outcomes within which it can enforce the outcome to fall, rather than particular individual outcomes it can force to come about. Thus, a coalition C may be able to guarantee that the outcome is among a and b but can not enforce either a or b separately. Suppose that C consists of two players, 1 and 2, and the utility vectors associated with a and b for C are given by $u(a) = (2,1)$ and $u(b) = (1,2)$. The coalition C could then demand a utility of 1 to both of its members on this basis, even if there is no outcome that yields precisely 1 to both player 1 and player 2. If the circumstances are such that such a demand can reasonably be made, the utility vector $(1,1)$ should be included in $V(C)$, calling for a comprehension condition that makes this possible. On the other hand, if no such claim can be made, coalition C should not be effective for $(1,1)$ and the situation should be modeled by means of a tighter comprehension condition.

Formally, a utility vector u in $\mathbb{R}^N$ is understood to be *feasible for* C if there is some $x_C \in V(C)$ with $x_C \geq u_C$. We also say that a coalition is *effective* for a utility vector u if u is feasible for C. We also say that a coalition is *effective for an outcome* a if C is effective for x_C and $x_C = u_C(a)$ for some outcome a. Now, the notion of *dominance* in NTU games is defined in terms of players' preferences and coalitional effectivity.

Definition 2 (Dominance)**.** Let (N, H, V) be a coalitional NTU game and let C be a coalition. For u and u' utility vectors in H, we say u *dominates* u' *via* C, in symbols $u \succ_C u'$, if u is feasible for C and $u_C > u'_C$, i.e., if there is some $x_C \in V(C)$ with $x_C \geq u_C > u'_C$. Utility vector u *dominates* u', in symbols $u \succ u'$, whenever u dominates u' via some coalition C.

Obviously no utility vector dominates itself, i.e., the dominance relation for NTU games is irreflexive.

Let H be a set of feasible utility vectors defined by a utility matrix U which defines the players' preferences over a finite set of outcomes A. Then, every dominance relation on H straightforwardly defines a dominance relation on A. Thus, we say that *outcome* a dominates outcome b whenever $u(a)$ dominates $u(b)$. More Formally, we say that a binary relation R on a finite set $A = \{a_1, \ldots, a_{|A|}\}$ *is induced by a finite coalitional NTU game* (N, H, V) whenever a utility matrix $U = (u_{ij})_{i \in N, j \in A}$ exists such that $H = H(U)$, $|H| = |A|$ and, for all $a, b \in A$, aRb if and only if a

dominates b, i.e., if the function which maps each $a \in A$ to the vector $u(a) \in H$ is an isomorphism between the graphs (A, R) and $(H, \succ)$. We now have the following useful lemma, which basically says that—as far as the structure of the dominance relations is concerned—we can restrict our attention to tight games without loss of generality.

Lemma 1. *Let (N, H, V) be a tight finite NTU game and χ a comprehension condition. Then, there is a finite NTU game (N, H, V') under χ such that the dominance relations of (N, H, V) and (N, H, V') coincide.*

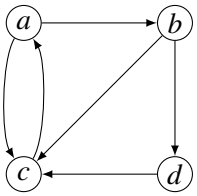
Proof. Define the characteristic function V' such that $V'(C) = \chi(V(C))$ for every coalition $C \subseteq N$. Consider the finite NTU game (N, H, V') and let $\succ$ and $\succ'$ denote the dominance relations of (N, H, V) and (N, H, V'), respectively. We show that $u \succ u'$ if and only if $u \succ' u'$ for all $u, u' \in H$. First assume $u \succ u'$. Then there is some coalition C and some $x_C \in V(C)$ such that $x_C \geq u_C > u'_C$. Hence, $x_C \in \chi(V(C))$ and, accordingly, $x_C \in V'(C)$. Therefore also $u \succ' u'$. For the other direction, assume $u \succ' u'$. Then there is some coalition C and some $x_C \in V'(C)$ such that $x_C \geq u_C > u'_C$. As $x_C \in \chi(V(C))$ and comprehension conditions being downward, there is some $y_C \in V(C)$ such that $y_C \geq x_C$. It follows that $u_C \in V(C)$ and $u \succ u'$. □

3 Dominance Relations of Finite Coalitional NTU Games

We are now in a position to prove our first result, which states that every irreflexive relation on a finite set of outcomes can be induced as the dominance relation of some coalitional NTU game. The idea behind the proof is to construct a coalitional game for each irreflexive relation R on a set of outcomes A. We introduce two players i_a and j_a for each $a \in A$ as well as an appropriate utility matrix U, which depends on R and represents the players' preferences over A. We then set the set feasible utility vectors $H(U) = \{u(a) \colon a \in A\}$. Each coalition that contains both i_a and j_a for some $a \in A$, is defined to be universally effective, i.e., any such coalition C feasible for any vector in H_C, whereas any other coalition D is so ineffective that no vector in H dominates any other via D. The reader is referred to Figure 1 for an illustration of this construction, which we will formally define in the proof below. Theorem 1 establishes that the dominance relation on A as induced by this game coincides with R.

Theorem 1. *Let R be an irreflexive relation on a finite set A of outcomes χ a comprehension condition. Then, R is induced as the dominance relation of some finite coalitional NTU game under χ.*

Proof. By virtue of Lemma 1, we may assume without loss of generality that χ is tight. We define a finite NTU game $V_R = (N, H, V)$ as follows. With each $a \in A$ we associate two players i_a and j_a, and say that $\{i_a, j_a\}$ are *a pair* and that i_a and j_a are *partners*. Formally, $N = \{1, \ldots, 2 \cdot |A|\}$ and let $\{X_1, \ldots, X_{|A|}\}$ a partitioning of N



$$\begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccccccc} i_a & j_a & i_b & j_b & i_c & j_c & i_d & j_d \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \end{array}$$

Figure 1: The graph of an irreflexive relation that is induced by a coalitional NTU game (N, H, V) where N is given by $\{i_a, j_a, i_b, j_b, i_c, j_c, i_d, j_d\}$. The players' preferences are given by the utility matrix U. For reasons of readibility we depicted the transposal $U^\top$ of U on the right.

with $|X_k| = 2$ for all $1 \leq k \leq |A|$. We associate each X_k with outcome a_k and write $X_k = \{i_{a_k}, j_{a_k}\}$. Let $U = (u_{ij})_{i\in N, j\in\{1,\dots,|A|\}}$ be the $|N| \times |A|$ utility matrix such that for all $a, b \in A$,

$$u_{\{i_a,j_a\}}(b) = \begin{cases} (1,1) & \text{if } a = b, \\ (0,0) & \text{if } a \neq b \text{ and } aRb, \\ (0,1) & \text{otherwise.} \end{cases}$$

Set $H = H(U) = \{u(a)\colon\ a \in A\}$. Observe that $u(a) = u(b)$ if and only if $a = b$. Hence, $|A| = |H|$. For each player i, let $\tilde{u}^i \in H$ denote some outcome such that $\tilde{u}^i_i \in \min_{u\in H} u_i$. Since H is finite such a "minimal" outcome $\tilde{u}^i$ exists for each player i. Now, we define the characteristic function V such that for each coalition C in N,

$$V(C) = \begin{cases} H_C & \text{if } \{i_a, j_a\} \subseteq C, \text{ for some } a \in A, \\ \{\tilde{u}^i_C\colon\ i \in C\} & \text{otherwise.} \end{cases}$$

It now suffices to show for arbitrary $a, b \in A$ that aRb if and only if $a \succ b$. First assume that aRb. Observe that, by construction, $u_{\{i_a,j_a\}}(a) = (1,1)$ and $u_{\{i_a,j_a\}}(b) = (0,0)$. Moreover, $(1,1) \in V(\{i_a, j_a\})$. Accordingly, $u(a)$ dominates $u(b)$ via the coalition $\{i_a, j_a\}$ and we may conclude that $u(a) \succ u(b)$. Hence $a \succ b$.

For the opposite direction, assume that $a \succ b$ and, for a contradiction, that not aRb. Thus, there is some coalition C and some $x_C \in V(C)$ such that $x_C \geq u_C(a) > u_C(b)$. Because V_R is binary we have $x_i = u_i(a) = 1$ and $u_i(b) = 0$ for each $i \in C$. Because not aRb, by construction, $u_{j_a}(a) = 1$ and $u_{j_a}(b) = 1$. Hence, $j_a \notin C$. Observe, however, that $\tilde{u}^i_i = 0$ for each player $i \in C$. Since $x_C \in V(C)$, it follows that $i_c, j_c \in C$, for some $c \in A$. But then, however, $u_{\{i_c,j_c\}}(a) = (1,1)$. By definition of U, it follows that $c = a$ and, in particular, that $j_a \in C$, a contradiction. □

It is readily appreciated that the finite NTU game V_R constructed in the proof above is binary, ordinary and monotonic. Moreover, the construction used in the proof of Lemma 1 can easily be seen to preserve these properties. Accordingly, under every comprehension condition there is some finite NTU game that corresponds to some irreflexive binary relation and, moreover, satisfies these natural properties.

4 Coalitional Effectivity in Non-cooperative Games

Although in the formal definition of a coalitional NTU game the players' strategies are abstracted away from, they are still implicit in the characteristic function. A coalition is assumed to be effective for a particular utility vector if its members have a joint strategy that guarantees all of them the utility specified in that vector. However, the question keeps lingering how this guarantee should be given a formal and precise interpretation. In a setting without transferable utility, α- and *β-effectivity* provide two standard ways of determining the value of a coalition in a non-cooperative game [see, e.g., 2, 3, 4]. After having introduced the appropriate formal definitions of a non-cooperative game in normal form and those of α-effectivity and β-effectivity, we show that irreflexivity characterizes dominance relations of NTU games obtained through β-effectivity. Dominance relations of NTU games obtained through α-effectivity, however, are subject to more restrictive constraints.

Definition 3 (Normal form games)**.** A *game G in normal form* is a tuple (N, S, Ω, g, U), where N is a set $\{1, \ldots, |N|\}$ of players, $S = \bigtimes_{i \in N} S_i$ is an $|N|$-dimensional space of strategy profiles, $\Omega = \{\omega_1, \ldots, \omega_{|\Omega|}\}$ a set of outcomes, and $g\colon S \to \Omega$ an *outcome function* associating each strategy profile s with an outcome $g(s)$ in Ω. Finally, $U = (u_{ij})_{i \in N, j \in \Omega}$ is an $|N| \times |\Omega|$ utility matrix.

Observe that this definition makes no specific assumptions as to whether the players have mixed strategies at their disposal. We have (s_C, t_{-C}) denote the strategy profile s^* such that $s^*_i = s_i$ if $i \in C$ and $s^*_i = t_i$ if $i \notin C$.

A coalition C is said to be *α-effective* for a particular utility vector $x_C \in \mathbb{R}^C$, if in the normal form game coalition C has a joint strategy that guarantees each of its members i at least a utility of x_i, no matter which strategies the players not in C may adopt. By contrast, C is said to be β-effective for a particular utility vector $x_C \in \mathbb{R}^C$, if the players that are not in C have no joint strategy that precludes the coalition C from obtaining a utility of at least x_i to each of its members i.

Definition 4 (α-effectivity and β-effectivity)**.** Let $G = (N, S, \Omega, g, U)$ be a game in normal form, C a coalition in N and $x_C \in \mathbb{R}^C$. Then,

C is α-effective for x_C if there is an $s \in S$ such that for all $t \in S$, $u_C(g(s_C, t_{-C})) \geq x_C$,
C is β-effective for x_C if for all $s \in S$, there is a $t \in S$ such that $u_C(g(t_C, s_{-C})) \geq x_C$.

For $\gamma \in \{\alpha, \beta\}$ and χ a comprehension condition, a finite coalitional NTU game (N, H, V) is said to *γ-correspond to a normal form game $G = (N, S, \Omega, g, U)$ under χ*, whenever $H = \{u(g(s)) \in \mathbb{R}^N\colon\ s \in S\}$ and for each coalition C in N,

$$V(C) = \{x_C \in \chi(H_C)\colon C \text{ is } \gamma\text{-effective for } x_C\}.$$

Also, if a binary relation R on a set A can be induced as the dominance relation of some finite coalitional NTU game under χ that γ-corresponds to some normal

form game, we say that R is *obtainable through γ-effectivity under χ*. If the comprehension condition χ is clear from the context, we usually omit the reference to χ.

The following example concerns a class of normal-form games that evince a particularly sharp contrast between the sets of outcome for which coalition are α- and β-effective.

Example 1. Let $\Omega = \{\omega_1, \ldots, \omega_{|\Omega|}\}$ be a set of outcomes, N a set of players, and $k \in \mathbb{N}$. Let further $\phi\colon \mathbb{N}^N \to \mathbb{N}$ such that for each $x \in \mathbb{N}^N$, $\phi(x) = 1 + (\sum_{i \in N} x_i \bmod |\Omega|)$. The *modulo game* $M(\Omega, k) = (N, S, \Omega, g, U)$ *on* Ω *and* k is then a game in normal form such that for each player i, $S_i = \{1, \ldots, k\}$ and g such that for each strategy profile $s \in S$, $g(s) = a_k$ if and only if $k = \phi(s)$. Obviously, if $k \geq |\Omega|$, for every $\omega_m \in \Omega$ and every joint strategy s_{-C} of its non-members, every coalition C has a strategy t_C that yields a_m as the outcome of the modulo game $M(\Omega, k)$. Merely set t_C such that $m = 1 + ((\sum_{i \in C} t_i + \sum_{i \notin C} s_i) \bmod |\Omega|)$. As this is always possible, every coalition is β-effective for every outcome in Ω, whereas every coalition other than the grand coalition N is only α-effective for outcomes in Ω that minimize the utility of at least one of its members.

As for the general case, we find that the structure of dominance relations induced by finite NTU games do not in an important sense depend on the comprehension condition assumed. Rather, in order to establish the characterizing structural properties of the dominance relations obtained through either α- or β-effectivity, we can assume the comprehension condition to be tight.

Lemma 2. *Let $\gamma \in \{\alpha, \beta\}$, R be a binary relation on a finite set A and χ a comprehension condition. Then, if R is obtainable trough γ-effectivity under a tight comprehension condition, R is also obtainable through γ-effectivity under χ.*

Proof. Let (N, H, V) and $G = (N, S, A, g, U)$ be such that aRb if and only if $u(a)$ dominates $u(b)$ in (N, H, V) and (N, V, H) β-corresponds to G under the tight comprehension condition. Let further $V'(C) = \{x_C \in \chi(H_C)\colon C \text{ is } \gamma\text{-effective for } x_C \text{ in } G\}$ for all coalitions C. Consider arbitrary $a, b \in A$ and let $\succ'$ denote the dominance relation of (N, H, V'). First assume aRb. Then, $u(a) \succ_C u(b)$ for some coalition C, where $\succ$ is the dominance relation of (N, H, V). Hence, there is some $x_C \in V(C)$ such that $x_C \geq u(a)_C > u(b)_C$. Observe that $V(C) \subseteq V'(C)$ and so $u_C(a) \in V'(C)$. It follows that $u_C(a) \succ' u_C(b)$ as well. For the opposite direction assume $u(a) \succ' u(b)$. Then, there is some $x_C \in \chi(H)$ with $x_C \geq u_C(a) > u_C(b)$ and C is γ-effective for x_C. Because both α- and β-effectivity of x_C are defined in terms of the existence of strategy profiles s such that $u_C(g(s)) \geq x_C$, some reflection reveals that C is now also γ-effective for $u_C(a)$ and aRb follows. □

Characteristic functions based on either α-effectivity or β-effectivity are perforce monotonic. However, if no restrictions are imposed on the comprehension

$$\begin{bmatrix} (1,0,0) & (1,2,0) \\ (0,0,0) & (0,1,0) \end{bmatrix} \quad \begin{bmatrix} (1,0,0) & (2,1,0) \\ (0,0,0) & (0,1,0) \end{bmatrix}$$

Figure 2: A three-player game, in which player 1 chooses rows, player 2 chooses columns and player 3 chooses matrices, showing that if the comprehension condition is tight, α-effectivity does not generally imply superadditivity.

conditions, superadditivity does not generally hold for characteristic functions obtained through either α- or β-effectivity. Even if comprehensiveness is being assumed, only α-effectivity guarantees superadditivity.[3]

Example 2. Consider the normal form game depicted in Figure 2. Then, if the comprehension condition is tight, we find that both player 1 and player 2 are α-effective for 1, Player 1 in virtue of the top row, player 2 because of the right column. Yet, $V(\{1,2\}) = \{(0,0),(1,0),(0,1)\}$, i.e., there is no outcome enforceable by the coalition consisting of both player 1 and player 2 which yields a utility of at least 1 to both players. It is worth observing that if $(1,2,0)$ and $(2,1,0)$ were both replaced by $(1,1,0)$, player 1 and 2 together could enforce an outcome that guarantees a utility of 1 to both of them and superadditivity would have been satisfied. If comprehensiveness is assumed, however, the coalition $\{1,2\}$ is effective for $(1,1)$. This is so because, in that case, both $(1,1) \in \chi(H_C)$ and player 1 and 2 could enforce an outcome with at least a utility of 1 to both players by playing top row and right column.

Intuitively, superadditivity is a particularly natural property in the context of α-effectivity. If two disjoint coalitions C and D can guarantee particular utilities to their members by playing particular strategies, then each member of either coalition should also be guaranteed that utility, if both coalitions play those strategies simultaneously. Accordingly, examples like the above suggest that tight comprehension conditions are conceptually dubious in the context of α-effectivity and that looser ones are more appropriate. As the topic does not affect the issues at hand, we will not pursue it here. Moreover, our model also allows for comprehension conditions, e.g., full comprehensiveness, that do guarantee superadditivity of NTU games obtained through α-effectivity. Thus, we rather point at another important structural property related to superadditivity that all finite NTU games obtained through α-effectivity do satisfy. We will call a finite NTU game (N, H, V) *α-consistent* if, for disjoint coalitions C and D, $x_C \in V(C)$ and $y_D \in V(D)$ imply that there is some $u \in H$ such that $u_{C\cup D} \geq (x_C, y_D)$. Superadditivity is stronger than α-consistency in that it additionally requires the coalition $C \cup D$ also to be effec-

[3]This follows from Bergin and Duggan's characterization of NTU games supported by α- and β-effectivity in the comprehensive settting. However, if, for each coalition C, $V(C)$ is compact and convex, superadditivity is also satisfied for NTU games obtained through β-effectivity. The proof of this result is non-trivial and relies on Kakutani's fixed point theorem [2, 4]. Also, in settings where comprehensiveness is assumed, this follows from Bergin and Duggan's Theorem 2.

tive for $u_{C\cup D}$. We find that every finite NTU game obtained through α-effectivity satisfies the weaker property of α-consistency.

Lemma 3. *Let χ be a comprehension condition. Then, every finite NTU game that α-corresponds to a normal form game under χ is α-consistent.*

Proof. Let (N, H, V) be an arbitrary finite NTU game and $G = (N, S, \Omega, g, U)$ be an equally arbitrary normal form game such that (N, H, V) α-corresponds to G. Let C and D be disjoint coalitions in N with $x_C \in V(C)$ and $y_D \in V(D)$. Then there are strategy profiles $s, t \in S$ such that for all $r \in S$ both $u(g(s_C, r_{-C})) \geq x_C$ and $u(g(t_D, r_{-D})) \geq y_D$. Let $\tilde{s}$ be defined such that $\tilde{s}_C = s_C$ and $\tilde{s}_D = t_D$. Then, for all $r \in S$, $u(g(\tilde{s}_{C\cup D}, r_{-(C\cup D)})) \geq (x_C, y_D)$. Then observe that $u(g(\tilde{s}_{C\cup D}, r_{-(C\cup D)}))$ is in H, and a fortiori also in $\chi(H)$, which concludes the proof. □

5 Dominance Relations through Coalitional Effectivity

The finite NTU games that are obtainable through α- and β-effectivity constitute two distinct subclasses of coalitional games. Theorem 2 shows that a restriction to the latter class of games does not impose additional constraints on the dominance relations that are obtainable. On the other hand, we find α-effectivity only yields dominance relations that also satisfy the *edge-mapping property*, which is defined in Section 5.2.

5.1 Dominance Relations through β-Effectivity

We are now in a position to state and prove our second result, which says that every irreflexive relation on a set of outcomes A is obtainable through β-effectivity and vice versa.

Theorem 2. *Let R be a binary relation on a finite set A and χ an arbitrary comprehension condition. Then, R is obtainable through β-effectivity under χ if and only if R is irreflexive.*

Proof. The only-if direction is trivial as the dominance relation of any finite NTU game is irreflexive.

For the opposite direction, observe that, by virtue of Lemma 2, it suffices to give the proof for the case in which χ is tight. So consider an arbitrary irreflexive relation R on a set A along with the finite NTU game $V_R = (N, H, V)$ as defined in the proof of Theorem 1. It then suffices to prove that V_R can be obtained through β-effectivity. To this end we construct a game $G^R_\beta = (N, S, \Omega, g, U)$, with $\Omega = A = \{a_1, \ldots, a_{|A|}\}$, $N = \{i_a, j_a \colon a \in A\}$, and $U = (u_{ij})_{i\in N, j\in A}$ as in the proof of Theorem 1. Then, with χ assumed to be tight, for each coalition C in N,

$$V(C) = \begin{cases} H_C & \text{if } \{i_a, j_a\} \subseteq C, \text{ for some } a \in A, \\ \{\tilde{u}^i_C \colon\ i \in C\} & \text{otherwise.} \end{cases}$$

For each player $i \in N$ we now define an outcome $\tilde{a}_i \in A$ as follows. Let $a \in A$ be such that $i \in \{i_a, j_a\}$. Then set

$$\tilde{a}_i = a_k, \quad \text{where} \quad k = \begin{cases} \min\{m\colon aRa_m\} & \text{if } aRb \text{ for some } b \in A, \\ \min\{m\colon a_m \neq a\} & \text{otherwise.} \end{cases}$$

The definition of $\tilde{a}_i$ has been chosen in such a way that $u_i(\tilde{a}_i) = \tilde{u}_i(i)$. Thus, without loss of generality, we may assume that $\tilde{u}_i^i = u_i(\tilde{a}_i)$ for each $i \in N$. Also observe that $\tilde{a}_{i_a} = \tilde{a}_{j_a}$ for each $a \in A$.

Further, for each player i in N, we have $S_i = \{0, 1\} \times \{0, 1\} \times \{1, \ldots, |A|\}$, with representative element $s_i = (s_i^1, s_i^2, s_i^3)$. This leaves us with the definition of the outcome function g. Suppose strategy profile s is played. Intuitively, the coalition $C(s)$ consisting of all pairs i and j with $s_i^1 = s_j^1 = 1$ is then formed. Formally we define

$$C(s) = \bigcup_{a \in A} \{i \in \{i_a, j_a\}\colon s_{i_a}^1 = s_{j_a}^1 = 1\}.$$

The members of $C(s)$ then decide whether all players in N continue to play the modulo game $M(A, |A|)$ or the modulo game $M(\{\tilde{a}_i\colon i \notin C(s)\}, |A|)$. The latter is played if $s_i^2 = 0$ for all $i \in C(s)$ and $C(s) \neq N$, the former, otherwise. Observe that this also covers the case in which $C(s) = \emptyset$. Irrespective of which modulo game is played, the outcome is then determined by $s^3 = (s_i^3)_{i \in N}$. Accordingly, let for each $B \subseteq A$ the function $\phi_B\colon \mathbb{N}^N \to N$ be defined such that for $x = (x_i)_{i \in N} \in \mathbb{N}^N$,

$$\phi_B(x) = 1 + \left(\sum_{i \in N} x_i \bmod |B| \right).$$

Formally define the outcome function g, such that for all strategy profiles $s \in S$,

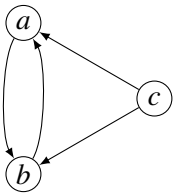
$$g(s) = a_m, \quad \text{where} \quad m = \begin{cases} \phi_A(s^3) & \text{if } C(s) = N \text{ or } s_i^2 = 1 \text{ for some } i \in C(s), \\ \phi_B(s^3) & \text{where } B = \{\tilde{a}_i\colon i \notin C(s)\}\text{, otherwise.} \end{cases}$$

Let for each coalition C in N,

$$V_\beta(C) = \{x_C \in \chi(H_C)\colon C \text{ is } \beta\text{-effective for } x_C \text{ in } G_\beta^R\}.$$

We first prove that $H = \{u(g(s))\colon s \in S\}$. As it is obvious that $\{u(g(s))\colon s \in S\} \subseteq H$, consider an arbitrary $u \in H$. Then $u = u(a_m)$ for some $1 \leq m \leq |A|$. Now consider the strategy profile $(s_i)_{i \in N}$ such that $s_i = (1, 1, m)$ if $i = 1$ and $s_i = (1, 1, |A|)$, otherwise. Then $C(s) = N$ and, informally, the modulo game $M(A, |A|)$ is played. Hence, $g(s) = \phi_A(s^3) = a_m$ and $u(a_m) \in \{u(g(s))\colon s \in S\}$, as desired.

It remains to be shown that that for each coalition C, $V(C) = V_\beta(C)$. To this end, first consider an arbitrary coalition C containing a pair i and j. Also consider an arbitrary $u \in H$ along with an equally arbitrary strategy profile $s = (s_i)_{i \in N}$. Without loss of generality we may assume $u = u(a_m)$, for some $1 \leq m \leq |A|$. Let $t = (t_i)_{i \in N}$ be such that $t_i = (1, 1, s_i^3)$ and $t_j = (1, 1, m')$, with m' satisfying $m' + 1 + ((\sum_{i \neq j} s_i) \bmod |A|) = m$. Set $\tilde{s} = (s_{-\{i,j\}}, t_{\{i,j\}})$. Informally, then the modulo



$$\begin{bmatrix} (1,1) & (1,1) & (1,1) \\ (1,1) & (1,0) & (0,1) \\ (1,1) & (0,1) & (1,0) \end{bmatrix}$$

Figure 3: Example showing that it can occur that two outcomes dominate each another in a graph obtained through α-effectivity, where $u(a) = (1,0)$, $u(b) = (0,1)$ and $u(c) = (1,1)$. In the two-player non-cooperative game depicted on the right player 1 chooses rows and player 2 columns.

game $M(A, |A|)$ is played. Accordingly, $\phi_A(\tilde{s}^3) = m$ and $g(\tilde{s}) = a_m$. With u having been chosen arbitrarily, it follows that $V_\beta(C) = \{u_C \colon u \in H\} = V_C$.

Finally, consider an arbitrary coalition C that contains no pairs, i.e., for no $a \in A$, $\{i_a, j_a\} \subseteq C$. Also consider an arbitrary $x_C \in \mathbb{R}^C$. First assume that $x_C \in V(C)$. We prove that also $x_C \in V_\beta(C)$. By virtue of monotonicity, we may assume without loss of generality that $C = \{i^*\}$ for some $i^* \in N$. Then, $x_C = x_{i^*} = u_{i^*}(\tilde{a}_{i^*})$. Also, $\tilde{a}_{i^*} = a_m$ for some $1 \leq m \leq |A|$. Let s be an arbitrary strategy profile. Without loss of generality we may assume that $s^1_{i^*} = 0$. Then, there is some $B \subseteq A$ such that $a_m \in B$ and $g(s) = a_k$ where $k = \phi_B(s^3)$. Let m' be such that $m' + 1 + ((\sum_{i \neq j} s^3_i) \bmod |B|) = m$ and set $t_{i^*} = (0, 0, m')$. Then, $g(t_{i^*}, s_{-i^*}) = a_m$, as desired.

For the other direction, assume $x_C \notin V(C)$. Without loss of generality we may assume that $x_C = u_C(a_m)$ for some $1 \leq m \leq |A|$. We may also assume that $a_m = \tilde{a}_i$ for no $i \in C$ and $u_i(a_m) > u_i(\tilde{a}_i)$ for some $i \in C$. Let D be the set partners of the members in C, i.e., $D = \{j \in N \colon j \text{ is the partner of some } i \in C\}$. As C contains no pair, C and D are obviously disjoint. Further, let $E = N \setminus (C \cup D)$. Thus, E only contains pairs, i.e., if i and j are partners, then $i \in E$ if and only if $j \in E$. Also observe that $\{\tilde{a}_i \colon i \in C\} = \{\tilde{a}_i \colon i \in D\}$. Hence, $a_m \notin \{\tilde{a}_i \colon i \in C \cup D\}$. Let s be the strategy profile such that $s_i = (1,1,1)$ for all $i \in E$. Then, informally, no matter which strategy $C \cup D$ adopts, the modulo game $M(\{\tilde{a}_i \colon i \in C \cup D\}, |A|)$ is played. Formally, for all $t \in S$, $g(s_E, t_{-E}) \in \{\tilde{a}_i \colon i \in C \cup D\}$. Hence, $g(s_E, t_{-E}) \neq a_m$ and it follows that $x_C = u_C(a_m) \notin V_\beta(C)$, which concludes the proof. □

5.2 Edge-Mappings and the Edge-Mapping Property

For dominance graphs obtained through α-effectivity matters are slightly more complicated than for those obtained through β-effectivity. For instance, it is not the case that every irreflexive dominance relation can be obtained through α-effectivity. Consider for instance the dominance graph on two alternatives a and b such that a and b dominate one another, i.e., $a \succ b$ as well as $b \succ a$. Now assume for a contradiction that this graph can be induced through α-effectivity. As $a \succ b$, there must be some coalition C that is α-effective for outcome a and such that all of its mem-

bers strictly prefer outcome a to outcome b, i.e., $u_C(a) > u_C(b)$. Similarly, because $b \succ a$ there is some coalition D that is α-effective for outcome b and $u_D(b) > u_D(a)$. It follows that C and D are disjoint. Moreover, by α-consistency, there is some outcome c such that $u_{C\cup D}(c) \geq (u_C(a), u_D(b))$. Clearly, c has to be distinct from both a and b, a contradiction.

On the other hand, a dominance graph containing alternatives that dominate one another does not preclude that dominance relation being obtainable through α-effectivity. Consider, for instance, the dominance graph on three alternatives, a, b and c, depicted in Figure 3. There the alternatives a and b dominate one another. Nevertheless, the graph is obtainable through α-effectivity from the non-cooperative game depicted on the right.

Intuitively, this can be understood as follows. Two outcomes a and b dominating one another indicates that there are two disjoint coalitions C and D such that C is α-effective for $u_C(a)$ the former whereas D is α-effective for $u_D(b)$. Accordingly, both C and D have strategies that guarantee these utility vectors to themselves, respectively, no matter which strategies the other players adopt. Moreover, C and D being disjoint, they can play these strategies simultaneously. If they do so some outcome c results with both $u_C(c) \geq u_C(a)$ and $u_D(c) \geq u_D(b)$. However, c has to be distinct from both a and b as both $u_C(a) > u_C(b)$ and $u_D(b) > u_D(a)$.

If a dominance relation $\succ$ is obtained through α-effectivity from a normal form game, it is worth remarking that $a \succ_C b$ does not so much mean that coalition C has a strategy that, no matter what strategies the other players adopt, a is the outcome of the game. Rather, $a \succ_C b$ signifies that in the normal form game coalition C has a strategy at her disposal which guarantees, irrespective of the strategies the other players adopt, that the outcome *falls within a set of outcomes* each outcome of which is at least as good for the members of C as a and strictly better than b. (There may even be several such strategies for C, but we leave this issue aside.) Accordingly, with each *edge* (a, b) *in* $\succ$, i.e., each pair $a, b \in A$ such that $a \succ b$, it thus is possible to associate a coalition C, a *witnessing coalition*, along with such a set of outcomes that are at least as good as a and strictly better than b for all members in C. Moreover, for all distinct edges (a, b) and (c, d) with disjoint witnessing coalitions these sets should have a non-empty intersection. Otherwise the witnessing coalitions of (a, b) and (c, d) could each play a strategy that guarantees the outcome to fall within disjoint sets, an absurdity. Accordingly, if a binary relation R is obtainable via α effectivity, it must at least be possible to associate with each edge (x, y) a set of outcomes containing x but not y. Moreover, any such so-called *edge mapping* has to satisfy a number of consistency conditions. For instance, if both (a, b) and (b, a) are edges in R they have disjoint witnessing coalitions and the edge mapping should duly associate overlapping sets of outcomes with (a, b) and (b, a). The full set of consistency conditions is summarized in the *edge-mapping property (EMP)* below. We then find that binary relations obtainable through α-effectivity are completely characterized by irreflexivity and the edge-mapping property.

Formally, we define an *edge mapping* for a given irreflexive binary relation R

on a set A as a function $\psi\colon A \times A \to 2^A$ such that for each *edge* $(a, b) \in R$ we have $a \in \psi(a, b)$ and $b \notin \psi(a, b)$. Observe that an edge mapping ψ is not in general commutative, i.e., it does not generally hold that $\psi(a, b) = \psi(b, a)$. Given an edge mapping ψ for R we say that two edges (a, b) and (c, d) in R are *ψ-exclusive* whenever at least one of the following two conditions holds:

(*i*) $\{a, b\} \cap \psi(c, d) \neq \emptyset$ and $d \in \psi(a, b)$,

(*ii*) $\{c, d\} \cap \psi(a, b) \neq \emptyset$ and $b \in \psi(c, d)$.

In this context it is worth observing that for every asymmetric relation there is an edge mapping ψ such that no two edges are ψ-exclusive. Merely set $\psi(a, b) = \{a\}$. Then, for all edges (a, b) and (c, d) such that $\{a, b\} \cap \psi(c, d) \neq \emptyset$, either $a = c$ or $b = c$. If the former we have $d \neq a$ immediately, if the latter this follows by asymmetry. In either case $d \notin \psi(a, b)$ (also see Corollary 1, below). On the other hand, for any alternatives a, b in A, if both $(a, b) \in R$ and $(b, a) \in R$, (a, b) and (b, a) are ψ-exclusive for any edge mapping ψ. Intuitively, (a, b) and (c, d) being ψ-exclusive means that the witnessing coalitions of (a, b) and (c, d) cannot be other than disjoint given ψ. To appreciate this, recall the intuitive interpretation of $\psi(a, b)$ as a set of outcomes that is at least as desirable as a and strictly more desirable than b for the members of the coalition witnessing (a, b). Accordingly, $\{a, b\} \cap \psi(c, d) \neq \emptyset$ implies that the coalition witnessing (a, b) strictly prefers a to d. If also $d \in \psi(c, d)$ this means that the witnessing coalitions of (a, b) and (c, d) have to be disjoint. An analogous argument holds for (*ii*). This idea is made precise in Lemma 5, below.

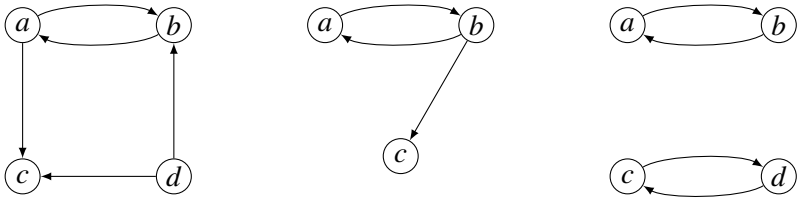
The *edge-mapping property (EMP)* is then defined as follows.

Definition 5 (Edge-mapping property, EMP)**.** Let $R \subseteq A \times A$ a binary relation on a set A. R is said to satisfy the *edge-mapping property (EMP)* if an edge mapping $\psi\colon A \times A \to 2^A$ exists such that $\bigcap_{(a,b)\in R'} \psi(a, b) \neq \emptyset$ for each subset $R' \subseteq R$ of which the edges are pairwise ψ-exclusive.

Informally, the edge-mapping property guarantees that disjoint witnessing coalitions under the edge-mapping ψ cannot force the game to end in different outcomes. Also take notice of the fact that every asymmetric relation vacuously satisfies the edge-mapping property in virtue of the edge-mapping that maps each edge (a, b) to $\{a\}$.

Example 3. Consider the three binary relations, R_1, R_2 and R_3 depicted in Figure 4. Only R_1 satisfies the edge-mapping property (EMP) in virtue of the edge mapping ψ summarized in the table below.

(x, y)	$\psi(x, y)$
(a, b)	$\{a, d\}$
(a, c)	$\{a, d\}$
(b, a)	$\{b, d\}$
(d, b)	$\{d\}$
(d, c)	$\{d\}$



(R_1) (R_2) (R_3)

Figure 4: Of the three binary relations depicted, only R_1 satisfies the edge-mapping property.

Then $d \in \psi(x, y)$ for all $x, y \in \{a, b, c, d\}$ and, therefore, cannot fail to satisfy EMP. In R_2, the edges (a, b) and (b, a) are ψ-exclusive for every edge mapping ψ. Suppose ψ were an edge mapping with respect to which R_2 satisfies EMP. Accordingly, $\psi(a, b) = \{a, c\}$, $\psi(b, a) = \{b, c\}$, and $b \in \psi(b, c)$. Now (a, b) and (b, c) are ψ-exclusive because $c \in \psi(a, b)$ and $b \in \psi(b, c)$ whereas (b, a) and (b, c) are ψ-exclusive because $b \in \psi(b, c)$ and $c \in \psi(b, a)$. As, however, $a \notin \psi(b, a)$, $b \notin \psi(a, b)$ and $c \notin \psi(b, c)$, $\psi(a, b) \cap \psi(b, a) \cap (b, c) = \emptyset$. Hence, R_2 does not satisfy EMP. We leave it to the avid reader to verify that R_3 does not satisfy EMP either.

We find that the dominance relation of every α-consistent finite NTU game satisfies the edge-mapping property.

Lemma 4. *The dominance relation of every α-consistent finite NTU game satisfies the edge-mapping property.*

Proof. Let $\succ$ be the dominance relation of an α-consistent finite NTU game (N, H, V). For all $u, v \in H$ with $u \succ v$, there is some coalition $C(u, v)$ such that $u_{C(u,v)} > v_{C(u,v)}$. Moreover, there is some $x \in V(C(u, v))$ with $x_{C(u,v)} \geq u_{C(u,v)} > v_{C(u,v)}$. Now define $\psi\colon H \times H \to 2^H$ such that, for all $u, v \in H$,

$$\psi(u, v) = \{x \in H\colon\ x_{C(u,v)} \geq u_{C(u,v)} > v_{C(u,v)}\}.$$

Obviously, ψ is an edge mapping for the dominance relation $\succ$, as for all $u, v \in H$ we have $u \in \psi(u, v)$ and $v \notin \psi(u, v)$. Let $u^1, v^1, \ldots, u^m, v^m \in H$ such that $(u^1, v^1), \ldots, (u^m, v^m)$ are pairwise ψ-exclusive edges in the dominance relation $\succ$. For each $1 \leq k \leq m$, we have C_k denote the coalition $C(u^k, v^k)$.

First, we establish that the coalitions C_i and C_j are disjoint, for all $1 \leq i < j \leq m$. Without loss of generality we may assume that $\{u^i, v^i\} \cap \psi(u^j, v^j) \neq \emptyset$ and $v^j \in \psi(u^i, v^i)$. Now observe that the former implies that $u^i_{C_j} > v^j_{C_j}$ or $v^i_{C_j} > v^j_{C_j}$, whereas the latter yields both $v^j_{C_i} \geq u^i_{C_i}$ and $v^j_{C_i} \geq v^i_{C_i}$. Hence, $C_i \cap C_j = \emptyset$. We now show that $\bigcap_{1 \leq i \leq m} \psi(u^i, v^i) \neq \emptyset$. For each $1 \leq k \leq m$ there is some $x_{C_k} \in V(C_k)$ such that $x_{C_k} \geq u^k_{C_k} > v^k_{C_k}$. Because $C_1, \ldots, C_k$ are pairwise disjoint and V is α-consistent, there is some $u^* \in H$ such that $u_{C_1 \cup \cdots \cup C_m} \geq x_{C_1 \cup \cdots \cup C_m}$. It follows that

$u^* \in \psi(u^i, v^i)$ for all $1 \leq i \leq m$. Hence, $\bigcap_{1 \leq i \leq m} \psi(u^i, v^i) \neq \emptyset$, which concludes the proof. □

Since every coalitional NTU game obtained through α-effectivity is α-consistent, Lemma 4 implies that the dominance relation of any such game satisfies the edge-mapping property.

5.3 Dominance Relations through α-Effectivity

In order to obtain a full characterization of the dominance relation that can be obtained through α-effectivity, we construct for each irreflexive relation R with the edge-mapping property a non-cooperative game G_α^R. We then show that R is induced as the dominance relation of the NTU game that α-corresponds to G_α^R. The players of this game are defined by the weak, i.e., reflexive, transitive and complete, orders over the outcome set A. Thus, by contrast to the constructions used in the proofs of Theorems 1 and 2, the number of players is exponential, rather than linear, in the number of vertices.

Let $\mathscr{P}(A)$ denote the set of weak orders P over the set A. We write $a \gtrsim_P b$ to signify that $(a, b) \in P$. Also $a \sim_P b$ denotes that both $a \gtrsim_P b$ and $b \gtrsim_P a$ in P, and $a >_P b$ that $a \gtrsim_P b$ but not $b \gtrsim_P a$. We omit the subscript whenever P is clear from the context. Then, define, for each $X \subseteq A$ with $a \in X$ and $b \notin X$,

$$C(X, a, b) = \{P \in \mathscr{P}(A) \colon x \gtrsim_P a >_P b \text{ for all } x \in X\}.$$

If an edge mapping ψ is fixed in the context, we will also denote $C(\psi(a, b), a, b)$ by $C(a, b)$. Before we give our characterization result, we first prove a lemma, establishing the exact conditions under which two coalitions $C(a, b)$ and $C(c, d)$ are disjoint.

Lemma 5. *Let ψ be an edge mapping for an irreflexive relation R on A. Then, for all $(a, b), (c, d) \in R$,*

$C(\psi(a, b), a, b) \cap C(\psi(c, d), c, d) = \emptyset$ *if and only if (a, b) and (c, d) are ψ-exclusive.*

Proof. For the if-direction assume (a, b) and (c, d) to be ψ-exclusive. Without loss of generality we may assume that $\{a, b\} \cap \psi(c, d) \neq \emptyset$ and $d \in \psi(a, b)$. From the former follows that either $a >_P d$ or $b >_P d$ for each $P \in C(\psi(c, d), c, d)$, whereas the latter yields both $d \gtrsim_P a$ and $d \gtrsim_P b$ for each $P \in C(\psi(a, b), a, b)$. Hence, $C(\psi(a, b), a, b) \cap C(\psi(c, d), c, d) = \emptyset$.

For the opposite direction assume that (a, b) and (c, d) are not ψ-exclusive. Then,

(*i*) $d \in \psi(a, b)$ implies $\{a, b\} \cap \psi(c, d) = \emptyset$, and

(*ii*) $b \in \psi(c, d)$ implies $\{c, d\} \cap \psi(a, b) = \emptyset$.

We distinguish three cases: (1) $d \in \psi(a,b)$, (2) $b \in \psi(c,d)$ and (3) neither $d \in \psi(a,b)$ nor $b \in \psi(c,d)$. First assume $d \in \psi(a,b)$. Then, $d \neq b$. By (i), moreover, $\{a,b\} \cap \psi(c,d) = \emptyset$. Hence, also $a \neq c$. Therefore, there exists a weak order P on A such that for all $x \in (\psi(a,b) \cup \psi(c,d)) \setminus \{a\}$,

$$x > a \sim d > b.$$

Observe that both $P \in C(a,b)$ and $P \in C(c,d)$. Hence, $C(a,b) \cap C(c,d) \neq \emptyset$. Case (2) is covered by an analogous argument.

Finally, if (3) obtains, we have $a \neq b$, $c \neq d$, $a \neq d$ as well as $b \neq c$. Moreover, $\{b,d\} \cap (\psi(a,b) \cup \psi(c,d)) = \emptyset$. It follows that there is a weak order P on A such that for all $x \in (\psi(a,b) \cup \psi(c,d)) \setminus \{a,c\}$,

$$x > a \sim c > b \sim d.$$

Then, both $P \in C(a,b)$ and $P \in C(c,d)$ and we may conclude $C(a,b) \cap C(c,d) \neq \emptyset$. □

We are now in a position to prove our characterization result for dominance graphs obtained through α-effectivity. The edge mapping property may appear a bit contrived. Even if that is the case, the important thing to observe is that it is a property of binary relations that is defined independently of their interpretation as dominance relations. Moreover, Theorem 3 can be used to obtain more intuitive results. Three of these are captured in Corollaries 1, 2, and 3.

Theorem 3. *Let R be a binary relation on a finite set A and χ a comprehension condition. Then, R is obtainable through α-effectivity under χ if and only if R is irreflexive and satisfies the edge-mapping property.*

Proof. The only-if direction is an immediate consequence of Lemma 3 and Lemma 4: every finite NTU game that α-corresponds to some normal form game under χ is α-consistent, and the dominance relation of every α-consistent finite NTU game satisfies the edge-mapping property.

For the opposite direction, assume R to satisfy the edge-mapping property and let $\psi\colon A \times A \to 2^A$ be the witnessing edge mapping. We first construct a normal form game $G^R_\alpha = (N, S, \Omega, g, U)$, where the set of players N is given by the set $\mathscr{P}(A)$ of weak orders over A and $\Omega = A = \{a_1, \ldots, a_{|A|}\}$. By virtue of Lemma 2 we may assume without loss of generality that χ is tight. Each player P defines his own preference relation over A, i.e., P weakly prefers a to b if and only if $(a,b) \in P$. We have the utility matrix $U = (u_{ij})_{i \in N, j \in A}$ represent these preferences. Let, moreover, set of strategies for each player $i \in N$ be defined as as $S_i = A \times A \times \{1, \ldots, |A|\}$, with typical element $s_i = (s_i^1, s_i^2, s_i^3)$.

This leaves us with the definition of the outcome function $g\colon S \to A$. For a better understanding, however, we first give an informal description of the game G^R_α and introduce a number of notational conventions. The game G^R_α can be understood as follows. By choosing the strategies s_i^1 and s_i^2 a player announces which

coalition $C(x, y)$, where $x, y \in A$ and xRy, he wishes to belong to. Only if all players of a coalition express the wish to belong to that very coalition, it is actually formed. In this way the simultaneous formation of overlapping coalitions is precluded. Then a modulo game $M(X, |A|)$ is played, the outcome of which is determined by $(s_i^3)_{i \in N}$. The resulting outcome is also the outcome of G_α^R. Accordingly each coalition $C(a, b)$ in $\mathscr{C}(R)$ can force the outcome of the game to fall within the set $\psi(a, b)$ by choosing an appropriate joint strategy which guarantees the members i of $C(a, b)$ a minimum utility of $u_i(a)$, no matter which strategies the other players adopt.

Let $\mathscr{C}(R) = \{C(a, b) \colon aRb\}$ and for each strategy profile $s = (s^1, s^2, s^3)$ in S. We say that $C(a, b)$ *forms at* s if $s_i^1 = a$ and $s_i^2 = b$ for all $i \in C(a, b)$. Now define $\mathscr{C}(s)$ as the set of coalitions in $\mathscr{C}(R)$ that form at s, i.e.,

$$\mathscr{C}(s) = \{C(a, b) \in \mathscr{C}(R) \colon C(a, b) \text{ forms at } s\}.$$

Defined thus, all coalitions in $\mathscr{C}(s)$ are pairwise disjoint. Moreover, by virtue of Lemma 5, for any $a, b, c, d \in A$ with $C(a, b), C(c, d) \in \mathscr{C}(s)$, the edges $(a, b), (c, d) \in R$ are ψ-exclusive. With each strategy profile $s = (s^1, s^2, s^3)$ we associate a set $X(s) \subseteq A$ of outcomes defined as,

$$X(s) = \bigcap \{\psi(a, b) \colon C(a, b) \in \mathscr{C}(s)\}.$$

We postulate that $X(s) = A$ in case $\mathscr{C}(s) = \emptyset$. As R satisfies the edge-mapping property, it follows that for each strategy profile s the set $X(s)$ is non-empty.

We are now in a position to formally define the outcome function g, such that for all strategy profiles $s = (s^1, s^2, s^3)$ in S,

$$g(s) = a_m, \quad \text{where} \quad m = 1 + \left(\sum_{i \in N} s_i^3 \bmod |X(s)| \right).$$

Accordingly, by merely forming, each coalition $C(a, b)$ in $\mathscr{C}(R)$ has a strategy that, no matter which strategies the other players adopt, guarantees the the outcome of G_α^R to fall within $\psi(a, b)$.

Let V_α^R be the finite coalitional NTU game (N, H, V_α) where $H = \{u(g(s)) \colon s \in S\}$ and for each coalition C in N,

$$V_\alpha(C) = \{x_C \in H_C \colon C \text{ is } \alpha\text{-effective for } x_C \text{ in } G_\alpha^R\}.$$

Obviously, V_α^R α-corresponds to G_α^R, so it remains to be shown that V_α^R induces R. Observe that $u(a) = u(b)$ if and only if $a = b$. The if-direction is trivial. For the other direction, observe that there is some weak order $P \in \mathscr{P}(A)$ such that $a > b$ in P. As $P \in N$ also $u_P(a) > u_P(b)$ signifying that $a \neq b$. Consequently, $|H| = |A|$ and it suffices to prove that for all $a, b \in A$, aRb if and only if $u(a) \succ u(b)$.

Consider arbitrary $a, b \in A$ and assume aRb. Then, $C(a, b) \in \mathscr{C}(R)$. Let the strategy profile s be defined such that $s_i = (a, b, 1)$ for all $i \in C(a, b)$ and let t

be an arbitrary strategy profile. Then $C(a,b)$ is formed at $s^* = (s_{C(a,b)}, t_{-C(a,b)})$, i.e., $C(a,b) \in \mathscr{C}(s^*)$. Hence, $X(s^*) \subseteq \psi(a,b)$ and $g(s^*) \in \psi(a,b)$. Moreover, $u_{C(a,b)}(g(s^*)) \geq u_{C(a,b)}(a) > u_{C(a,b)}(b)$. As t had been chosen arbitrarily, it follows that $u_{C(a,b)}(a) \in V_\alpha(C(a,b))$. Therefore, $u(a)$ dominates $u(b)$ via $C(a,b)$, which yields $u(a) \succ u(b)$.

For the opposite direction, assume that aRb does *not* hold for some $a, b \in A$. Then, $C(a,b)$ is not in $\mathscr{C}(R)$ and consequently never forms. Because the dominance relation of any finite NTU game is irreflexive, without loss of generality, we may assume that $a \neq b$. For a contradiction, moreover, assume that nevertheless $u(a) \succ u(b)$. Then, there is some coalition C in N and some $\tilde{s} \in S$ such that,

$$u_C(g(\tilde{s}_C, s_{-C})) \geq u_C(a) > u_C(b), \quad \text{for all for all } s \in S. \tag{$*$}$$

Consider this $\tilde{s}$ and observe that $C \neq N$, otherwise C would also have contained weak orders P with $u_P(b) \geq u_P(a)$. So, let i^* be a player with $i^* \notin C$. Observe that there are $c, d \in A$ with cRd such that $C(c,d) \subseteq C$. To appreciate this let $b = a_m$ and let s be such that $s_i = (a, b, x_i)$ and $1 + ((\sum_{i \in C} \tilde{s}^3 + \sum_{i \notin C} s^3) \bmod |A|) = m$. Because $C \neq N$, s is well-defined. If C contains no coalition $C(c,d)$ then $X(\tilde{s}_C, s_{-C}) = A$ and $g(\tilde{s}_C, s_{-C}) = a_m = b$. Consequently, $u_C(b) \geq u_C(g(\tilde{s}_C, s_{-C}))$, which is at variance with $(*)$.

We now show that $(c,d) = (a,b)$. Recall that $a \neq b$ and that, because R is irreflexive, also $c \neq b$. For a contradiction assume $(c,d) \neq (a,b)$. We distinguish the following cases: (*i*) $a \neq c$ and $b = d$, (*ii*) $a \neq c$ and $b \neq d$, and (*iii*) $a = c$ and $b \neq d$. If (*i*), $b \neq c$ and there is a weak order P_1 over A with $c > b = d > a$. Hence, $P_1 \in C(c,d)$. If (*ii*) there is a weak order P_2 over A with $c \sim b > a \sim d$. Now, $P_2 \in C(c,d)$. Finally, if (*iii*), both $c \neq b$ and $a \neq d$. Accordingly a weak order P_3 over A with $b > a = c > d$ exists. Moreover, $P_3 \in C(c,d)$. In all three cases together with $C(c,d) \subseteq C$, it follows that it is not the case that $u_C(a) > u_C(b)$, contradicting $(*)$.

Recall that cRd. Hence aRb as well, because $(a,b) = (c,d)$. This, however, contradicts our previous assumption. □

The following corollaries show how the edge-mapping can be employed. The first two can also easily be obtained by other means, but are included for illustrative purposes. Corollary 3 is slightly more substantial.

Corollary 1. *Let χ be a comprehension condition. Then, every asymmetric relation on a finite set A is obtainable through α-effectivity under χ.*

Proof. Define $\psi\colon A \times A \to 2^A$ such that $\psi(a,b) = \{a\}$ for all $(a,b) \in R$. Obviously, ψ is an edge mapping. Asymmetry of R, moreover, guarantees that no two edges in R are ψ-exclusive. Hence, R satisfies the edge-mapping property trivially. Theorem 3 then yields the desired result. □

Corollary 2. *Let χ be a comprehension condition and R an irreflexive relation on a finite set A such that there is some $a \in A$ with xRa for no $x \in A$. Then, R is obtainable through α-effectivity under χ.*

Proof. Define $\psi\colon A \times A \to 2^A$ so that $\psi(b,c) = \{a,b\}$ for all $(b,c) \in R$. Then, obviously, $b \in \psi(b,c)$. By irreflexivity of R, we have $c \neq b$ and since a is undominated, also $c \neq a$. Hence, $c \notin \psi(b,c)$ and we may conclude that ψ is an edge mapping for R. Now observe that $a \in \bigcap_{(b,c)\in R} \psi(b,c)$. Accordingly, the relation R has the edge-mapping property and Theorem 3 yields the desired result. □

Corollary 3. *Let χ be a comprehension condition and A a finite set of at least two alternatives. Then, the maximal irreflexive relation $R^u = \{(a,b) \in A \times A\colon\ a \neq b\}$ on A is not obtainable through α-effectivity under χ.*

Proof. Assume for a contradiction that R^u satisfies the edge-mapping property and let ψ be the witnessing edge mapping. We first prove by induction on k that for each $0 \leq k$ there are distinct $c_1, \ldots, c_k \in A$ such that the edges (a,b), (b,a), $(a,c_1), \ldots, (a,c_k)$ are ψ-exclusive. For $k = 0$ merely observe that (a,b) and (b,a) are ψ-exclusive as $a \in \psi(a,b)$ and $b \in \psi(b,a)$ by the definition of an edge mapping. For the induction step, assume that $a, b, c_1, \ldots, c_k \in A$ exist such that (a,b), (b,a), $(a,c_1), \ldots, (a,c_k)$ are pairwise ψ-exclusive. Having assumed R^u to satisfy the edge-mapping property in virtue of ψ, we have $\bigcap_{(x,y)\in X_k} \psi(x,y) \neq \emptyset$, where $X_k = \{(a,b),(b,a)\} \cup \{(a,c_i)\colon\ 1 \leq i \leq k\}$. Observe, however, that $a \notin \psi(b,a)$, $b \notin \psi(a,b)$ and $c_i \notin \psi(a,c_i)$ for each $1 \leq i \leq k$. Hence, there is some c_{k+1} distinct from $a, b, c_1, \ldots, c_k$ such that $c_{k+1} \in \psi(a,b)$, $c_{k+1} \in \psi(b,a)$ and $c_{k+1} \in \psi(a,c_i)$ for each $1 \leq i \leq k$. Now consider the edge (a,c_{k+1}). Obviously, $a \in \psi(a,c_{k+1})$. It follows that the edges $(a,b),(b,a),(a,c_1),\ldots,(a,c_{k+1})$ are pairwise ψ-exclusive.

To conclude, consider the case in which $k = |A|-2$. Then, $\{a,b,c_1,\ldots,c_{|A|-2}\} = A$ coincides with A and the edges $(a,b),(b,a),(a,c_1),\ldots,(a,c_{|A|-2})$ are pairwise ψ-exclusive. However, for each $x \in A$, $x \notin \psi(a,x)$. Hence, $\bigcap_{(x,y)\in X_{|A|-2}} \psi(x,y) = \emptyset$, contradicting the assumption that R^u satisfies the edge-mapping property. □

6 Conclusion

We characterized the structural restrictions of dominance relations in coalitional games that denote whether there is an effective coalition that unanimously prefers one outcome to another. We have shown that *any* irreflexive relation over a finite set can be obtained as the dominance relation of some ordinary, monotonic, and simple coalitional NTU game V, even if we require V to be induced by a *non-cooperative* game via β-effectivity. Dominance relations obtainable via α-effectivity are characterized by a more restrictive condition, which we refer to as the *edge-mapping property*.

Many well-known dominance-based solution concepts from coalitional game theory (e.g., the core or stable sets) lack existence, uniqueness, or even both. Social choice theory, on the other hand, has produced solution concepts—e.g., the Banks set, the uncovered set, or the minimal covering set—of which existence, uniqueness, and several other desirable properties are guaranteed for *asymmetric*

dominance relations on a finite set of alternatives. An important question for future work is whether there are extensions of these concepts that retain most of their attractive properties for dominance relations that are not anti-symmetric.

Acknowledgements

We thank Felix Fischer and three anonymous referees for valuable comments on previous versions of this paper. This material is based upon work supported by the Deutsche Forschungsgemeinschaft under grants BR 2312/3-1 and BR 2312/3-2. Preliminary results were presented at the 5th International Conference on Logic, Game Theory and Social Choice, at the Dagstuhl Seminar on Computational Issues in Social Choice and at the 8th Conference on Logic and the Foundations of Game and Decision Theory.

References

[1] J. Abdou and H. Keiding. *Effectivity Functions in Social Choice*. Kluwer Academic Publishers, 1991.

[2] R. J. Aumann. Acceptable points in general n-person games. In A. W. Tucker and R. D. Luce, editors, *Contributions to the Theory of Games IV*, volume 40 of *Annals of Mathematics Studies*, pages 287–324. Princeton University Press, 1959.

[3] R. J. Aumann. The core of a cooperative game without side payments. *Transactions of the American Mathematical Society*, 98:539–552, 1961.

[4] R. J. Aumann and B. Peleg. Von Neumann-Morgenstern solutions to cooperative games without side payments. *Bulletin of the American Society*, 66: 173–179, 1960.

[5] J. Bergin and J. Duggan. An implementation-theoretic approach to non-cooperative foundations. *Journal of Economic Theory*, 86:50–76, 1999.

[6] B. Dutta and J.-F. Laslier. Comparison functions and choice correspondences. *Social Choice and Welfare*, 16(4):513–532, 1999.

[7] R. Farquharson. *Theory of Voting*. Yale University Press, 1969.

[8] P. C. Fishburn. Condorcet social choice functions. *SIAM Journal on Applied Mathematics*, 33(3):469–489, 1977.

[9] S. Hart and M. MasColell. Bargaining and value. *Econometrica*, 64(2):357–380, 1996.

[10] E. Kalai and D. Samet. Monotonic solution concepts to general cooperative games. *Econometrica*, 53(2):307–328, 1985.

[11] Ö. Kıbrıs and M. R. Sertel. Bargaining over a finite set of alternatives. *Social Choice and Welfare*, 28:421–437, 2007.

[12] S. Lahiri. A weak bargaining set for contract choice problems. *Research in Economics*, 61(4):185–190, 2007.

[13] J.-F. Laslier. *Tournament Solutions and Majority Voting*. Springer-Verlag, 1997.

[14] R. D. Luce and H. Raiffa. *Games and Decisions: Introduction and Critical Survey*. John Wiley & Sons Inc., 1957.

[15] D. C. McGarvey. A theorem on the construction of voting paradoxes. *Econometrica*, 21(4):608–610, 1953.

[16] J. Nash. The bargaining problem. *Econometrica*, 18:155–162, 1950.

[17] J. Nash. Two-person cooperative games. *Econometrica*, 21:128–140, 1953.

[18] A. Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica*, 50 (1):97–109, 1982.

[19] T. Schwartz. Cyclic tournaments and cooperative majority voting: A solution. *Social Choice and Welfare*, 7:19–29, 1990.

[20] R. Serrano. A comment on the nash program and the theory of implementation. *Economic Letters*, 55:203–208, 1997.

[21] A. D. Taylor and W. S. Zwicker. *Simple Games. Desirability Relations, Trading, Pseudoweightings*. Princeton University Press, 1999.

[22] J. von Neumann and O. Morgenstern. *The Theory of Games and Economic Behavior*. Princeton University Press, 2nd edition, 1947.

AGM-consistent beliefs in branching time

Giacomo Bonanno
Department of Economics,
University of California,
Davis, CA 95616-8578 - USA
e-mail: gfbonanno@ucdavis.edu

January 2009

Abstract

In previous work belief change over time was modeled by means of branching-time structures; a corresponding modal logic with operators for next-time, information and belief was proposed and some aspects of the relationship between this logic and the AGM theory of belief revision were discussed. In this paper we establish a stronger correspondence between the semantics of temporal belief revision frames and AGM belief revision. The addition of a valuation to a temporal belief revision frame gives rise - for every state-instant pair (ω, t) - to a belief set K (at (ω, t)) and a partial belief revision function based on K (constructed from the beliefs at the immediate successors of t and at state ω). We investigate under what conditions such a partial belief revision function can be extended to a full AGM revision function. We find that a necessary and sufficient condition (when the set of states Ω is finite) is that there exist a total pre-order R of Ω that rationalizes belief revision at (ω, t), in the sense that at t and at the immediate successors of t (and at state ω) the states that the agent considers possible are the R-maximal states among the ones that are compatible with the information received. We also provide a set of axioms that characterizes this class of temporal belief revision frames.

Keywords: information, belief, branching time, AGM belief revision, plausibility ordering.

1 Introduction

In [4] and [5] belief change over time was modeled by means of branching-time structures; a corresponding modal logic with operators for next-time, information and belief was proposed and some aspects of the relationship between this logic and the AGM theory of belief revision ([1]) were discussed. In this paper we establish a stronger correspondence between the semantics of temporal belief revision frames and AGM belief revision. The addition of a valuation to a temporal belief revision frame gives

rise - for every state-instant pair (ω, t) - to a belief set K (the agent's beliefs at (ω, t)) and a partial belief revision function based on K (constructed from the agent's beliefs at the immediate successors of instant t and at state ω). We investigate under what conditions such a partial belief revision function can be extended to a full AGM revision function. We find that a necessary and sufficient condition (when the set of states Ω is finite) is that there exist a total pre-order R of Ω that rationalizes belief revision at (ω, t), in the sense that at instant t and at its immediate successors (and at state ω) the states that the agent considers possible are the R-maximal states among the ones that are compatible with the information received. We also provide a set of axioms in a modal logic that characterizes this class of temporal belief revision frames.

2 Temporal belief revision frames

A *next-time branching frame* is a pair $\langle T, \rightarrowtail \rangle$ where T is a set of instants or dates and $\rightarrowtail$ is a binary "next-time" relation on T satisfying the following properties: $\forall t_1, t_2, t_3 \in T$,

1. if $t_1 \rightarrowtail t_3$ and $t_2 \rightarrowtail t_3$ then $t_1 = t_2$,
2. if $\langle t_1, ..., t_n \rangle$ is a sequence with $t_i \rightarrowtail t_{i+1}$, for every $i = 1, ..., n-1$, then $t_n \neq t_1$.

The interpretation of $t_1 \rightarrowtail t_2$ is that t_2 is an *immediate successor* of t_1 or t_1 is the *immediate predecessor* of t_2: every instant has at most a unique immediate predecessor but can have several immediate successors. If $t \in T$ we denote the set of immediate successors of t by $t^{\rightarrowtail}$, that is, $t^{\rightarrowtail} = \{t' \in T : t \rightarrowtail t'\}$.

A *temporal belief-information structure* is a tuple $\langle T, \rightarrowtail, \Omega, \{\mathcal{B}_t, \mathcal{I}_t\}_{t \in T} \rangle$ where $\langle T, \rightarrowtail \rangle$ is a next-time branching frame, Ω is a set of states (or possible worlds) and, for every $t \in T$, $\mathcal{B}_t$ and $\mathcal{I}_t$ are binary relations on Ω, the first capturing beliefs and the latter information. The interpretation of $\omega \mathcal{I}_t \omega'$ is that at state ω and time t − according to the information received − it is possible that the true state is ω'. On the other hand, the interpretation of $\omega \mathcal{B}_t \omega'$ is that at state ω and time t, in light of the information

received, the agent considers state ω' possible (an alternative expression is "ω' is a doxastic alternative to ω at time t"). We shall use the following notation:

$$\mathcal{B}_t(\omega) = \{\omega' \in \Omega : \omega \mathcal{B}_t \omega'\} \text{ and, similarly, } \mathcal{I}_t(\omega) = \{\omega' \in \Omega : \omega \mathcal{I}_t \omega'\}.$$

Thus $\mathcal{B}_t(\omega)$ is the set of states that are reachable from ω according to the relation $\mathcal{B}_t$ and similarly for $\mathcal{I}_t(\omega)$.

Definition 1 *A temporal belief-information structure* $\langle T, \rightarrowtail, \Omega, \{\mathcal{B}_t, \mathcal{I}_t\}_{t \in T} \rangle$ *is a* temporal belief revision frame *if it satisfies the following properties:* $\forall \omega \in \Omega$, $\forall t, t', t'' \in T$*:*

1. $\mathcal{B}_t(\omega) \subseteq \mathcal{I}_t(\omega)$,
2. $\mathcal{B}_t(\omega) \neq \varnothing$,
3. *if* $t \rightarrowtail t'$ *then* $\mathcal{I}_{t'}(\omega) \subseteq \mathcal{I}_t(\omega)$,
4. *if* $t \rightarrowtail t'$, $t \rightarrowtail t''$ *and* $\mathcal{I}_{t'}(\omega) = \mathcal{I}_{t''}(\omega)$ *then* $\mathcal{B}_{t'}(\omega) = \mathcal{B}_{t''}(\omega)$,
5. *if* $t \rightarrowtail t'$ *and* $\mathcal{B}_t(\omega) \cap \mathcal{I}_{t'}(\omega) \neq \varnothing$ *then* $\mathcal{B}_{t'}(\omega) = \mathcal{B}_t(\omega) \cap \mathcal{I}_{t'}(\omega)$.

Property 1 says that information is believed and Property 2 that beliefs are consistent. Thus the two together imply that information itself is consistent.[1] By Property 3, later information never contradicts earlier information: at any given state, information becomes more refined as time goes by. Property 4 requires that at any two instants that share the same immediate predecessor, if information is the same then beliefs must be the same. Property 5 was called the 'Qualitative Bayes Rule' (QBR) in [4], based on the following observation. In a probabilistic setting, let $P_{\omega,t}$ be the probability measure over a set of states Ω representing the agent's probabilistic beliefs at state ω and time t, let $F \subseteq \Omega$ be an event representing the information received by the agent at a later date t' and let $P_{\omega,t'}$ be the posterior probability measure representing the revised beliefs at state ω and date t'. Bayes' rule requires that, if $P_{\omega,t}(F) > 0$, then, for every event $E \subseteq \Omega$, $P_{\omega,t'}(E) = \frac{P_{\omega,t}(E \cap F)}{P_{\omega,t}(F)}$. Bayes' rule thus implies the following (where $supp(P)$ denotes the support of the probability measure P):

$$\text{if } \; supp(P_{\omega,t}) \cap F \neq \varnothing, \text{then } \; supp(P_{\omega,t'}) = supp(P_{\omega,t}) \cap F.$$

[1] As pointed out by Friedman and Halpern [9], it is not clear how one could be informed of a contradiction.

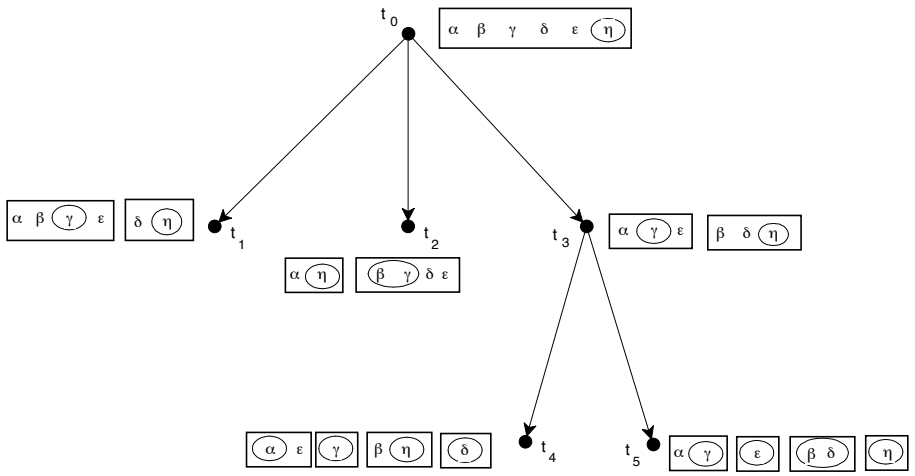


Figure 1: A temporal belief revision frame

If we set $\mathcal{B}_t(\omega) = supp(P_{\omega,t})$, $F = \mathcal{I}_{t'}(\omega)$, with $t \rightarrowtail t'$, and $\mathcal{B}_{t'}(\omega) = supp(P_{\omega,t'})$ then we get Property 5. Thus in a probabilistic setting the proposition "at date t the agent believes ϕ" would be interpreted as "the agent assigns probability 1 to the set of states where ϕ is true".

Figure 1 shows a temporal belief revision frame. For simplicity, in all the figures we assume that, for every instant t, the information relation $\mathcal{I}_t$ is an equivalence relation (whose equivalence classes are denoted by rectangles) and the belief relation $\mathcal{B}_t$ is transitive and euclidean[2] (we represent this fact by enclosing states in ovals and, within an equivalence class of $\mathcal{I}_t$, we have that $\omega' \in \mathcal{B}_t(\omega)$ if and only if ω' belongs to an oval). Note, however, that none of the results below require $\mathcal{I}_t$ to be an equivalence relation, nor do they require $\mathcal{B}_t$ to be transitive and euclidean.

For example, in Figure 1 at state α and time t_3 the agent is informed that the true state is either α, γ or ε ($\mathcal{I}_{t_3}(\alpha) = \{\alpha, \gamma, \varepsilon\}$) and (incorrectly) believes that it is γ ($\mathcal{B}_{t_3}(\alpha) = \{\gamma\}$). At the next instant t_4 (and still at state α) the agent is now informed that the true state is either α or ε ($\mathcal{I}_{t_4}(\alpha) = \{\alpha, \varepsilon\}$) and forms the revised (correct) belief that the true state is α. On the other hand, t_5 is an alternative next instant to t_3 and at t_5 (and still at state α) the agent's information is $\mathcal{I}_{t_5}(\alpha) = \{\alpha, \gamma\}$ and, according

[2] $\mathcal{B}_t$ is transitive if $\omega' \in \mathcal{B}_t(\omega)$ implies that $\mathcal{B}_t(\omega') \subseteq \mathcal{B}_t(\omega)$; it is euclidean if $\omega' \in \mathcal{B}_t(\omega)$ implies that $\mathcal{B}_t(\omega) \subseteq \mathcal{B}_t(\omega')$. Property 1 of Definition 1 is usually referred to as seriality.

to the Qualitative Bayes' Rule (Property 5 of Definition 1), she maintains the earlier (incorrect) belief that the true state is γ ($\mathcal{B}_{t_5}(\alpha) = \{\gamma\}$).[3]

We want to relate temporal belief revision frames to the AGM theory of belief revision ([1]), which is reviewed in the following section.[4]

3 Belief revision functions

Let Φ_0 be the set of formulas of a propositional language based on a countable set S_0 of atomic propositions.[5] Given a subset $K \subseteq \Phi_0$, its PL-deductive closure $[K]^{PL}$ (where 'PL' stands for Propositional Logic) is defined as follows: $\psi \in [K]^{PL}$ if and only if there exist $\phi_1, ..., \phi_n \in K$ (with $n \geq 0$) such that $(\phi_1 \wedge ... \wedge \phi_n) \rightarrow \psi$ is a tautology (that is, a theorem of Propositional Logic). A set $K \subseteq \Phi_0$ is *consistent* if $[K]^{PL} \neq \Phi_0$ (equivalently, if there is no formula ϕ such that both ϕ and $\neg\phi$ belong to $[K]^{PL}$). A set $K \subseteq \Phi_0$ is *deductively closed* if $K = [K]^{PL}$. A *belief set* is a set $K \subseteq \Phi_0$ which is deductively closed.

Let K be a consistent belief set representing the agent's initial beliefs and let $\Psi_0 \subseteq \Phi_0$ be a set of formulas representing possible items of information. A *belief revision function based on K* is a function $\circledast_K : \Psi_0 \rightarrow 2^{\Phi_0}$ (where 2^{Φ_0} denotes the set of subsets of Φ_0) that associates with every formula $\psi \in \Psi_0$ (thought of as new information) a set $\circledast_K(\psi) \subseteq \Phi_0$ (thought of as the revised beliefs).[6] If $\Psi_0 \neq \Phi_0$ then $\circledast_K$ is called a *partial* belief revision function, while if $\Psi_0 = \Phi_0$ then $\circledast_K$ is called a *full* belief revision function.

Definition 2 *Let $\circledast_K : \Psi_0 \rightarrow 2^{\Phi_0}$ be a (partial) belief revision function and $\circledast'_K : \Phi_0 \rightarrow 2^{\Phi_0}$ a full belief revision function. We say that $\circledast'_K$ is an* extension *of $\circledast_K$ if, for every $\psi \in \Psi_0$, $\circledast'_K(\psi) = \circledast_K(\psi)$.*

A *full* belief revision function is called an *AGM revision function* if it satisfies the

[3] As further illustration, focusing on state ε and the immediate successors of t_0, we have that $\mathcal{I}_{t_1}(\varepsilon) = \{\alpha, \beta, \gamma, \varepsilon\}$, $\mathcal{B}_{t_1}(\varepsilon) = \{\gamma\}$, $\mathcal{I}_{t_2}(\varepsilon) = \{\beta, \gamma, \delta, \varepsilon\}$, $\mathcal{B}_{t_2}(\varepsilon) = \{\beta, \gamma\}$, $\mathcal{I}_{t_3}(\varepsilon) = \{\alpha, \gamma, \varepsilon\}$ and $\mathcal{B}_{t_3}(\varepsilon) = \{\gamma\}$. This collection of sets will be used later to illustrate the notion of AGM-consistency.

[4] For a more detailed account see, for example, [10] or [8].

[5] Thus Φ_0 is defined recursively as follows: if $p \in S_0$ then $p \in \Phi_0$ and if $\phi, \psi \in \Phi_0$ then $\neg\phi \in \Phi_0$ and $(\phi \vee \psi) \in \Phi_0$.

[6] In the literature it is common to use the notation K^*_ψ instead of $\circledast_K(\psi)$.

following properties, known as the AGM axioms: $\forall \phi, \psi \in \Phi_0$,

(AGM1) $\circledast_K(\phi) = [\circledast_K(\phi)]^{PL}$

(AGM2) $\phi \in \circledast_K(\phi)$

(AGM3) $\circledast_K(\phi) \subseteq [K \cup \{\phi\}]^{PL}$

(AGM4) if $\neg\phi \notin K$, then $[K \cup \{\phi\}]^{PL} \subseteq \circledast_K(\phi)$

(AGM5) $\circledast_K(\phi) = \Phi_0$ if and only if ϕ is a contradiction

(AGM6) if $\phi \leftrightarrow \psi$ is a tautology then $\circledast_K(\phi) = \circledast_K(\psi)$

(AGM7) $\circledast_K(\phi \wedge \psi) \subseteq [\circledast_K(\phi) \cup \{\psi\}]^{PL}$

(AGM8) if $\neg\psi \notin \circledast_K(\phi)$, then $[\circledast_K(\phi) \cup \{\psi\}]^{PL} \subseteq \circledast_K(\phi \wedge \psi)$.

AGM1 requires the revised belief set to be deductively closed.

AGM2 requires that the information be believed.

AGM3 says that beliefs should be revised minimally, in the sense that no new formula should be added unless it can be deduced from the information received and the initial beliefs.[7]

AGM4 says that if the information received is compatible with the initial beliefs, then any formula that can be deduced from the information and the initial beliefs should be part of the revised beliefs.

AGM5 requires the revised beliefs to be consistent, unless the information ϕ is a contradiction (that is, $\neg\phi$ is a tautology).

AGM6 requires that if ϕ is propositionally equivalent to ψ then the result of revising by ϕ be identical to the result of revising by ψ.

AGM7 and AGM8 are a generalization of AGM3 and AGM4 that

> "applies to *iterated* changes of belief. The idea is that if $\circledast_K(\phi)$ is a revision of K [prompted by ϕ] and $\circledast_K(\phi)$ is to be changed by adding further sentences, such a change should be made by using expansions of $\circledast_K(\phi)$ whenever possible. More generally, the minimal change of K to include both ϕ and ψ (that is, $\circledast_K(\phi \wedge \psi)$) ought to be the same as the expansion of $\circledast_K(\phi)$ by ψ, so long as ψ does not contradict the beliefs in $\circledast_K(\phi)$" (Gärdenfors [10], p. 55; notation changed to match ours).

[7] Note that, for every formula ψ, $\psi \in [K \cup \{\phi\}]^{PL}$ if and only if $(\phi \rightarrow \psi) \in K$ (since, by hypothesis, $K = [K]^{PL}$).

4 Temporal models and AGM revision

We now return to the semantic structures of Definition 1 and interpret them by adding a valuation that associates with every atomic proposition p the set of states at which p is true. Note that, by defining a valuation this way, we frame the problem as one of belief revision, since the truth value of an atomic proposition depends only on the state and not on the time.[8]

Let S_0 be a countable set of atomic propositions and Φ_0 the set of propositional formulas built from S_0 (see Footnote 5). Given a temporal belief revision frame $\mathcal{F} = \langle T, \rightarrowtail, \Omega, \{\mathcal{B}_t, \mathcal{I}_t\}_{t\in T}\rangle$, a *model based on* (or an interpretation of) $\mathcal{F}$ is obtained by adding to $\mathcal{F}$ a *valuation* $V : S_0 \to 2^\Omega$ (where 2^Ω denotes the set of subsets of Ω).[9] Truth of an arbitrary formula $\phi \in \Phi_0$ at state ω in model $\mathcal{M}$ is denoted by $\omega \models_\mathcal{M} \phi$ and is defined recursively as follows:

(1) for $p \in S_0, \omega \models_\mathcal{M} p$ if and only if $\omega \in V(p)$,

(2) $\omega \models_\mathcal{M} \neg\phi$ if and only if $\omega \not\models_\mathcal{M} \phi$, and

(3) $\omega \models_\mathcal{M} (\phi \vee \psi)$ if and only if either $\omega \models_\mathcal{M} \phi$ or $\omega \models_\mathcal{M} \psi$ (or both).

The truth set of formula ϕ in model $\mathcal{M}$ is denoted by $|\phi|_\mathcal{M}$; thus $|\phi|_\mathcal{M} = \{\omega \in \Omega : \omega \models_\mathcal{M} \phi\}$.

Definition 3 *Given a model $\mathcal{M} = \langle T, \rightarrowtail, \Omega, \{\mathcal{B}_t, \mathcal{I}_t\}_{t\in T}, V\rangle$, a state $\omega \in \Omega$, an instant $t \in T$ and formulas $\phi, \psi \in \Phi_0$ we say that*

- *at (ω, t) the agent is informed that ψ if and only if $\mathcal{I}_t(\omega) = |\psi|_\mathcal{M}$,*
- *at (ω, t) the agent believes that ϕ if and only if $\mathcal{B}_t(\omega) \subseteq |\phi|_\mathcal{M}$.*

Note that for information we require *equality* of the two sets (this corresponds to the notion of 'all the agent knows': see [4] for references), while for belief we impose the standard requirement that $\mathcal{B}_t(\omega)$ be a *subset* of the truth set of a formula.

[8] In principle, the temporal structures of Definition 1 can be used to describe either a situation where the objective facts describing the world do not change − so that only the beliefs of the agent change over time − or a situation where both the facts and the doxastic state of the agent change. In the literature the first situation is called *belief revision*, while the latter is called *belief update* (see [12]). We restrict attention to belief revision.

[9] If instead of belief revision we were interested in belief update (see Footnote 8), then we would need to define a valuation as a function $V : S_0 \to 2^{\Omega\times T}$.

Given a model $\mathcal{M}$ and a state-instant pair (ω, t), according to Definition 3 we can associate with (ω, t) a belief set and a (typically partial) belief revision function as follows. Let

$$K_{\mathcal{M},\omega,t} = \{\phi \in \Phi_0 : \mathcal{B}_t(\omega) \subseteq |\phi|_{\mathcal{M}}\}, \tag{1}$$

denote the set of formulas that the agent believes at (ω, t), that is, his belief set at (ω, t). It is straightforward to show that $K_{\mathcal{M},\omega,t}$ is a consistent and deductively closed set. Let

$$\Psi_{\mathcal{M},\omega,t} = \left\{\psi \in \Phi_0 : |\psi|_{\mathcal{M}} = \mathcal{I}_{t'}(\omega) \text{ for some } t' \in t^{\rightarrowtail}\right\} \tag{2}$$

be the possible items of information that the agent might receive next time (that is, at some immediate successor of t: recall that $t^{\rightarrowtail} = \{t \in T : t \rightarrowtail t'\}$). Finally let $\circledast_{K_{\mathcal{M},\omega,t}} : \Psi_{\mathcal{M},\omega,t} \rightarrow 2^{\Phi_0}$ be defined as[10]

$$\circledast_{K_{\mathcal{M},\omega,t}}(\psi) = \left\{\phi \in \Phi_0 : \mathcal{B}_{t'}(\omega) \subseteq |\phi|_{\mathcal{M}} \text{ for } t' \in t^{\rightarrowtail} \text{ such that } \mathcal{I}_{t'}(\omega) = |\psi|_{\mathcal{M}}\right\}. \tag{3}$$

That is, if at the immediate successor t' of t, the agent is informed that ψ ($\mathcal{I}_{t'}(\omega) = |\psi|_{\mathcal{M}}$), then his revised belief set is given by the set of formulas that he believes at (ω, t'): $\{\phi \in \Phi_0 : \mathcal{B}_{t'}(\omega) \subseteq |\phi|_{\mathcal{M}}\}$.

What properties must a choice frame $\mathcal{F}$ satisfy in order for it to be the case that the belief revision functions associated with an *arbitrary* interpretation (or model) of $\mathcal{F}$ can be extended to full AGM belief revision functions?

Definition 4 *A temporal belief revision frame $\mathcal{F} = \langle T, \rightarrowtail, \Omega, \{\mathcal{B}_t, \mathcal{I}_t\}_{t\in T}\rangle$ is* AGM-consistent *if, for every model $\mathcal{M} = \langle \mathcal{F}, V\rangle$ based on it and for every state-instant pair (ω, t), the associated belief revision function $\circledast_{K_{\mathcal{M},(\omega,t)}}$ (see (3) above) can be extended (see Definition 2) to a full belief revision function that satisfies the AGM axioms.*

The temporal belief revision frame illustrated in Figure 1 is *not* AGM consistent. To see this, consider the following model based on it. The set of atomic propositions

[10]This function is well defined because of Property 4 of Definition 1.

is $S_0 = \{p, q, r, s\}$ and the valuation $V : S_0 \to 2^{\Omega}$ is as follows: $V(p) = \{\alpha, \beta, \gamma, \varepsilon\}$, $V(q) = \{\beta, \gamma, \delta, \varepsilon\}$, $V(r) = \{\beta, \gamma\}$ and $V(s) = \{\gamma\}$. The initial beliefs are given by the consistent and deductively closed set $K = \{\phi \in \Phi_0 : \eta \in |\phi|\}$. If ψ is a formula such that $|\psi| = \mathcal{I}_t(\omega)$ for some $t \in \{t_1, t_2, t_3\}$ and $\omega \in \Omega$, let $\circledast_K(\psi) = \{\phi \in \Phi_0 : \mathcal{B}_t(\omega) \subseteq |\phi|\}$ be the revised beliefs at state ω and instant t after receiving information ϕ. Thus, for example, $\{\neg p, \neg q, \neg r, \neg s\} \subseteq K$; furthermore, since $\mathcal{I}_{t_1}(\varepsilon) = \{\alpha, \beta, \gamma, \varepsilon\} = |p|$ and $\mathcal{B}_{t_1}(\varepsilon) = \{\gamma\} \subseteq |p| \cap |q| \cap |r| \cap |s|$, $\{p, q, r, s\} \subseteq \circledast_K(p)$. Similarly, since $\mathcal{I}_{t_2}(\varepsilon) = \{\beta, \gamma, \delta, \varepsilon\} = |q|$ and $\mathcal{B}_{t_2}(\varepsilon) = \{\beta, \gamma\}$, $\{p, q, r\} \subseteq \circledast_K(q)$ but $s \notin \circledast_K(q)$. Since $s \in \circledast_K(p)$ and $s \notin \circledast_K(q)$

$$\circledast_K(p) \neq \circledast_K(q). \tag{4}$$

Suppose that $\circledast'_K : \Phi_0 \to 2^{\Phi_0}$ is an AGM belief revision function that extends $\circledast_K$. Since $(q \wedge r) \in \circledast_K(p) = \circledast'_K(p)$ and $\circledast'_K(p)$ is consistent, $\neg(q \wedge r) \notin \circledast'_K(p)$. It follows from AGM axioms AGM7 and AGM8 that $\circledast'_K(p \wedge (q \wedge r)) = [\circledast'_K(p) \cup \{(q \wedge r)\}]^{PL} = [\circledast'_K(p)]^{PL} = \circledast'_K(p)$. Similarly, since $(p \wedge r) \in \circledast_K(q) = \circledast'_K(q)$, by AGM axioms AGM7 and AGM8 $\circledast'_K(q \wedge (p \wedge r)) = \circledast'_K(q)$. Furthermore, since $(p \wedge (q \wedge r)) \leftrightarrow (q \wedge (p \wedge r))$ is a tautology, it follows from AGM axiom AGM6 that $\circledast'_K(p \wedge (q \wedge r)) = \circledast'_K(q \wedge (p \wedge r))$. Hence $\circledast'_K(p) = \circledast'_K(q)$. Since $\circledast'_K$ is an extension of $\circledast_K$, $\circledast'_K(p) = \circledast_K(p)$ and $\circledast'_K(q) = \circledast_K(q)$, yielding a contradiction with (4).

The main result of this section is that, when Ω is finite, AGM-consistency is equivalent to the property that, at every state-instant pair (ω, t), belief revision can be rationalized by a "plausibility" ordering of the set of states, in the sense that at t and at the immediate successors of t (and a state ω) the states that the agent considers possible are the most plausible among the ones that are compatible with the information received.

Definition 5 *A* plausibility ordering *of the set of states Ω is a total pre-order of Ω, that is, a binary relation $R \subseteq \Omega \times \Omega$ which is complete ($\forall \omega, \omega' \in \Omega$, either $\omega R \omega'$ or $\omega' R \omega$) and transitive ($\forall \omega, \omega', \omega'' \in \Omega$, if $\omega R \omega'$ and $\omega' R \omega''$ then $\omega R \omega''$). We interpret $\omega R \omega'$ as "ω is at least as plausible as ω'". Given a plausibility ordering R of Ω and a subset*

$E \subseteq \Omega$, *let*

$$\max_R E = \{\omega \in E : \omega R \omega',\ \forall \omega' \in E\}$$

Thus $\max_R\ E$ is the set of states in E that are maximal ("most plausible") according to the ordering R.

Definition 6 *A temporal belief revision frame* $\langle T, \rightarrowtail, \Omega, \{\mathcal{B}_t, \mathcal{I}_t\}_{t \in T}\rangle$ *is a* plausibility frame *if it satisfies the following property:* $\forall \omega \in \Omega$, $\forall t \in T$, *there exists a plausibility ordering* $R_{(\omega,t)}$ *of* Ω *that rationalizes the agent's beliefs at* t *and its immediate successors (and state* ω*) in the sense that*

1. $\mathcal{B}_t(\omega) = \max_{R_{(\omega,t)}}\ \mathcal{I}_t(\omega)$, *and*
2. *for every* $t' \in T$ *such that* $t \rightarrowtail t'$, $\mathcal{B}_{t'}(\omega) = \max_{R_{(\omega,t)}}\ \mathcal{I}_{t'}(\omega)$.

The temporal belief revision frame illustrated in Figure 1 is *not* a plausibility frame. In fact, consider the state-instant pair (ε, t_0). Suppose that R is a total pre-order of Ω that rationalizes the agent's beliefs at (ε, t_0). Then, it must be that $\mathcal{B}_{t_1}(\varepsilon) = \{\gamma\} = \max_R\ \mathcal{I}_{t_1}(\varepsilon) = \{\alpha, \beta, \gamma, \varepsilon\}$, which implies that $(\beta, \gamma) \notin R$; furthermore, it must be that $\mathcal{B}_{t_2}(\varepsilon) = \{\beta, \gamma\} = \max_R\ \mathcal{I}_{t_2}(\varepsilon) = \{\beta, \gamma, \delta, \varepsilon\}$,which implies that $(\beta, \gamma) \in R$, yielding a contradiction.

Remark 1 *In Definition 6 the total pre-order* R *that rationalizes beliefs at* (ω, t) *is required to be a "global" relation on the entire set of states* Ω*. Alternatively one could merely require the existence of a "local" total pre-order of* $\mathcal{I}_t(\omega)$*. However, the two definitions are equivalent. If* R *is a total pre-order of* Ω*, then the restriction* R' *of* R *to* $\mathcal{I}_t(\omega)$ *[that is,* $R' = R \cap (\mathcal{I}_t(\omega) \times \mathcal{I}_t(\omega))$*] would provide the desired local total pre-order. Conversely, if* R' *is a local total pre-order of* $\mathcal{I}_t(\omega)$ *then define* $R \subseteq \Omega \times \Omega$ *as follows:* $R = R' \cup \{(x, y) : x \in \mathcal{I}_t(\omega)$ *and* $y \in \Omega \backslash \mathcal{I}_t(\omega)\} \cup \{(x, y) : x, y \in \Omega \backslash \mathcal{I}_t(\omega)\}$*. Then* R *is the desired global total preorder.*[11]

The next proposition provides a necessary and sufficient condition for a temporal belief revision frame to be a plausibility frame. All the proofs are given in the Appendix.

[11] Recall that, by Property 3 of Definition 1, if $t \rightarrowtail t'$ then $\mathcal{I}_{t'}(\omega) \subseteq \mathcal{I}_t(\omega)$.

Proposition 1 *A temporal belief revision frame* $\langle T, \rightarrowtail, \Omega, \{\mathcal{B}_t, \mathcal{I}_t\}_{t\in T}\rangle$ *is a plausibility frame if and only if it satisfies the following property:* $\forall \omega \in \Omega$, $\forall t \in T$, $\forall t_0, t_1, ..., t_n \in t^{\rightarrowtail}$ *with* $t_n = t_0$ *and* $n \geq 1$ *(recall that* $t^{\rightarrowtail}$ *is the set of immediate successors of* t*),*

$$\begin{array}{l} \textit{if } \mathcal{I}_{t_{k-1}}(\omega) \cap \mathcal{B}_{t_k}(\omega) \neq \varnothing, \forall k = 1, ..., n, \\ \textit{then } \mathcal{I}_{t_{k-1}}(\omega) \cap \mathcal{B}_{t_k}(\omega) = \mathcal{B}_{t_{k-1}}(\omega) \cap \mathcal{I}_{t_k}(\omega), \forall k = 1, ..., n. \end{array} \quad (PLS)$$

For example, in the frame of Figure 1, property PLS is violated since $\mathcal{I}_{t_1}(\varepsilon) \cap \mathcal{B}_{t_2}(\varepsilon) = \{\beta, \gamma\} \neq \mathcal{I}_{t_2}(\varepsilon) \cap \mathcal{B}_{t_1}(\varepsilon) = \{\gamma\}$.

The following proposition says that, when the set of states is finite, a temporal belief revision frame is AGM-consistency if and only if it is a plausibility frame.

Proposition 2 *Let* $\mathcal{F} = \langle T, \rightarrowtail, \Omega, \{\mathcal{B}_t, \mathcal{I}_t\}_{t\in T}\rangle$ *be a temporal belief frame where* Ω *is finite. Then the following are equivalent.*

1. $\mathcal{F}$ *is AGM-consistent.*
2. $\mathcal{F}$ *is a plausibility frame.*

We now turn to a modal logic characterization of AGM-consistent frames.

5 A temporal logic for belief revision

We briefly review the modal language introduced in [4], which contains the following modal operators: the next-time operator $\bigcirc$, the belief operator B, the information operator I and the "all state" operator A. The intended interpretation is as follows:

$\bigcirc\phi$: "at every next instant it will be the case that ϕ"

$B\phi$: "the agent believes that ϕ"

$I\phi$: "the agent is informed that ϕ"

$A\phi$: "it is true at every state that ϕ".

Fix a model $\mathcal{M} = \langle T, \rightarrowtail, \Omega, \{\mathcal{B}_t, \mathcal{I}_t\}_{t\in T}, V\rangle$, where $V : S_0 \rightarrow 2^{\Omega}$ is a valuation. Given a state ω, an instant t and a formula ϕ, we write $(\omega, t) \models_{\mathcal{M}} \phi$ to denote that ϕ is true at (ω, t) in model $\mathcal{M}$. Let $\|\phi\|_{\mathcal{M}} \subseteq \Omega \times T$ denote the truth set of ϕ, that is, $\|\phi\|_{\mathcal{M}} = \{(\omega, t) \in \Omega \times T : (\omega, t) \models_{\mathcal{M}} \phi\}$ and let $\lceil\phi\rceil_{\mathcal{M},t} \subseteq \Omega$ denote the set of states at which ϕ is true *at time* t, that is, $\lceil\phi\rceil_{\mathcal{M},t} = \{\omega \in \Omega : (\omega, t) \models_{\mathcal{M}} \phi\}$. Truth at (ω, t)

is defined as usual for Boolean (that is, non-modal) formulas. For the modal formulas we have

$(\omega, t) \models_{\mathcal{M}} \bigcirc\phi$ if and only if $(\omega, t') \models_{\mathcal{M}} \phi$ for every t' such that $t \rightarrowtail t'$.

$(\omega, t) \models_{\mathcal{M}} B\phi$ if and only if $\mathcal{B}_t(\omega) \subseteq \lceil\phi\rceil_{\mathcal{M},t}$

$(\omega, t) \models_{\mathcal{M}} I\phi$ if and only if $\mathcal{I}_t(\omega) = \lceil\phi\rceil_{\mathcal{M},t}$

$(\omega, t) \models_{\mathcal{M}} A\phi$ if and only if $\lceil\phi\rceil_{\mathcal{M},t} = \Omega$.

Note that, while the truth condition for the operator B is the standard one, the truth condition for the operator I is non-standard: instead of simply requiring that $\mathcal{I}_t(\omega) \subseteq \lceil\phi\rceil_{\mathcal{M},t}$ we require equality: $\mathcal{I}_t(\omega) = \lceil\phi\rceil_{\mathcal{M},t}$ (for an explanation see [4], where the role of the "all state" operator is also discussed).

A formula ϕ is *valid in a model* if $\|\phi\|_{\mathcal{M}} = \Omega \times T$, that is, if ϕ is true at every state-instant pair (ω, t). A formula ϕ is *valid in a frame* if it is valid in every model based on it. A property of frames *characterizes* (or is *characterized by*) an axiom if the axiom is valid in every frame that satisfies the property and, conversely, if the frame violates the property then there is a model based on that frame and a state-instant pair at which the axiom is falsified.

Let $\Diamond$ be an abbreviation for $\neg\bigcirc\neg$ (thus $(\omega, t) \models_{\mathcal{M}} \Diamond\phi$ if and only if $(\omega, t') \models_{\mathcal{M}} \phi$ for some t' such that $t \rightarrowtail t'$); furthermore, let $\bigwedge_{j=1,...,m} \phi_j$ denote the conjunction $(\phi_1 \wedge ... \wedge \phi_m)$. In the following proposition all the formulas are restricted to be Boolean, that is, formulas that do not contain any modal operators.

Proposition 3 *The class of plausibility frames (see Definition 6) is characterized by the following axioms (in Axiom 6 let $\phi_0 = \phi_n$ and $\chi_0 = \chi_n$):*

1. $I\phi \rightarrow B\phi$

2. $B\phi \rightarrow \neg B\neg\phi$

3. $I\psi \rightarrow \bigcirc(I\phi \rightarrow A(\phi \rightarrow \psi))$

4. $\Diamond(I\psi \wedge B\phi) \rightarrow \bigcirc(I\psi \rightarrow B\phi)$

5a. $(\neg B\neg\phi \wedge B\psi) \rightarrow \bigcirc(I\phi \rightarrow B\psi)$

5b. $\neg B\neg(\phi \wedge \neg\psi) \rightarrow \bigcirc(I\phi \rightarrow \neg B\psi)$

6. $\bigwedge_{j=1,...,n} \Diamond\left(I\phi_j \wedge \neg B\neg\phi_{j-1} \wedge B\chi_j\right) \rightarrow$
$\bigwedge_{j=1,...,n} \bigcirc\Big((I\phi_j \rightarrow B(\phi_{j-1} \rightarrow \chi_{j-1})) \wedge (I\phi_{j-1} \rightarrow B(\phi_j \rightarrow \chi_j))\Big)$

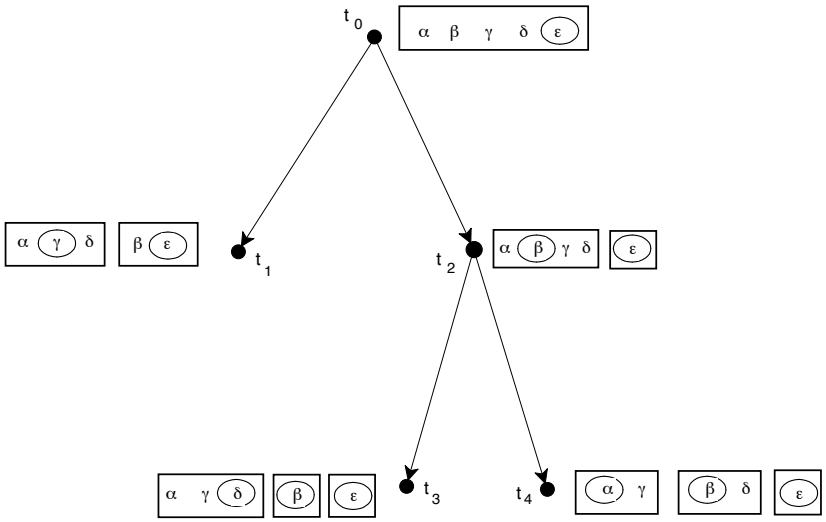


Figure 2: A plausibility frame

Axiom 1 says that information is believed and Axiom 2 that beliefs are consistent. Axioms 3 and 4 correspond to Properties 3 and 4 of Definition 1. It is shown in [5] that the conjunction of $5a$, $5b$ and a weaker version of 2 provide a characterization of Property 5 of Definition 1 (the Qualitative Bayes Rule). Axiom 6 characterizes Property PLS (see Proposition 1).

6 Concluding remarks

In general, in a plausibility frame the ordering $R_{(\omega,t)}$ that rationalizes belief revision at state-instant pair (ω, t) may be different from the ordering $R_{(\omega,t')}$ that rationalizes belief revision at (ω, t') with $t \rightarrowtail t'$. To see this, consider the plausibility frame illustrated in Figure 2.

Consider state α. At instant t_0 and its immediate successors the agent's beliefs are rationalized by the ordering $R_{(\alpha.t_0)} = (\{\varepsilon\}\times\Omega)\cup(\{\beta\}\times\Omega\backslash\{\varepsilon\})\cup(\{\gamma\}\times\Omega\backslash\{\beta,\varepsilon\})\cup(\{\alpha,\delta\}\times\{\alpha,\delta\})$;[12] in particular, for any ordering R that rationalizes beliefs at (α, t_0) it must be that $(\gamma,\alpha)\in R$ and $(\alpha,\gamma)\notin R$. On the other hand, at instant t_2 the agent's beliefs are rationalized by the ordering $R_{(\alpha.t_2)} = (\{\beta,\varepsilon\}\times\Omega)\cup(\{\delta\}\times\Omega\backslash\{\beta,\varepsilon\})\cup$

[12]That is, ε is more plausible than every other state, β is the second most plausible state, followed by γ; α and β are the least plausible states.

$(\{\alpha\} \times \{\alpha, \gamma\}) \cup \{(\gamma, \gamma)\}$; in particular, for any ordering R that rationalizes beliefs at (α, t_2) it must be that $(\alpha, \gamma) \in R$ and $(\gamma, \alpha) \notin R$. Thus the ranking of α versus γ must be the opposite at t_0 relative to t_2.

Indeed, the literature on iterated belief revision has pointed out that, in general, the entrenchment ordering of formulas that rationalizes AGM belief revision at an instant can be quite different from the entrenchment ordering that rationalizes belief revision at a later instant. Several proposals have been made concerning the restrictions that should be placed on iterated belief revision (see, for example, [7] and [13]). The analysis proposed in this paper allows one to frame the discussion both semantically, in terms of plausibility frames and associated ordering of states, and syntactically, in terms of modal axioms.[13]

A Appendix

In order to prove Proposition 1 we need some preliminary definitions and lemmas.

Definition 7 *A* choice structure *is a triple $\langle \Omega, \mathcal{E}, f \rangle$ where Ω is a set, $\mathcal{E} \subseteq 2^{\Omega}$ is a collection of non-empty subsets of Ω and $f : \mathcal{E} \to 2^{\Omega}$ is a function that satisfies the following property: $\forall E \in \mathcal{E}, \varnothing \neq f(E) \subseteq E$.*

Give a choice structure $\mathcal{C} = \langle \Omega, \mathcal{E}, f \rangle$, a Hansson sequence *in $\mathcal{C}$ is a sequence $\langle E_1, ..., E_n, E_{n+1} \rangle$ ($n \geq 1$) such that, $\forall k = 1, ..., n$: (1) $E_k \in \mathcal{E}$, (2) $E_{n+1} = E_1$ and (3) $E_k \cap f(E_{k+1}) \neq \varnothing$.*

Give two choice structures $\mathcal{C} = \langle \Omega, \mathcal{E}, f \rangle$ and $\mathcal{C}' = \langle \Omega, \mathcal{E}', f' \rangle$, we say that $\mathcal{C}'$ is a QBR-extension of *$\mathcal{C}$ if (1) $\mathcal{E}' = \mathcal{E} \cup \{\Omega\}$, (2) f' is an extension of f, that is, $\forall E \in \mathcal{E}$, $f'(E) = f(E)$ and (3) $\forall E \in \mathcal{E}$, if $E \cap f'(\Omega) \neq \varnothing$ then $f(E) = E \cap f'(\Omega)$.*

Definition 8 *Given a choice structure $\mathcal{C} = \langle \Omega, \mathcal{E}, f \rangle$ and a total pre-order $R \subseteq \Omega \times \Omega$, we say that R* rationalizes *$\mathcal{C}$ if and only if, for every $E \in \mathcal{E}$, $f(E) = \max_R E$.*

The following result is due to Hansson ([11], Theorem 7, p. 455).

Proposition 4 *Let $\mathcal{C} = \langle \Omega, \mathcal{E}, f \rangle$ be a choice structure. The following are equivalent:*

1. there exists a total pre-order $R \subseteq \Omega \times \Omega$ that rationalizes $\mathcal{C}$,

2. for every Hansson sequence $\langle E_1, ..., E_n, E_{n+1} \rangle$, $E_k \cap f(E_{k+1}) = f(E_k) \cap E_{k+1}$, $\forall k = 1, ..., n$.

Lemma 5 *Let $\mathcal{C} = \langle \Omega, \mathcal{E}, f \rangle$ be a choice structure and $\mathcal{C}' = \langle \Omega, \mathcal{E}', f' \rangle$, a QBR-extension of $\mathcal{C}$ (see Definition 7). Then the following are equivalent:*

(a) if $\langle E_1, ..., E_n, E_{n+1} \rangle$ is a Hansson sequence in $\mathcal{C}$ then, $\forall k = 1, ..., n$, $E_k \cap f(E_{k+1}) = f(E_k) \cap E_{k+1}$;

(b) if $\langle E'_1, ..., E'_n, E'_{n+1} \rangle$ is a Hansson sequence in $\mathcal{C}'$ then, $\forall k = 1, ..., n$, $E'_k \cap f'(E'_{k+1}) = f'(E'_k) \cap E'_{k+1}$.

[13] For a different, but related, approach see [2] and [3]

Proof. That $(b) \Rightarrow (a)$ is obvious, since the set of Hansson sequences in $\mathcal{C}'$ contains the set of Hansson sequences in $\mathcal{F}$ (they are those where $E'_k \in \mathcal{E}$ for all k) Thus we only need to prove $(a) \Rightarrow (b)$. Consider first the case where, $\forall E \in \mathcal{E}, E \cap f'(\Omega) \neq \varnothing$. Then, since $\mathcal{C}'$ is a QBR-extension of $\mathcal{C}$, $f(E) = E \cap f'(\Omega), \forall E \in \mathcal{E}$. Define the following relation R' on Ω: for all $x, y \in \Omega$, $xR'y$ if and only if either (1) $x \in f'(\Omega)$ or (2) $x \notin f'(\Omega)$ and $y \notin f'(\Omega)$. R' is clearly a total pre-order[14] and, furthermore, for every $E \in \mathcal{E}'$, $f'(E) = \max_{R'} E$.[15] Thus, by Proposition 4, (b) holds. Suppose now that $E \cap f'(\Omega) = \varnothing$ for some $E \in \mathcal{E}$. Let $\mathcal{E}_0 = \{E \in \mathcal{E} : E \cap f'(\Omega) = \varnothing\}$ and let $\Omega_0 = \bigcup_{E \in \mathcal{E}_0} E$. Then $\Omega_0 \cap f'(\Omega) = \varnothing$. By Proposition 4 it follows from (a) that there is a total pre-order R of Ω such that, for all $E \in \mathcal{E}$, $f(E) = \max_R E$. Fix such a total pre-order R and define the following relation R' on Ω:

$$\begin{array}{rl} R' = & (R \cap (\Omega_0 \times \Omega_0)) \bigcup \{(x, y) \in \Omega \times \Omega : x \in f'(\Omega)\} \\ & \bigcup \{(x, y) \in \Omega \times \Omega : y \in \Omega \backslash (\Omega_0 \cup f'(\Omega))\} \end{array} \tag{5}$$

We want to show that R' is a total pre-order of Ω and is such that, for every $E \in \mathcal{E}'$, $f'(E) = \max_{R'} E$. If we establish this then, by Proposition 4, (b) holds.

Proof that R' is complete. Fix arbitrary $x, y \in \Omega$. We need to show that either $xR'y$ or $yR'x$. If $x \in f'(\Omega)$ then, by (5), $xR'y$; similarly, if $y \in f'(\Omega)$ then $yR'x$. If $x, y \in \Omega_0$ then it follows from (5) and completeness of R. If $y \in \Omega \backslash (\Omega_0 \cup f'(\Omega)$ then, by (5), $xR'y$; similarly, if $x \in \Omega \backslash (\Omega_0 \cup f'(\Omega)$ then $yR'x$.

Proof that R' is transitive. Fix arbitrary $x, y, z \in \Omega$ and suppose that $xR'y$ and $yR'z$. We need to show that $xR'z$. If $x \in f'(\Omega)$, then, by (5), $xR'z$. Assume that $x \notin f'(\Omega)$. Two cases are possible: (1) $x \in \Omega_0$ and (2) $x \in \Omega \backslash (\Omega_0 \cup f'(\Omega)$. In Case 1, since $xR'y$, it must be that either (1a) $y \in \Omega_0$ or (1b) $y \in \Omega \backslash (\Omega_0 \cup f'(\Omega)$. In Case 1a, since $yR'z$, it must be that either $z \in \Omega_0$, in which case $xR'z$ by (5) and transitivity of R, or $z \in \Omega \backslash (\Omega_0 \cup f'(\Omega)$, in which case $xR'z$ by (5). In Case 1b, since $yR'z$ by (5) it must be that $z \in \Omega \backslash (\Omega_0 \cup f'(\Omega)$ and thus, by (5), $xR'z$. Consider now Case 2, where $x \in \Omega \backslash (\Omega_0 \cup f'(\Omega)$. Then, since $xR'y$, it must be that $y \in \Omega \backslash (\Omega_0 \cup f'(\Omega)$ and thus, since $yR'z$, it must be that also $z \in \Omega \backslash (\Omega_0 \cup f'(\Omega)$. Hence $xR'z$ by (5).

Thus R' is a total pre-order of Ω. It remains to show that, for every $E \in \mathcal{E}'$, $f'(E) = \max_{R'} E$. It is clear from (5) that $f'(\Omega) = \max_{R'} \Omega$ (recall that $\Omega_0 \cap f'(\Omega) = \varnothing$). Thus we only need to show that $f(E) = \max_{R'} E$ for all $E \in \mathcal{E}$. If $E \in \mathcal{E}_0$ (that is, $E \cap f'(\Omega) = \varnothing$) then, since $f(E) = \max_R E$, it follows from (5) that $f(E) = \max_{R'} E$. Suppose, therefore, that $E \notin \mathcal{E}_0$, that is, $E \cap f'(\Omega) \neq \varnothing$. Then, since $\mathcal{C}'$ is a QBR-extension of $\mathcal{C}$, $f(E) = E \cap f'(\Omega)$. Hence, since $f'(\Omega) = \max_{R'} \Omega$ and $\max_{R'} \Omega \cap E = \max_{R'} E$ (because $\max_{R'} \Omega \cap E \neq \varnothing$), it follows that $f(E) = \max_{R'} E$.

■

[14] Proof of completeness. Fix arbitrary $x, y \in \Omega$. We need to show that either $xR'y$ or $yR'x$. If $x \in f'(\Omega)$ then $xR'y$; if $y \in f'(\Omega)$ then $yR'x$; if both $x \notin f'(\Omega)$ and $y \notin f'(\Omega)$ then $xR'y$ and $yR'x$.

Proof of transitivity. Fix arbitrary $x, y, z \in \Omega$ and suppose that $xR'y$ and $yR'z$. We need to show that $xR'z$. If $x \in f'(\Omega)$, then $xR'z$. If $x \notin f'(\Omega)$ then, since $xR'y$, it must be that $y \notin f'(\Omega)$ and thus, since $yR'z$, it must be that also $z \notin f'(\Omega)$. Thus $xR'z$.

[15] By definition of R', $\max_{R'} \Omega = f'(\Omega)$. Let $E \in \mathcal{E}$. Then, since $f(E) = E \cap f'(\Omega) = E \cap \max_{R'} \Omega$, $f(E) = \max_{R'} E$.

Proof of Proposition 1. Fix a temporal belief revision frame $\langle T, \rightarrowtail, \Omega, \{\mathcal{B}_t, \mathcal{I}_t\}_{t\in T}\rangle$, an arbitrary state $\hat{\omega}$ and an arbitrary instant $\hat{t}$. Condition PLS states that

$$\begin{array}{l}\forall t_0, t_1, ..., t_n \in \hat{t}^{\rightarrowtail} \text{ with } t_n = t_0 \text{ and } n \geq 1, \text{ if } \mathcal{I}_{t_{k-1}}(\hat{\omega}) \cap \mathcal{B}_{t_k}(\hat{\omega}) \neq \varnothing, \forall k = 1, ..., n, \\ \text{then } \mathcal{I}_{t_{k-1}}(\hat{\omega}) \cap \mathcal{B}_{t_k}(\hat{\omega}) = \mathcal{B}_{t_{k-1}}(\hat{\omega}) \cap \mathcal{I}_{t_k}(\hat{\omega}), \forall k = 1, ..., n\end{array} \tag{6}$$

Associate with $(\hat{\omega}, \hat{t})$ the following choice structure $\mathcal{C} = \left\langle \hat{\Omega}, \mathcal{E}, f \right\rangle$: $\hat{\Omega} = \mathcal{I}_{\hat{t}}(\hat{\omega})$, $\mathcal{E} = \{\mathcal{I}_t(\hat{\omega}) : t \in \hat{t}^{\rightarrowtail}\}$[16] and, for every $E \in \mathcal{E}$, if $E = \mathcal{I}_t(\hat{\omega})$ for some $t \in \hat{t}^{\rightarrowtail}$ then $f(E) = \mathcal{B}_t(\hat{\omega})$.[17]. Then (6) is equivalent to the following (see Definition 7)

$$\begin{array}{l}\text{for every Hansson sequence } \langle E_1, ..., E_n, E_{n+1}\rangle \text{ in } \mathcal{C} \\ E_j \cap f(E_{j+1}) = f(E_j) \cap E_{j+1}, \forall j = 1, ..., n.\end{array} \tag{7}$$

Let $\mathcal{C}' = \left\langle \hat{\Omega}, \mathcal{E}', f' \right\rangle$ be the extension of $\mathcal{C}$ given by $\mathcal{E}' = \mathcal{E} \cup \hat{\Omega}$ and $f'(\hat{\Omega}) = \mathcal{B}_t(\hat{\omega})$. Then, by Property 5 of Definition 1, $\mathcal{C}'$ is a QBR extension of $\mathcal{C}$ (see Definition 7). Thus, by Lemma 5, (7) is equivalent to

$$\begin{array}{l}\text{for every Hansson sequence } \langle E_1, ..., E_n, E_{n+1}\rangle \text{ in } \mathcal{C}' \\ E_j \cap f(E_{j+1}) = f(E_j) \cap E_{j+1}, \forall j = 1, ..., n\end{array} \tag{8}$$

By Proposition 4, (8) is equivalent to the existence of a total pre-order $\hat{R} \subseteq \hat{\Omega} \times \hat{\Omega}$ that rationalizes $\mathcal{C}'$ and thus - by construction of $\mathcal{C}'$ - $\hat{R}$ that rationalizes beliefs at $(\hat{\omega}, \hat{t})$ restricting the set of states to $\mathcal{I}_{\hat{t}}(\hat{\omega})$. By Remark 1 this is equivalent to the existence of a total pre-order $R \subseteq \Omega \times \Omega$ that rationalizes beliefs at $(\hat{\omega}, \hat{t})$. ■

Proof of Proposition 2. The proof is a "pointwise" application of the following result proved in [6]: if $\mathcal{C} = \langle \Omega, \mathcal{E}, f\rangle$ is a choice structure with Ω finite then $\mathcal{C}$ is rationalizable by a total pre-order of Ω (that is, for every $E \in \mathcal{E}$, $f(E) = \max_R E$) if and only if $\mathcal{C}$ is AGM-consistent.[18] One only needs to apply this result to the choice structure associated with each state-instant pair (ω, t) (as explained in the Proof of Proposition 1).

■

Proof of Proposition 3. It is shown in [5] that, for $j = 1, 2, 3$, Axiom j of Proposition 3 characterizes Property j of Definition 1.

Next we show that Axiom 4 of Proposition 3 characterizes Property 4 of Definition 1. Fix an arbitrary frame that satisfies Property 4 of Definition 1. Fix arbitrary $\hat{\omega} \in \Omega$, $\hat{t} \in T$ and Boolean formulas ϕ and ψ and suppose that $(\hat{\omega}, \hat{t}) \models \Diamond(I\psi \wedge B\phi)$. Then there exists a t' such that $\hat{t} \rightarrowtail t'$ and $(\hat{\omega}, t') \models I\psi \wedge B\phi$, that is, $\mathcal{I}_{t'}(\hat{\omega}) = \lceil\psi\rceil_{t'}$ and $\mathcal{B}_{t'}(\hat{\omega}) \subseteq \lceil\phi\rceil_{t'}$. We have to show that $(\hat{\omega}, \hat{t}) \models \bigcirc(I\psi \rightarrow B\phi)$. Fix an arbitrary $t \in T$ such that $\hat{t} \rightarrowtail t$ and suppose that $(\hat{\omega}, t) \models I\psi$. Then $\mathcal{I}_t(\hat{\omega}) = \lceil\psi\rceil_t$. Since ψ is a Boolean formula, by Proposition 5 in [4], $\lceil\psi\rceil_{t'} = \lceil\psi\rceil_t$. Hence $\mathcal{I}_{t'}(\hat{\omega}) = \mathcal{I}_t(\hat{\omega})$ and thus, by Property 4 of Definition 1, $\mathcal{B}_{t'}(\hat{\omega}) = \mathcal{B}_t(\hat{\omega})$. Hence $\mathcal{B}_t(\hat{\omega}) \subseteq \lceil\phi\rceil_{t'}$. Since

[16] Recall that, by Property 3 of Definition 1, if $\hat{t} \rightarrowtail t$ then $\mathcal{I}_t(\hat{\omega}) \subseteq \mathcal{I}_{\hat{t}}(\hat{\omega})$.

[17] The function f is well-defined because of Property 4 of Definition 1.

[18] $\mathcal{C}$ is AGM-consistent if, for every valuation $V : S_0 \rightarrow 2^{\Omega}$, the (partial) belief revision function $\circledast_K : \Psi_0 \rightarrow 2^{\Phi_0}$ where $K = \{\phi \in \Phi_0 : f(\Omega) \subseteq |\phi|\}$, $\Psi_0 = \{\phi \in \Phi_0 : |\phi| \in \mathcal{E}\}$ and, for every $\phi \in \Psi_0$, $\circledast_K(\phi) = \{\psi \in \Phi_0 : f(|\phi|) \subseteq |\psi|\}$, can be extended to a full AGM belief revision function.

ϕ is a Boolean formula, $\lceil\phi\rceil_{t'} = \lceil\phi\rceil_t$, so that $\mathcal{B}_t(\hat{\omega}) \subseteq \lceil\phi\rceil_t$, that is, $(\hat{\omega}, t) \models B\phi$. Hence $(\hat{\omega}, t) \models I\psi \rightarrow B\phi$ and thus, since t was chosen arbitrarily with $\hat{t} \rightarrowtail t$, $(\hat{\omega}, \hat{t}) \models \bigcirc(I\psi \rightarrow B\phi)$. Conversely, fix a frame that violates Property 4 of Definition 1. Then there exist $\omega \in \Omega$ and $t, t_1, t_2 \in T$ such that $t \rightarrowtail t_1, t \rightarrowtail t_2, \mathcal{I}_{t_1}(\omega) = \mathcal{I}_{t_2}(\omega)$ and $\mathcal{B}_{t_1}(\omega) \neq \mathcal{B}_{t_2}(\omega)$. Without loss of generality we can assume that

$$\text{there exists an } \alpha \in \mathcal{B}_{t_2}(\omega) \text{ such that } \alpha \notin \mathcal{B}_{t_1}(\omega) \tag{9}$$

(otherwise renumber the two instants). Construct a model where, for some atomic propositions p and q, $\|p\| = \mathcal{I}_{t_1}(\omega) \times T$ and $\|q\| = \mathcal{B}_{t_1}(\omega) \times T$. Then $(\omega, t_1) \models Ip \wedge Bq$ and thus, since $t \rightarrowtail t_1$, $(\omega, t) \models \Diamond(Ip \wedge Bq)$. Furthermore, since $\mathcal{I}_{t_1}(\omega) = \mathcal{I}_{t_2}(\omega)$, $(\omega, t_2) \models Ip$ and, by (9), $(\omega, t_2) \not\models Bq$, so that $(\omega, t_2) \not\models (Ip \rightarrow Bq)$. Hence, since $t \rightarrowtail t_2$, $(\omega, t) \not\models \bigcirc(Ip \rightarrow Bq)$ and thus Axiom 6 is falsified at (ω, t).

It is shown in [5] that Axiom $5a$ of Proposition 3 (called ND in [5]) is characterized by the following property

$$\text{if } t \rightarrowtail t' \text{ and } \mathcal{B}_t(\omega) \cap \mathcal{I}_{t'}(\omega) \neq \varnothing \text{ then } \mathcal{B}_{t'}(\omega) \subseteq \mathcal{B}_t(\omega) \tag{10}$$

and Axiom $5b$ of Proposition 3 (called NA in [5]) is characterized by the following property

$$\text{if } t \rightarrowtail t' \text{ then } \mathcal{B}_t(\omega) \cap \mathcal{I}_{t'}(\omega) \subseteq \mathcal{B}_{t'}(\omega) \tag{11}$$

Since Property 5 of Definition 1 implies both (10) and (11), it follows that a frame that satisfies Property 5 validates Axioms $5a$ and $5b$. Furthermore, in the presence of Property 1 of Definition 1, the conjunction of (10) and (11) implies Property 5. Thus, in the presence of Property 1, violation of Property 5 implies violation of either (10) or (11) (or both) and thus leads to the possibility of falsifying either Axiom $5a$ or Axiom $5b$ (or both).

We conclude the proof of Proposition 3 by showing that Axiom 6 is characterized by Property PLS of Proposition 1. Fix a temporal belief revision frame that satisfies property PLS, an arbitrary model based on it, arbitrary Boolean formulas $\phi_1, ..., \phi_n$ and $\chi_1, ..., \chi_n$ and arbitrary $\hat{\omega} \in \Omega$ and $\hat{t} \in T$ and suppose that (letting $\phi_0 = \phi_n$)

$$(\hat{\omega}, \hat{t}) \models \bigwedge_{j=1,2,...,n} \Diamond \left(I\phi_j \wedge \neg B \neg \phi_{j-1} \wedge B\chi_j\right) \tag{12}$$

We have to show that, for every $j = 1, ..., n$ (letting $\phi_0 = \phi_n$ and $\chi_0 = \chi_n$)

$$(\hat{\omega}, \hat{t}) \models \bigcirc \Big((I\phi_j \rightarrow B(\phi_{j-1} \rightarrow \chi_{j-1})) \wedge (I\phi_{j-1} \rightarrow B(\phi_j \rightarrow \chi_j)) \Big).$$

By (12) there exist $t_1, ..., t_n \in T$ such that $\hat{t} \rightarrowtail t_j$ for all $j = 1, ..., n$ and

$$\begin{array}{l} (\hat{\omega}, t_1) \models I\phi_1 \wedge \neg B \neg \phi_n \wedge B\chi_1 \text{ and} \\ (\hat{\omega}, t_j) \models I\phi_j \wedge \neg B \neg \phi_{j-1} \wedge B\chi_j \text{ for all } j = 2, ..., n. \end{array} \tag{13}$$

Thus

$$\begin{array}{l}\text{(a) } \mathcal{I}_{t_j}(\hat{\omega}) = \left\lceil \phi_j \right\rceil_{t_j} \text{ for all } j = 1, ..., n, \\ \text{(b) } \mathcal{B}_{t_j}(\hat{\omega}) \cap \mathcal{I}_{t_{j-1}}(\hat{\omega}) \neq \varnothing \text{ for all } j = 2, ..., n, \\ \text{(c) } \mathcal{B}_{t_1}(\hat{\omega}) \cap \mathcal{I}_{t_n}(\hat{\omega}) \neq \varnothing \\ \text{(d) } \mathcal{B}_{t_j}(\hat{\omega}) \subseteq \left\lceil \chi_j \right\rceil_{t_j} \text{ for all } j = 1, ..., n. \end{array} \tag{14}$$

Fix arbitrary $j \in \{1, ..., n\}$ and $t \in T$ with $\hat{t} \rightarrowtail t$. We have to show that if $(\hat{\omega}, t) \models I\phi_j$ then $(\hat{\omega}, t) \models B(\phi_{j-1} \rightarrow \chi_{j-1})$ and if $(\hat{\omega}, t) \models I\phi_{j-1}$ then $(\hat{\omega}, t) \models B(\phi_j \rightarrow \chi_j)$. Suppose first that $(\hat{\omega}, t) \models I\phi_j$, that is, $\mathcal{I}_t(\hat{\omega}) = \left\lceil \phi_j \right\rceil_t$. Since ϕ_j is a Boolean formula, by Proposition 5 in [4], $\left\lceil \phi_j \right\rceil_t = \left\lceil \phi_j \right\rceil_{t_j}$, so that, by (a) of (14), $\mathcal{I}_t(\hat{\omega}) = \mathcal{I}_{t_j}(\hat{\omega})$. It follows from this and Property 4 of Definition 1, that $\mathcal{B}_t(\hat{\omega}) = \mathcal{B}_{t_j}(\hat{\omega})$. Thus without loss of generality we can take $t = t_j$. Similarly, if $(\hat{\omega}, t) \models I\phi_{j-1}$ then, without loss of generality, we can take $t = t_{j-1}$. Thus it will be sufficient to show that if $(\hat{\omega}, t_j) \models I\phi_j$ then $(\hat{\omega}, t_j) \models B(\phi_{j-1} \rightarrow \chi_{j-1})$ and if $(\hat{\omega}, t_{j-1}) \models I\phi_{j-1}$ then $(\hat{\omega}, t_{j-1}) \models B(\phi_j \rightarrow \chi_j)$. By (b) and (c) of (14) and property PLS we have that (letting $t_0 = t_n$)

$$\mathcal{I}_{t_{j-1}}(\hat{\omega}) \cap \mathcal{B}_{t_j}(\hat{\omega}) = \mathcal{B}_{t_{j-1}}(\hat{\omega}) \cap \mathcal{I}_{t_j}(\hat{\omega}). \tag{15}$$

By (d) of (14), $\mathcal{B}_{t_{j-1}}(\hat{\omega}) \subseteq \left\lceil \chi_{j-1} \right\rceil_{t_{j-1}}$ and, since χ_{j-1} is a Boolean formula, by Proposition 5 in [4], $\left\lceil \chi_{j-1} \right\rceil_{t_{j-1}} = \left\lceil \chi_{j-1} \right\rceil_{t_j}$. Thus

$$\mathcal{B}_{t_{j-1}}(\hat{\omega}) \subseteq \left\lceil \chi_{j-1} \right\rceil_{t_j} \tag{16}$$

Hence, by (15) and (16),

$$\mathcal{I}_{t_{j-1}}(\hat{\omega}) \cap \mathcal{B}_{t_j}(\hat{\omega}) \subseteq \left\lceil \chi_{j-1} \right\rceil_{t_j} \tag{17}$$

Now (letting $\neg E$ denote the complement E, that is, $\neg E = \Omega \backslash E$),

$$\mathcal{B}_{t_j}(\hat{\omega}) \subseteq \ \neg \mathcal{I}_{t_{j-1}}(\hat{\omega}) \cup \left(\mathcal{I}_{t_{j-1}}(\hat{\omega}) \cap \mathcal{B}_{t_j}(\hat{\omega}) \right). \tag{18}$$

By (a) of (14), $\mathcal{I}_{t_{j-1}}(\hat{\omega}) = \left\lceil \phi_{j-1} \right\rceil_{t_{j-1}}$ and, since ϕ_{j-1} is a Boolean formula, $\left\lceil \phi_{j-1} \right\rceil_{t_{j-1}} = \left\lceil \phi_{j-1} \right\rceil_{t_j}$. Thus

$$\neg \mathcal{I}_{t_{j-1}}(\hat{\omega}) = \ \neg \left\lceil \phi_{j-1} \right\rceil_{t_j} = \left\lceil \neg \phi_{j-1} \right\rceil_{t_j} \tag{19}$$

Putting together (18), (19) and (17) we get that $\mathcal{B}_{t_j}(\hat{\omega}) \subseteq \left\lceil \neg \phi_{j-1} \right\rceil_{t_j} \cup \left\lceil \chi_{j-1} \right\rceil_{t_j} = \left\lceil \phi_{j-1} \rightarrow \chi_{j-1} \right\rceil_{t_j}$, that is, $(\hat{\omega}, t_j) \models B(\phi_{j-1} \rightarrow \chi_{j-1})$. The proof that if $(\hat{\omega}, t_{j-1}) \models I\phi_{j-1}$ then $(\hat{\omega}, t_{j-1}) \models B(\phi_j \rightarrow \chi_j)$ is along the same lines. [19]

Conversely, fix a frame that violates property PLS. Then there exist $\hat{\omega} \in \Omega, \hat{t} \in T$, $t_1, ..., t_n \in \hat{t}^{\rightarrowtail}$, and a $k^* \in \{1, ..., n\}$ such that (letting $t_0 = t_n$)

[19] By (d) of (14) $\mathcal{B}_{t_j}(\hat{\omega}) \subseteq \left\lceil \chi_j \right\rceil_{t_j}$ and since χ_j is Boolean, $\left\lceil \chi_j \right\rceil_{t_j} = \left\lceil \chi_j \right\rceil_{t_{j-1}}$. Thus, using (15), we get that $\mathcal{B}_{t_{j-1}}(\hat{\omega}) \cap \mathcal{I}_{t_j}(\hat{\omega}) \subseteq \left\lceil \chi_j \right\rceil_{t_{j-1}}$. Since $\mathcal{B}_{t_{j-1}}(\hat{\omega}) \subseteq \ \neg \mathcal{I}_{t_j}(\hat{\omega}) \cup \left(\mathcal{I}_{t_j}(\hat{\omega}) \cap \mathcal{B}_{t_{j-1}}(\hat{\omega}) \right)$ and $\mathcal{I}_{t_j}(\hat{\omega}) = \left\lceil \phi_j \right\rceil_{t_j} = \left\lceil \phi_j \right\rceil_{t_{j-1}}$, it follows that $\mathcal{B}_{t_{j-1}}(\hat{\omega}) \subseteq \left\lceil \neg \phi_j \right\rceil_{t_{j-1}} \cup \left\lceil \chi_j \right\rceil_{t_{j-1}} = \left\lceil \phi_j \rightarrow \chi_j \right\rceil_{t_{j-1}}$.

$$\begin{array}{l} \text{(a) } \mathcal{I}_{t_{k-1}}(\omega) \cap \mathcal{B}_{t_k}(\omega) \neq \varnothing, \forall k = 1, ..., n, \\ \text{(b) } \mathcal{I}_{t_{k^*-1}}(\hat{\omega}) \cap \mathcal{B}_{t_{k^*}}(\hat{\omega}) \neq \mathcal{B}_{t_{k^*-1}}(\hat{\omega}) \cap \mathcal{I}_{t_{k^*}}(\hat{\omega}). \end{array} \tag{20}$$

Let $p_1, ..., p_n, q_1, ..., q_n$, be atomic propositions and construct a model where, for every $k = 1, ..., n$, $\|p_k\| = \mathcal{I}_{t_k}(\hat{\omega}) \times T$ and $\|q_k\| = \mathcal{B}_{t_k}(\hat{\omega}) \times T$. Then, by (a) of (20) (letting $p_0 = p_n$)

$$(\hat{\omega}, \hat{t}) \models \bigwedge_{j=1,2,...,n} \Diamond \left(Ip_j \wedge \neg B \neg p_{j-1} \wedge Bq_j \right). \tag{21}$$

By (b) of (20), either
(A) there is an $\alpha \in \mathcal{I}_{t_{k^*-1}}(\hat{\omega}) \cap \mathcal{B}_{t_{k^*}}(\hat{\omega})$ such that $\alpha \notin \mathcal{B}_{t_{k^*-1}}(\hat{\omega}) \cap \mathcal{I}_{t_{k^*}}(\hat{\omega})$ or
(B) there is a $\beta \in \mathcal{B}_{t_{k^*-1}}(\hat{\omega}) \cap \mathcal{I}_{t_{k^*}}(\hat{\omega})$ such that $\beta \notin \mathcal{I}_{t_{k^*-1}}(\hat{\omega}) \cap \mathcal{B}_{t_{k^*}}(\hat{\omega})$.

Consider Case A first. Since $\alpha \in \mathcal{B}_{t_{k^*}}(\hat{\omega})$ and, by property (1) of Definition 1, $\mathcal{B}_{t_{k^*}}(\hat{\omega}) \subseteq \mathcal{I}_{t_{k^*}}(\hat{\omega})$, it must be that $\alpha \notin \mathcal{B}_{t_{k^*-1}}(\hat{\omega})$, so that $(\alpha, t) \models \neg q_{k^*-1}$, for every $t \in T$. Since $\alpha \in \mathcal{I}_{t_{k^*-1}}(\hat{\omega})$, $(\alpha, t) \models p_{k^*-1}$, for every $t \in T$. Thus $(\alpha, t) \models \neg(p_{k^*-1} \rightarrow q_{k^*-1})$, for every $t \in T$, in particular $(\alpha, t_{k^*}) \models \neg(p_{k^*-1} \rightarrow q_{k^*-1})$. Since $\alpha \in \mathcal{B}_{t_{k^*}}(\hat{\omega})$, it follows that $(\hat{\omega}, t_{k^*}) \models \neg B(p_{k^*-1} \rightarrow q_{k^*-1})$, so that, since $(\hat{\omega}, t_{k^*}) \models Ip_{k^*}$, $(\hat{\omega}, t_{k^*}) \models \neg(Ip_{k^*} \rightarrow B(p_{k^*-1} \rightarrow q_{k^*-1}))$. It follows from this and the fact that $\hat{t} \rightarrowtail t_{k^*}$ that $(\hat{\omega}, \hat{t}) \models \neg \bigcirc (Ip_{k^*} \rightarrow B(p_{k^*-1} \rightarrow q_{k^*-1}))$. This, together with (21) falsifies Axiom 6 of Proposition 3 at $(\hat{\omega}, \hat{t})$.

Now consider Case B. Since $\beta \in \mathcal{B}_{t_{k^*-1}}(\hat{\omega})$ and $\mathcal{B}_{t_{k^*-1}}(\hat{\omega}) \subseteq \mathcal{I}_{t_{k^*-1}}(\hat{\omega})$, it must be that $\beta \notin \mathcal{B}_{t_{k^*}}(\hat{\omega})$, so that $(\beta, t) \models \neg q_{k^*}$, for every $t \in T$. Since $\beta \in \mathcal{I}_{t_{k^*}}(\hat{\omega})$, $(\beta, t) \models p_{k^*}$, for every $t \in T$. Thus $(\beta, t) \models \neg(p_{k^*} \rightarrow q_{k^*})$, for every $t \in T$, in particular $(\beta, t_{k^*-1}) \models \neg(p_{k^*} \rightarrow q_{k^*})$. Since $\beta \in \mathcal{B}_{t_{k^*-1}}(\hat{\omega})$, it follows that $(\hat{\omega}, t_{k^*-1}) \models \neg B(p_{k^*} \rightarrow q_{k^*})$, so that, since $(\hat{\omega}, t_{k^{-1}*}) \models Ip_{k^*-1}$, $(\hat{\omega}, t_{k^*-1}) \models \neg(Ip_{k^*-1} \rightarrow B(p_{k^*} \rightarrow q_{k^*})$. It follows from this and the fact that $\hat{t} \rightarrowtail t_{k^*-1}$ that $(\hat{\omega}, \hat{t}) \models \neg \bigcirc (Ip_{k^*-1} \rightarrow B(p_{k^*} \rightarrow q_{k^*}))$. This, together with (21) falsifies Axiom 6 of Proposition 3 at $(\hat{\omega}, \hat{t})$.

■

References

[1] Alchourrón, Carlos, Peter Gärdenfors and David Makinson, On the logic of theory change: partial meet contraction and revision functions, *The Journal of Symbolic Logic*, 1985, 50: 510-530.

[2] van Benthem, Johan, Dynamic logics for belief change, *Journal of applied non-classical logics,* 17 (2007), 129–155.

[3] van Benthem, Johan and Cédric Dégremont, Multi-agent belief dynamics: bridges between dynamic doxastic and doxastic temporal logics, Technical report, ILLC, University of Amsterdam, 2008.

[4] Bonanno, Giacomo, Axiomatic characterization of the AGM theory of belief revision in a temporal logic, *Artificial Intelligence*, 171 (2007), 144-160.

[5] Bonanno, Giacomo, Belief revision in a temporal framework, in Krzysztof R. Apt and Robert van Rooij (editors), *New Perspectives on Games and Interaction*, Texts in Logic and Games Series, Amsterdam University Press, 2008, pp. 45-79.

[6] Bonanno, Giacomo, Rational choice and AGM belief revision, Working Paper, University of California, Davis, October 2008

[7] Darwiche, Adnan and Judea Pearl, On the logic of iterated belief revision, *Artificial Intelligence,* 89 (1997), 1-29.

[8] van Ditmarsch, Hans, Wiebe van der Hoek and Barteld Kooi, *Dynamic epistemic logic*, Springer, 2008.

[9] Friedman, Nir and Joseph Halpern, Belief revision: a critique, *Journal of Logic, Language, and Information*, 1999, 8: 401–420.

[10] Gärdenfors, Peter, *Knowledge in flux: modeling the dynamics of epistemic states*, MIT Press, 1988.

[11] Hansson, Bengt, Choice structures and preference relations, *Synthese*, 1968, 18: 443-458.

[12] Katsuno, Hirofumi and Alberto O. Mendelzon, On the difference between updating a knowledge base and revising it, in Peter Gärdenfors (editor), *Belief revision*, Cambridge University Press, 1992, 183–203.

[13] Nayak, A., M. Pagnucco and P. Peppas, Dynamic belief revision operators, *Artificial Intelligence*, 146 (2003), 193-228.

But what will everyone say? – Public Announcement Games*

Thomas Ågotnes† Hans van Ditmarsch‡

February 27, 2009

Abstract

Dynamic epistemic logics describe the epistemic consequences of actions. Public announcement logic, in particular, describe the consequences of public announcements. As such, these logics are *descriptive* – they describe what agents *can* do. In this paper we discuss what rational agents *will* or *should* do. We consider situations where each agent has a goal, a typically epistemic formula he or she would like to become true, and where the available actions are public announcements. What will each agent announce, assuming common knowledge of the situation? The truth value of the goal formula typically depends on the announcements made by several agents, hence we have a game theoretic scenario. We discuss possible solutions of such *public announcement games*.

1 Knowledge and Games, Wiebe and Us

Thomas. I knew Wiebe long before he knew me; first and foremost from his published work and reputation but also from seeing him at a conference or two. But our first real contact was at a *Knowledge and Games* workshop in Liverpool. I was a fresh PhD graduate, and had discovered what I thought to be a mistake in a paper by Wiebe and Mike Wooldridge, and this was the topic of my talk at the workshop. One could say that there was some nervousness: I didn't know how it would be received, if it really was a good idea or more like an academic *harakiri*. And while the problem was important in my mind then, in hindsight it admittedly looks slightly less significant. But Wiebe took it all very graciously. And it turned out very well in the end, because it was the starting point of a happy collaboration over many years now, in particular on formalising aspects of knowledge and games.

*For the workshop on *Reasoning about Knowledge and Rational Action* in honour of Professor Wiebe van der Hoek on his 50th birthday (*Wiebe Fest 2009*).

†Bergen University College, Norway, tag@hib.no, and University of Bergen, Norway, thomas.agotnes@infomedia.uib.no

‡University of Aberdeen, UK, and University of Otago, New Zealand, hans@cs.otago.ac.nz

Five years later, Hans and I are trying to secretly write this paper for *Wiebe Fest*. But what do you tell Wiebe when he asks what you are currently working on, if you are busy, or indeed what stops you from concentrating more on certain other projects? My naïve strategy is to not say anything, and blame laziness. But, Hans argues, if all of Wiebe's colleagues suddenly become quiet right before his anniversary, Wiebe will understand that something is going on. So a better strategy is perhaps to say that you are working on *something*, vaguely. But, then again, Wiebe would also know that that is the best strategy in case something is going on in secret; in that case isn't the first strategy better after all? It is not so easy to play knowledge games with Wiebe van der Hoek.

Formal models of knowledge and games have been central in Wiebe's research, and is the topic of this paper. In particular, we combine game theory with epistemic logic [7] and epistemic *dynamic* logic [12].

Hans. Even though I am known as Cluedo man, Wiebe is at the origin of this work. My first contact with Wiebe was in 1992 (or was it 1991?), when I was working at the Open University of the Netherlands, and Wiebe was finishing his PhD at the Vrije Universiteit Amsterdam. At the time, I was developing a follow-up course in logic for the Open University of the Netherlands, based on the Dutch-language textbook 'Logica voor informatici' ('Logic for Computer Science Students', currently named 'Logica voor informatica'—'for Computer Science'). The course would have a part on epistemic logic and Wiebe volunteered—for a nice fee—to write that part. I had come to Amsterdam to discuss matters with him. I recall entering a crowded and overly warm PhD office with this guy sitting there, surrounded by machinery and lots of books: Wiebe. Must have been sometime in winter, it was already getting somewhat dark.

Now one of these Open University things is that apart from feeding students *content*, epistemic logical content in this case, you also give them a *case study*, something to apply that content to. Wiebe told me about this game I had never heard of, called Cluedo, that might constitute a suitable case study in epistemic logic... I liked the idea, and it got into the tentative plans for that course. Things did not run that course exactly at the time. Eventually we had too much course material (other contributors to that course were, to name a few, Catholijn Jonker, and Maarten de Rijke), the material on epistemic logic was cut down a bit, and the case study on Cluedo never materialized. 'A bit' is kind of an understatement: at a public meeting with various contributors I proposed to butcher large chunks of already developed and written up material, acting in true fashion of one of my lesser known very impatient personae. Maarten almost exploded (and I assure you he had reason to), Wiebe never raised an eyebrow.

The topic of Cluedo kept roaming my brain, and came out again when, a few years later, I discussed PhD topics with possible supervisors. I am grateful that my paths crossed with Wiebe again at that stage, as, informally, he became much involved with the supervision of my PhD. The rest of *that* is history. I remember

having him murdered at my PhD defence—oh no, it was Jan van Maanen who got killed. Wiebe played a part in solving the murder prior to the actual defence.

The epistemic logic of knowledge games such as Cluedo is now well understood, also in a temporal epistemic setting [2]. But the game theory never really got started. More *knowledge* games than knowledge *games* therefore. Minor results are found in [10, 11]. Recently I thought to be pleasantly surprised, when reading `http://www.scientificblogging.com/news_releases/game_theory_solves_clue_and_maybe_improves_robot_mine_sweepers_too`. Game theory solves Clue! (Clue is the American name for the (British) game of Cluedo.) Alas, not so. This turned out to be the usual media hype. In the underlying contribution we address the matter of knowledge *games*. This approach is applicable to Cluedo.

2 Introduction

Dynamic epistemic logics describe the epistemic consequences of actions. Public announcement logic, in particular, describe the consequences of public announcements. As such, these logics are *descriptive* – they describe what agents *can* do, what pre- and post- conditions are, and so on. However, there is little *predictive* work in this area, describing what rational agents *will* do. In this paper we consider situations where each agent has a goal, a typically epistemic formula he or she would like to become true, and where the available actions are public announcements. What will each agent announce, assuming common knowledge of the situation? The truth value of the goal formula typically depends on the announcements made by several agents, hence we have a game theoretic scenario.

We make the following assumptions:

- agents have incomplete information about the world;
- agents have goals in the form of epistemic formulae, and agents' goals are common knowledge among all agents;
- each agent choose a (truthful) announcement (a formula she knows to be true);
- all agents make their announcements simultaneously; and
- all agents act rationally, i.e., they try to obtain their goals.

What can we say about how such agents will, or should, act?

In the next section we review the syntax and semantics of public announcement logic and some concepts from game theory. In Section 4 we introduce a formal model of *public announcement games*, and we discuss some possible solution concepts in Section 5 before we conclude in Section 6.

3 Background

3.1 Public Announcement Logic

The language $\mathcal{L}_{pal}$ of public announcement logic (PAL) [9] over a set of agents $N = \{1, \ldots, n\}$ and a set of primitive propositions Θ is defined as follows, where i is an agent and $p \in \Theta$:

$$\varphi ::= p \mid K_i\varphi \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid [\varphi_1]\varphi_2$$

We write $\langle\phi_1\rangle\phi_2$ resp. $\hat{K}_i\varphi$ for the duals $\neg[\phi_1]\neg\phi_2$ and $\neg K_i\neg\varphi$.

A *Kripke structure* over N and Θ is a tuple $M = (S, \sim_1, \ldots, \sim_n, V)$ where S is a set of states, $\sim_i \subseteq S \times S$ is an epistemic indistinguishability relation and is assumed to be an equivalence relation for each agent i, and $V : \Theta \to S$ assigns primitive propositions to the states in which they are true. A *pointed Kripke structure* is a pair (M, s) where s is a state in M. The interpretation of formulae in a pointed Kripke structure is defined as follows (the other clauses are defined in usual truth-functional way).

$M, s \models K_i\phi$ iff for every t such that $s \sim_i t$, $M, t \models \phi$

$M, s \models [\phi]\psi$ iff $M, s \models \phi$ implies that $M|\phi, s \models \psi$

where $M|\phi = (S', \sim'_1, \ldots, \sim'_n, V')$ such that $S' = \{s' \in S : M, s' \models \phi\}$; $\sim'_i = \sim_i \cap (S' \times S')$; $V'(p) = V(p) \cap S'$.

The purely epistemic fragment of the language (i.e., formulae not containing public announcement operators $[\phi]$) is denoted $\mathcal{L}_{el}$. It was already shown in Plaza's original publication on that logic [9] that the language of PAL is no more expressive than the purely epistemic fragment.

In this paper we will implicitly assume that Kripke structures are *finite* and *connected*.

3.2 Strategic Games

An *strategic game* is a tuple $G = \langle N, \{A_i : i \in N\}, \{u_i : i \in N\}\rangle$ where

- N is the finite set of *players*
- for each $i \in N$, A_i is the set of *strategies* (or *actions*) available to i. $A = \times_{j \in N} A_j$ is the set of *strategy profiles*.
- for each $i \in N$, $u_i : A \to \mathbb{R}$ is the *payoff function* for i, mapping each strategy profile to a number.

A strategy profile is a (pure strategy) *Nash equilibrium* if every strategy is the *best response* of that agent to the strategies of the other agents, i.e., if the agent can not do any better by choosing a different strategy given that the strategies of the other agents are fixed. In this paper we do not consider mixed strategies, and by

"Nash equilibrium" we implicitly mean the pure strategy variant. A strategy for an agent is *weakly dominant* if it is as least as good for that again as any other strategy, no matter which strategies the other agents choose.

4 Public Announcement Games

Formally, a *public announcement game* models the agents' knowledge, and thereby available announcements, and goals:

Definition 1 (Public Announcement Game) *An (n-player) public announcement game (PAG) is a tuple*

$$AG = \langle M, \gamma_1, \ldots, \gamma_n \rangle$$

where M is an epistemic structure, and $\gamma_i \in \mathcal{L}_{el}$ is the goal formula *for agent i. A* pointed PAG *is a tuple (AG, s) where AG is a PAG and s a state in AG. A* strategy *for agent i in a pointed PAG is a formula ϕ_i such that $M, s \models K_i\phi_i$.*

It is now very natural to associate a strategic game with any pointed PAG (AG, s): strategies, or actions, correspond to the individual announcements the agents can choose between, and a goal is satisfied iff it is true after all agents simultaneously make their chosen announcement. Formally:

Definition 2 (State Game) *The state game $G(AG, s)$ associated with state s of PAG $AG = \langle M, \gamma_i, \ldots, \gamma_n \rangle$ is defined by $N = \{1, \ldots, n\}$, $A_i = \{\phi_i : M, s \models K_i\phi_i\}$ and*

$$u_i(\langle \phi_1, \ldots, \phi_n \rangle) = \begin{cases} 1 & M, s \models \langle K_1\phi_1 \wedge \cdots K_n\phi_n \rangle \gamma_i \\ 0 & \textit{otherwise} \end{cases}$$

Like in Boolean games [3, 8], binary utilities are implicit in public announcement games; an agent's goal is either satisfied or not.

A point to note is that all PAGs have infinitely many strategies. However, for all interesting purposes any PAGs over a finite epistemic structure can be *seen* as one with only finitely many strategies, since there can be only finitely many announcements with different epistemic content.

Example 3 *Consider the following formal model of a situation: a two-player pointed PAG $(\langle M, \gamma_{Ann}, \gamma_{Bill} \rangle, s)$, where M is the following structure*

$$\bullet_t^{\neg p_B, p_A} \overset{Ann}{\cdots} \bullet_s^{p_B, p_A} \overset{Bill}{\cdots} \bullet_u^{p_B, \neg p_A}$$

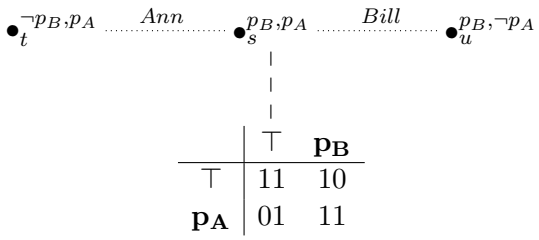
and

$$\gamma_{Ann} = (K_B p_A \vee K_B \neg p_A) \rightarrow (K_A p_B \vee K_A \neg p_B)$$
$$\gamma_{Bill} = (K_A p_B \vee K_A \neg p_B) \rightarrow (K_B p_A \vee K_B \neg p_A)$$

It is common knowledge that Ann's goal is that $Bill$ does not get to know whether p_A is true unless Ann gets to know whether p_B is true, and similarly the

other way around. Actually, p_A is true and Ann knows this, and the same for p_B and $Bill$. Ann does not know whether $Bill$ already knows p_A, and similarly for $Bill$/Ann/p_B. Furthermore, Ann does not know whether p_B is true, but she knows that if p_B is false then $Bill$ already knows that p_A is true, and similarly for $Bill$.

In s each agent can make two announcements with different information content, and the associated state game can thus be seen as a 2×2 matrix. We use the following picture to show that the game is associated with the point s:



	$\top$	$\mathbf{p_B}$
$\top$	11	10
$\mathbf{p_A}$	01	11

The figure above uses some notation we will use henceforth: Ann is assumed to be the row player and $Bill$ the column player; payoff is written xy where x is Ann's payoff and y is $Bill$'s.

Notice that the game above has two Nash equilibria: either both agents announce their private information, or neither say anything informative. A winning strategy, for either agent, is to say nothing.

So a pointed PAG models the type of situations described in Section 2, and it might be tempting at first sight to view a pointed PAG similarly to a Boolean game, and use the game theoretic tool chest to define rational outcomes based on the state game. For example, in Example 3 we identified two Nash equilibria in the state game. However, observe that in state s neither agent *knows* that the state actually is s – and thus they do not necessarily know what the state game is! It is a fundamental assumption behind solution concepts such as the Nash equilibrium that the strategic game is common knowledge. Since the state game is not common knowledge among the two agents, the identification of equilibria of the state game can therefore not be a reliable method of identifying rational outcomes. Figure 1 illustrates the state games associated with also the two other states of Example 3. Clearly, if the actual state is s, the state game is not known by any of the players – in fact, they don't even know all the actions available to the other player. Indeed, while (p_A, p_B) is a Nash equilibrium in the state game in s, it is not in the other state (t) which Ann considers possible – she does not even know for certain that p_B is a possible action for $Bill$.

Thus, the situations we are interested in can be modelled as a particular type of strategic games with imperfect information, where the strategies and information available in each state are closely interconnected (strategies *are* information) and where the same strategies are available in indiscernible states (but not necessarily in others). It is at least not immediately clear how standard models of strategic games with imperfect information, such as Bayesian games [4], can be applied to

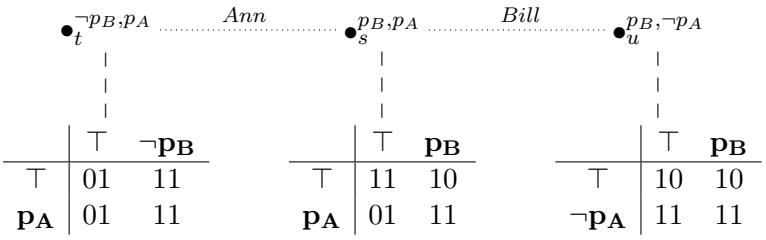


Figure 1: PAG of Example 3 with state games.

this setting (see also the discussion in Section 6). We now go on to discuss possible outcomes and solution concepts.

5 Solution Concepts

Let us first consider weakly dominant (wd) strategies. It should be clear from the discussion above that there is a crucial distinction, in a pointed PAG, between:

- the *existence* of a strategy for an agent which is weakly dominant for that agent on the one hand, and
- the existence of a strategy for an agent which that agent *knows* is weakly dominant on the other,

because it might be the case that there is a strategy which is weakly dominant in one of the states she considers possible, but not in another. For example, in state t of the model in Figure 1, p_A is weakly dominant for Ann, but Ann does not know this. Contrast this with the fact that $\top$ is also weakly dominant for Ann, but this Ann knows.

Knowledge of weakly dominant strategies is a natural solution concept for PAGs. But another important distinction must be made.

Definition 4 *Given a pointed PAG $(\langle M, \gamma_i, \ldots, \gamma_n \rangle, s)$ and an agent i, we say that i has a* weakly dominant strategy de dicto *iff i has a weakly dominant strategy in the state game of any state t such that $s \sim_i t$.*

Definition 5 *Given a pointed PAG $(\langle M, \gamma_i, \ldots, \gamma_n \rangle, s)$ and an agent i, we say that i has a* weakly dominant strategy de de *iff there is some strategy for i which is weakly dominant in the state game of any state t such that $s \sim_i t$.*

If an agent has a wd strategy *de dicto*, she knows *that* she has a wd strategy, i.e., she has a wd strategy in all states she considers possible, but she does not necessarily know *which* strategy is dominant; it is not necessarily the *same* strategy that is dominant in all the possible states. If she has a wd strategy *de re*, on the other hand, she knows *which* strategy is dominant; the same strategy is dominant in all

the states she considers possible. Of course, having a wd strategy *de re* implies having one *de dicto*. The *de dicto*/*de re* distinction is well known in the knowledge and action literature [6, 5]. In state s in the model in Figure 1, Ann has a wd strategy *de re*, namely $\top$ ($Bill$ also has one – which one?). An example where an agent has a wd strategy *de dicto* but not *de re* will be shown later (Example 11).

What about the Nash equilibrium? Clearly, we can have similar situations: there might be a Nash equilibrium without the agents knowing it; the agents might know that there is a Nash equilibrium but not necessarily know what it is (there might be different equilibria in different accessible states). However, what "know" means here is not as clear as in the case of dominant strategies where knowledge of a single agent was needed. In the case of the Nash equilibrium there are *several* agents involved. Group notions of knowledge, such everybody-knows, distributed knowledge and common knowledge, have been studied in the context of the *de dicto*/*de re* distinction before [6]. For our purpose, we argue that the proper type of group knowledge for knowing a Nash equilibrium *de re* is *common knowledge*, since that is the assumption in game theory. Common knowledge of an equilibrium among all agents corresponds to a common equilibrium in all states of the model (since we assume connectedness). Thus, the existence of a Nash equilibrium *de re* is a *model* property, rather than a *pointed* model property, unlike existence of dominant strategies.

Definition 6 *Given a PAG AG, we say that* there is a Nash equilibrium de re *if there exists a tuple of formulae, one for each agent, which constitutes a Nash equilibrium in the state game of every state in the PAG.*

For example, in the PAG in Figure 1, there is a Nash equilibrium *de re*, because the strategy profile $(\top, \top)$ is a Nash equilibrium in all the state games. An example where there are Nash equilibria in all the state games but no Nash equilibrium *de re* will be shown later (Example 11).

5.1 The Induced Game

Can a PAG be viewed as a (single) strategic game? We suggest the following definition.

Definition 7 *Given a PAG $AG = \langle M, \gamma_1, \ldots, \gamma_n \rangle$ with $M = (S, \sim_1, \ldots, \sim_n, V)$, the* induced game *$G_{AG}$ is defined as follows:*

- $N = \{1, \ldots, n\}$
- *A_i is the set of functions $a : S \to \mathcal{L}_{el}$ with the following properties:*
  - *Truthfulness: $M, s \models K_i a(s)$ for any s*
  - *Uniformity: $s \sim_i t \Rightarrow a(s) = a(t)$*

Thus, a strategy $a \in A_i$ gives a possible announcement for each state, but the same announcement for indiscernible states (note that the same announcements are always truthful in indiscernible states). Alternatively, a_i can be seen as a function mapping equivalence classes *to announcements.*

- *The payoffs are defined as follows. For any state s in AG, let $G(AG, s) = (N, \{A_i^s : i \in N\}, \{u_i^s : i \in N\})$ be the state game associated with s (Def. 2). Define, for any $(a_1, \ldots, a_n) \in A_1 \times \cdots \times A_n$:*

$$u_i(a_1, \ldots, a_n) = \frac{\sum_{s \in S} u_i^s(a_1(s), \ldots, a_n(s))}{|S|}$$

There are two important points to consider in the above definition.

First, strategies are defined as plans for action in *any* possible state. This may look counter-intuitive if we want to find rational actions in some particular state of a PAG: agents know the available actions in that state (the same actions are available in all the states an agent considers possible). However, even though the current state is a member of the equivalence class one agent currently considers as possible states, she might consider many possibilities for what *another* agent's current equivalence class might be. Thus, she must take into account what the other agent is likely to do in all of these circumstances. Thus, a strategy must be description of behaviour for any contingency; even though each agent will only choose actions that actually are available in the current state.

Second, payoff is computed by taking the average over *all states in the model*. It is clear that it does not suffice to look only in the current state, as each agent also might consider other states possible. But why not, then, compute an agent's payoff by taking the average over all the states that agent considers possible (the agent's equivalence class)? The reason is that the strategic game *must be common knowledge*, in order for solution concepts such that the Nash equilibrium to make sense. It might for example be that Ann considers it possible that $Bill$ considers state u possible, but that state u is not in either Ann's or $Bill$'s equivalence class for a current state s. If we take the average over only each agent's equivalence class for s, u will not be taken into account. Averaging over all reachable states corresponds to averaging over all states commonly considered possible (all states accessible according to the accessibility relation for common knowledge). This is also the reason that the induced game is not induced from a *pointed* PAG: the induced game is the same at all points. This is as it should be, since the game should be common knowledge at any state. The computed payoffs can be seen as *expected* payoffs, not expected by a particular agent in the game, but expected payoffs as computed by a *common knower* – an agent whose knowledge is exactly the common knowledge among all agents in the game.

We will shortly explain the induced game further through several examples.

Proposition 8 *If agent i has a weakly dominant strategy* de re *in (AG, s) for every state s in a PAG AG, there is a weakly dominant strategy for i in the induced game.*

Proof A weakly dominant strategy a in the induced game is defined by taking $a(s)$ to be a wd strategy in the state game in s and choosing the same strategy for all states int the same equivalence class (this is possible because the agent has a strategy *de re*). Wlog. assume that there are only two agents, and that $i = 1$. Suppose that a is not weakly dominant. Then there is some other strategy a' for 1, and some strategy b for 2 such that

$$\frac{\sum_{s \in M} u_1^s(a'(s), b(s))}{|M|} > \frac{\sum_{s \in M} u_1^s(a(s), b(s))}{|M|}$$

Since payoffs are positive, this implies that $u_1^s(a'(s), b(s)) > u_1^s(a(s), b(s))$ for some s. But then $a(s)$ is not weakly dominant in the state game in s after all, which is a contradiction. □

Definition 9 *A* Nash Announcement Equilibrium *(NAE) of a PAG is a Nash equilibrium of the induced game.*

Example 10 *Let us continue Example 3. We construct the induced game as follows (it is instructive to inspect the state games as illustrated in Fig. 1 on p. 7).* A_A *(for* Ann*) contains the following four strategies:*

- a_A^1*:* $t, s \mapsto \top$*;* $u \mapsto \top$
- a_A^2*:* $t, s \mapsto \top$*;* $u \mapsto \neg p_A$
- a_A^3*:* $t, s \mapsto p_A$*;* $u \mapsto \top$
- a_A^4*:* $t, s \mapsto p_A$*;* $u \mapsto \neg p_A$

A_B *(for* $Bill$*) is as follows:*

- a_B^1*:* $u, s \mapsto \top$*;* $t \mapsto \top$
- a_B^2*:* $u, s \mapsto \top$*;* $t \mapsto \neg p_B$
- a_B^3*:* $u, s \mapsto p_B$*;* $t \mapsto \top$
- a_B^4*:* $u, s \mapsto p_B$*;* $t \mapsto \neg p_B$

In order to compute the payoffs, we need to check the payoffs in the state games for

each state and combination of strategies. We have the following:

$\mathbf{a_A^x}, \mathbf{a_B^y}$	$\mathbf{t}$	$\mathbf{s}$	$\mathbf{u}$
$1,1$	01	11	10
$1,2$	11	11	10
$1,3$	01	10	10
$1,4$	11	10	10
$2,1$	01	11	11
$2,2$	11	11	10
$2,3$	01	10	11
$2,4$	11	10	11
$3,1$	01	01	10
$3,2$	11	01	10
$3,3$	01	11	10
$3,4$	11	11	10
$4,1$	01	01	11
$4,2$	11	01	11
$4,3$	01	11	11
$4,4$	11	11	11

We get the following payoff matrix. We will henceforth write the payoffs without *dividing by the number of states, for ease of presentation (the equilibria do of course not depend on this):*

	$\mathbf{a_B^1}$	$\mathbf{a_B^2}$	$\mathbf{a_B^3}$	$\mathbf{a_B^4}$
$\mathbf{a_A^1}$	$\underline{22}$	$\underline{32}$	21	31
$\mathbf{a_A^2}$	$\underline{23}$	32	22	32
$\mathbf{a_A^3}$	12	22	$\underline{22}$	$\underline{32}$
$\mathbf{a_A^4}$	13	23	$\underline{23}$	$\underline{33}$

The Nash equilibria are underlined.

Thus, the Nash announcement equilibria of this PAG are as follows, informally:

(1,1) *Both agents say nothing (informative), no matter what*

(1,2) *Ann says nothing, but $Bill$ says $\neg p_A$ if the state is t (which $Bill$ can discern from any other state) and nothing otherwise. Let us consider this in the case that the current state is s. Ann knows that the actual state is either s or t, but not which. Thus, in the equilibrium she will play $\top$ under the assumption that $Bill$ will play $\top$ if the actual state is s and $\neg p_A$ if the actual state is t ($Bill$ can discern between these two possibilities). Actually, $Bill$ will play $\top$.*

(2,1) *Similarly, with Ann and $Bill$ swapped*

(3,3) *Ann says p_A if she knows it, i.e., if the state is in Ann's equivalence class $\{s, t\}$. Similarly for $Bill$.*

(3,4) *Ann says p_A if she knows it, and Bill says p_B if he knows it and $\neg p_B$ if he knows that*

(4,3) *Similarly, for Ann and Bill swapped*

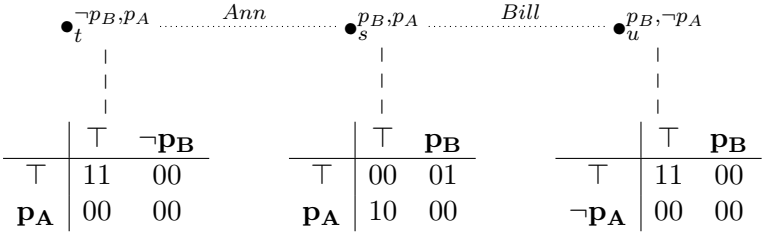
(4,4) *Both agents say everything they know*

Example 11 *Define a PAG AG as follows. Let the model be as in Example 3, but change the goals as follows:*

$$\gamma_{Ann} = (K_B(p_A \wedge p_B) \wedge \neg K_A p_B) \vee (K_B(\neg p_B \wedge p_A) \wedge \hat{K}_A \hat{K}_B \neg p_A) \vee (K_A(p_B \wedge \neg p_A) \wedge \hat{K}_B \hat{K}_A \neg p_B)$$

$$\gamma_{Bill} = (K_A(p_A \wedge p_B) \wedge \neg K_B p_A) \vee (K_B(\neg p_B \wedge p_A) \wedge \hat{K}_A \hat{K}_B \neg p_A) \vee (K_A(p_B \wedge \neg p_A) \wedge \hat{K}_B \hat{K}_A \neg p_B)$$

Perhaps the reader finds these long formulae hard to read, but it suffices to trust that they give the following state games:



	$\top$	$\neg \mathbf{p_B}$
$\top$	11	00
$\mathbf{p_A}$	00	00

	$\top$	$\mathbf{p_B}$
$\top$	00	01
$\mathbf{p_A}$	10	00

	$\top$	$\mathbf{p_B}$
$\top$	11	00
$\neg \mathbf{p_A}$	00	00

The PAG has some properties not found in the PAG in Example 3 (Figure 1). First, Ann has a weakly dominant strategy de dicto*, but not* de re*, in the pointed PAG (AG, s). The strategy p_A is weakly dominant in s, but not in t. There is, however, another weakly dominant strategy in t, namely $\top$. Second, while every state game has a Nash equilibrium, there does not exist a Nash equilibrium* de re *in AG.*

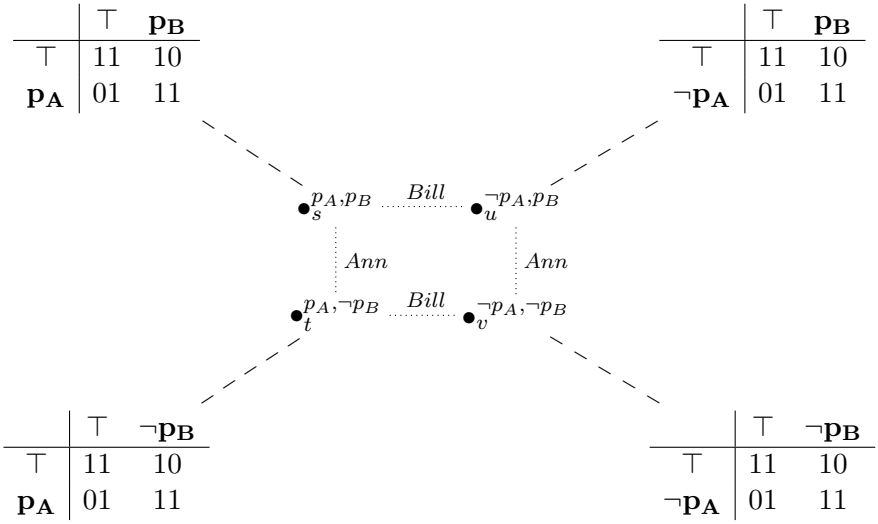
We get the following induced game, where the strategies are as in Example 10:

	$\mathbf{a^1_B}$	$\mathbf{a^2_B}$	$\mathbf{a^3_B}$	$\mathbf{a^4_B}$
$\mathbf{a^1_A}$	$\underline{22}$	11	$\underline{12}$	01
$\mathbf{a^2_A}$	11	00	$\underline{12}$	01
$\mathbf{a^3_A}$	$\underline{21}$	$\underline{21}$	00	00
$\mathbf{a^4_A}$	10	10	00	$\underline{00}$

The Nash announcement equilibria are underlined. Let us consider the situation in state s. There are several different Nash announcement equilibria, including: all agents announce $\top$ in all states (including in s). Note that $(\top, \top)$ is not *a Nash equilibrium in the state game in s. Another equilibrium is that both agents play "down" (i.e., not $\top$) in any state (also a Nash equilibrium in the state game in s). Note that (a^4_A, a^4_B)* is *a NAE while for example (a^3_A, a^4_B) is* not*. If the current state is s (or, from Ann's perspective the current equivalence class is $\{s, t\}$ and from Bill's perspective $\{s, u\}$), Ann will in fact do exactly the same if she uses strategy a^3_A or a^4_A. However, since Bill does not know whether Ann's equivalence class is $\{s, t\}$ or $\{u\}$, he must also consider what Ann does in u – which is exactly what differentiates a^3_A and a^4_A. Thus, the distinction between these two strategies is significant, even in a state (s) where they give the same action.*

Example 12 *Let us consider a more regular and symmetric PAG than the ones discussed so far. The situation is similar to the one in Example 3, but now Ann knows that $Bill$ does not know p_A, and similarly for $Bill$/Ann/p_B. The situation is modelled by the following goal formulae and Kripke structure. We have also shown the state games.*

$$\gamma_{Ann} = (K_B p_A \vee K_B \neg p_A) \rightarrow (K_A p_B \vee K_A \neg p_B)$$
$$\gamma_{Bill} = (K_A p_B \vee K_A \neg p_B) \rightarrow (K_B p_A \vee K_B \neg p_A)$$



Again, the induced game has four distinct strategies for each agent:

$\mathbf{x}$	$\mathbf{a^x_A}$		$\mathbf{a^x_B}$	
1	$s,t \mapsto \top$;	$u,v \mapsto \top$	$s,u \mapsto \top$;	$t,v \mapsto \top$
2	$s,t \mapsto \top$;	$u,v \mapsto \neg p_A$	$s,u \mapsto \top$;	$t,v \mapsto \neg p_B$
3	$s,t \mapsto p_A$;	$u,v \mapsto \top$	$s,u \mapsto p_B$;	$t,v \mapsto \top$
4	$s,t \mapsto p_A$;	$u,v \mapsto \neg p_A$	$s,u \mapsto p_B$;	$t,v \mapsto \neg p_B$

The induced game (Nash equilibria underlined):

	$\mathbf{a^1_B}$	$\mathbf{a^2_B}$	$\mathbf{a^3_B}$	$\mathbf{a^4_B}$
$\mathbf{a^1_A}$	$\underline{44}$	42	42	40
$\mathbf{a^2_A}$	24	33	33	42
$\mathbf{a^3_A}$	24	33	33	42
$\mathbf{a^4_A}$	04	24	24	$\underline{44}$

The game has two Nash equilibria. The first is that both agents say nothing, in all states*. The strategies in this equilibrium are both dominant strategies. The second equilibrium (a^4_A, a^4_B) is that both agents tell everything they know,* in all states*.*

In Example 12 the Nash announcement equilibria are all "composed" of Nash equilibria in the state game, in the following sense: for every NAE (a, b) and every

state s, $(a(s), b(s))$ is a Nash equilibrium in the state game in s (albeit not *all* such compositions of Nash equilibria in the state games are NAE in the example). Indeed, this is also the case in Example 3. Is this a general property of PAGs? No, and a counter example is found in Example 11: (a^1_A, a^1_B), because $(a^1_A(s), a^1_B(s))$ is not a Nash equilibrium in the state game in s.

We can establish a connection to having a Nash equilibrium *de re*, similarly to Proposition 8 for dominant strategies.

Proposition 13 *If there exists Nash equilibrium* de re *in a PAG, then there exists a Nash announcement equilibrium.*

Proof Assume wlog. that there are only two agents. If there is a Nash equilibrium *de re*, then there is a strategy profile (x, y) which is a Nash equilibrium in every state game. Let (a, b) be a strategy profile for the induced game such that $a(s) = x$ and $b(s) = y$ for any s. Clearly, a and b are both uniform and truthful. Suppose that (a, b) is not a Nash equilibrium in the induced game. Then there is a better response a' for one of the agents, again wlog. assume for agent 1. In other words, there is a strategy a' for agent 1 such that $u_1(a', b) > u_1(a, b)$. But this entails that $u^s_1(a'(s), b(s)) > u^s_1(a(s), b(s))$ for some state s, and thus that there is a strategy z for agent 1 in the state game in s such that $u^s_1(z, y) > u^s_1(x, y)$ – which contradicts the fact that (x, y) is a Nash equilibrium in the state game in s. □

Proposition 13 does not hold in the other direction. A counter example is found in Example 11.

6 Discussion

The intimate connection between knowledge and strategies in public announcement games distinguishes them from many other types of games. In *Boolean games* [3, 8], each agent has a goal formula like in PAGs, and each agent controls a set of primitive propositions which affects the truth value of the goal formulas. In contrast, in PAGs an agent "controls" common knowledge of any formula he or she knows. We have seen that we cannot simply view a pointed PAG as Boolean type game, because the agents do not necessarily have common knowledge about the game that is being played.

The most common model of strategic games with imperfect information is *Bayesian games* [4]. In Bayesian games it is, among other things, assumed that each agent has a probability measure over the set of states. This could perhaps be adapted to the possible worlds framework. However, it seems to be complicated by the fact that in the latter framework we can have situations like "Ann considers a state possible where B considers another state possible where..", while in the former framework there is only a "flat" probability measure and not different measures in different states. Still, it might be that our games can be seen as Bayesian

games, or the other way around. Our definition of the Nash announcement equilibrium has many similarities to the standard definition of Nash equilibria in Bayesian games. The exact relationship remains to be identified.

Is our definition of the Nash announcement equilibrium the right one? We have argued that it is reasonable and has desirable properties, e.g., it is an equilibrium of a game that is common knowledge among all agents, and the payoffs are expected payoffs computed by a "common knower". Further studies of its properties are needed, and this is work in progress. In future work we will also study mixed strategies, we will look at more fine grained goal models which do not necessarily give binary payoffs, for example lists of prioritised goals [1], and we will model situations with *sequential* announcements by using extensive form games.

References

[1] Thomas Ågotnes, Michael Wooldridge, and Wiebe van der Hoek. Normative system games. In M. Huhns and O. Shehory, editors, *Proceedings of the Sixth International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2007)*, pages 876–883. IFAMAAS, May 2007.

[2] C. Dixon. Using temporal logics of knowledge for specification and verification–a case study. *Journal of Applied Logic*, 4(1):50–78, 2006.

[3] P. Harrenstein. *Logic in Conflict*. PhD thesis, Utrecht University, 2004.

[4] J. C. Harsanyi. Games with Incomplete Information Played by 'Bayesian' Players, Parts I, II, and III. *Management Science*, 14:159–182, 320–334, 486–502, 1967–1968.

[5] Wojciech Jamroga and Thomas Ågotnes. Constructive knowledge: what agents can achieve under imperfect information. *Journal of Applied Non-Classical Logics*, 17(4):423–475, 2007.

[6] Wojciech Jamroga and Wiebe van der Hoek. Agents that know how to play. *Fundamenta Informaticae*, 63:185–219, 2004.

[7] J.-J.Ch. Meyer and W. van der Hoek. *Epistemic Logic for AI and Computer Science*. Cambridge Tracts in Theoretical Computer Science 41. Cambridge University Press, Cambridge, 1995.

[8] J.-J.Ch. Meyer P. Harrenstein, W. van der Hoek and C. Witteveen. Boolean games. In J. van Benthem, editor, *Proceeding of the Eighth Conference on Theoretical Aspects of Rationality and Knowledge (TARK VIII)*, pages 287–298, Siena, Italy, 2001.

[9] 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: Poster Session Program, pages 201–216. Oak Ridge National Laboratory, 1989. ORNL/DSRD-24.

[10] H.P. van Ditmarsch. The description of game actions in cluedo. In L.A. Petrosian and V.V. Mazalov, editors, *Game Theory and Applications*, volume 8, pages 1–28, Commack, NY, USA, 2002. Nova Science Publishers.

[11] H.P. van Ditmarsch. Some game theory of pit. In C. Zhang, H.W. Guesgen, and W.K. Yeap, editors, *Proceedings of PRICAI 2004 (Eighth Pacific Rim International Conference on Artificial Intelligence)*, pages 946–947. Springer, 2004. LNAI 3157.

[12] H.P. van Ditmarsch, W. van der Hoek, and B.P. Kooi. *Dynamic Epistemic Logic*, volume 337 of *Synthese Library*. Springer, 2007.

Dialogue Coherence: A Generation Framework

Robbert-Jan Beun and Rogier M. van Eijk
Utrecht University, Faculty of Science,
Department of Information and Computing Sciences,
P.O. Box 80089, 3508 TB Utrecht, The Netherlands
{rj, rogier}@cs.uu.nl

Abstract

This paper[1] presents a framework for the generation of coherent elementary conversational sequences at the speech act level. We will embrace the notion of a *cooperative dialogue game* in which two players produce speech acts to transfer relevant information with respect to their commitments. Central to the approach is that participants try to achieve some sort of balanced cognitive state as a result of speech act generation and interpretation. Cognitive states of the participants change as a result of the interpretation of speech acts and these changes provoke the production of a subsequent speech act. Describing the properties and the dynamics of the mental constructs that constitute the participants' cognitive states, such as beliefs and commitments, in relation to the various dialogue contributions is an essential aspect of the game. Although simple in its basic form, the framework enables us to produce abstract conversations with some properties that agree strikingly with coherence structures found in, for instance, Conversation Analysis.

1 Introduction

In its basic form, a dialogue can be conceived as a linear alternating sequence of symbolic elements between two participants who alternately play the role of sender and receiver (Hamblin, 1971). The utterances in a dialogue do not form independent segments of text, but show, like words in a sentence, a coherent structure of conversational units. In both the generation and the interpretation process, dialogue participants relate the content of new contributions to the previous discourse. The construction of a coherent representation of the conversational units in a dialogue is an essential cognitive activity in the process of understanding discourse.

Coherence as a relation in discourse has been studied from many different angles. Most studies in linguistics take an analytical stance and consider, for instance, the acceptability of written texts or dialogue sequences and the type of relations that should be taken into account to build a coherent representation of these discourses (e.g., [18, 31, 14, 33]). Inspired by computational theories on natural language processing and the need for computer systems that generate cooperative dialogue contributions, coherence is nowadays also studied from the perspective of language production

[1]This article has been published as Beun, R.J. & Eijk, R.M. van (2007). Dialogue Coherence: A Generation Framework *Journal of Logic, Language and Information* 16(4), pp. 365-385.

(see e.g., [4, 7, 20, 12, 22, 28]). In these 'synthetic' theories, questions have to be answered about, for instance, the appropriateness of responses in a dialogue, the adequate realisation of referential expressions or the linear ordering of discourse units as part of a dialogue turn. In these cases, coherence is often considered as a constraint on the communicative behaviour of the participants in terms of the production of speech acts or actual realization of the linguistic form.

In line with the 'synthetic' theories, the central goal of this paper is to present a computational framework that enables us to generate coherent elementary conversational sequences at the speech act level. For that, we will embrace the notion of a cooperative *dialogue game* [9] in which two players produce speech acts or 'moves' to transfer relevant information with respect to their cognitive states. Central to the approach, and in line with [23], is that the cognitive states of the players change as a result of the interpretation of the speech acts [8] and that these changes provoke the production of a subsequent move. The game works roughly like this: A speech act is generated on the basis of preconditions formed by the cognitive state of the sender, but changes the cognitive states of both sender and addressee after it has been manifested. In the next turn the addressee adopts the sender role and, subsequently, the changed mental constructs of his or her state function as the new preconditions for the next speech act. Conversational coherence of subsequent speech acts with previous utterances is established by the cognitive state of the participants and the rules for cognitive update and dialogue behaviour. As in realistic conversational situations, it is assumed that the information relevant with respect to a particular answer can be distributed among the participants.

Describing the properties and the dynamics of the mental constructs, such as beliefs and commitments [17, 35, 19], in relation to the various dialogue contributions is an essential part of the work. We will show that the coherence of the speech acts is tied to local interactions contingent to the agent's particular situation and that the coherence relations can be described in a situated sense, i.e., driven by the history of the speech acts and the dynamics of the mental constructs of the participants.

This paper is organized as follows. In Section 2 we will consider the notion of coherence in dialogue. In Section 3 we will describe the basic aspects of the generation model and the underlying theoretical principles of the dialogue game. In Section 4 we will describe the mental constructs used in the dialogue game and the notion of balance and in Section 5 the actual game is defined and in Section 6 an example is worked out. In Section 7, conclusions are given together with lines for further research.

2 Dialogue Coherence

Coherence relations can be described on a syntactic and a semantic level. Syntactically, most models of conversation include both a *linear* and a *hierarchical* conception of coherence relations. Linearity is established by a notion of pairing – two utterances that for some reason seem to be related to each other at the same level. In Conversation Analysis, for instance, the fundamental pairs of conversational organisation are sequences called 'adjacency pairs': a question is followed by an answer, a greeting by a counter-greeting, etcetera (e.g. [24]). From speech act theory we know the notion of 'uptake' [5], being the dependency of a successful performance of an illocutionary

act on the reaction of the addressee. Hierarchy, on the other hand, is established by embedded structures that may appear between paired units. In dialogue this can be created by so-called 'insertion sequences' – i.e. deviations from the main point that are usually expressed by the first part of an adjacency pair. Through the embedding structures of adjacency pairs, the *recursive* organisation of conversation becomes apparent. Similar structures can be found in, for instance, [30, 16, 29, 25]. A disadvantage of the syntactic approach is that it does not explain *why* particular sequences are conceived as being linear or hierarchical.

In line with, for instance, [31, 34, 6], we will assume that semantic coherence relations have an informational and/or an intentional nature. Informational coherence relations concern a 'believed' relation between the units that corresponds to an existing relation in the world described by the discourse (co-reference, spatio-temporal relations, causality, and the like). The informational view often refers to the participants' ontological and assertional knowledge about the discourse domain. Intentional relations do not lie in the state of affairs described, but in the assumed relations in the goals and plans of the participants (the illocutionary and perlocutionary relations, such as question-answer, offer-rejection and threat-defence).

In order to better define our intuition, we will first discuss some examples. All examples were taken from a small scale pilot experiment with 30 subjects who had to judge the degree of coherence on a scale from 1 (low coherence) to 5 (high coherence) of five short parts of Dutch discourse. The first two examples show (in)coherence at the informational level.

Ex1. *I read a book. The stone is black.* $(c = 1.4, sd = 0.6, n = 30)$

where c indicates the coherence score, sd the standard deviation and n the number of judgements. We can conclude from the low sd that in this case there was much agreement between the subjects. The next example had the highest score:

Ex2. *I read a book. The writer is a wise man.* $(c = 4.5, sd = 0.7, n = 30)$

The two examples show that neither object continuation nor coherence markers in the second sentence are decisive in the subjects' judgement. Subjects judged Ex2 as a coherent sequence, because they *know* that books are written by writers. In other words, since they know that there exists a relation between the objects in the first and the second sentence of Ex2, the subjects consider the two sentences as coherent utterances. In Ex1 the relation is absent, although, with some imagination, we could think of a situation where the 'I'-person is Fred Flintstone reading a book made of black stone. This implies that in the judgement of coherence the belief (or knowledge) state of the dialogue participants should be taken into account.

Let us now take a simple dialogue example:

Ex3. *A: Is John in the kitchen?*
B: What time is it?

$(c = 2, sd = 1.1, n = 30)$

Here, the coherence score is relatively low and there is less agreement between the subjects, but coherence is significantly higher than the coherence in Ex1 (Wilcoxon, $z = -2, 20, p < 0.05$). In the interview after the experiment, subjects who gave a relatively high score to Ex3 responded that they could imagine that B knows that John is always in the kitchen at a certain time. In other words, in a dialogue two subsequent utterances were considered as being coherent if there exists a informational relation in the last speaker's mind between the interpreted content of the utterance in the previous turn and the content of the utterance of this last speaker.

The next example from Sacks (discussed in [24]:

Ex4. *A: I have a fourteen year old son*
B: Well, that's all right
A: I also have a dog
B: Oh I'm sorry

$(c = 2.83, sd = 1.1, n = 30)$

seems a bizarre sequence in isolation, but when embedded in a context of goals and roles – A is trying to rent an apartment from landlord B – looks quite natural. Levinson uses the example to motivate the proposition that people have bad intuitions about the well-formedness of speech act sequences once they are taken out of the context. Although the coherence score is medium, we observe a relatively high score for sd (compared to Ex1 and Ex2). In the interview, subjects who gave Ex4 a high coherence score responded that they invented some sort of background, like the tenant-landlord situation; the other subjects could not imagine a situation where the dialogue makes sense. So, in order to understand the coherence relation, we also need information about the *goals* of the dialogue participants. Two subsequent utterances may seem incoherent, but apparently the sequence becomes coherent when people know the common goal of the dialogue.

In a related experiment, the following example was included:

Ex5. *A: What time is it?*
B: I have to go to the toilet.

$(c = 2.63, sd = 1.2, n = 30)$

The scores are in line with Ex4, but the utterances are not related in the same way. In Ex4, the landlord answers A's questions; in Ex5, B indicates that he gives priority to another goal, namely releasing the pressure of his bladder to answering A's question. In other words, B shifted from the common domain goal (i.e. answering the question) to a procedural goal by indicating that, for instance, the answer will be postponed. This was also the interpretation of subjects who gave the sequence a high score.

In general, dialogue coherence manifests itself as a subjective phenomenon and the judgement of coherence is a sliding scale that heavily depends on the background knowledge and goals of the dialogue participants. Also, coherence in a dialogue often cannot be established without the continuity or recurrence of semantic elements but the continuity does not have to be included in the surface structure of the utterances, as

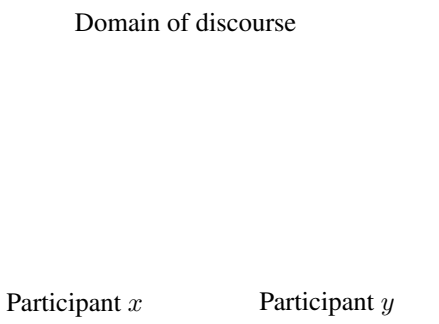


Figure 1: The triangle metaphor.

we have seen in the previous examples. So, in this paper, coherence is not considered as an intrinsic property of a text or a dialogue, but mainly as a mental phenomenon (c.f. [32, 13]). In the dialogues that will be generated below we will observe the elementary structural phenomenon, such as linear and hierarchical sequences based on the organization of the mental structure in terms of beliefs and goals of the participants and the rules that were used in the generation framework. Based on this framework, we will be able to show why particular sequences of discourse may be interpreted as linear and others as hierarchical. In what follows, we will describe the dialogue game and its underlying communication model.

3 The basic model

The dialogue game presented in this paper is based on a simple model employed in human-computer interaction ([21, 1]). Underlying the model is the recognition that humans interact naturally with their environment in two ways: symbolically and physically. On the one hand, if there is an intermediary interpreter, humans can interact symbolically and use language to give commands, ask for or provide information, etcetera. On the other hand, physically, one manipulates objects, for instance, by moving or fastening them, or observes them by seeing, feeling, hearing, tasting or smelling. The essential difference between the two types of interaction is that actions of the first type need an interpreter who can bridge the gap between the symbols and their actual meaning and purpose, while actions of the second type are related in a more direct manner to human perception and action.

In parallel with the distinction symbolical vs. physical, humans engaged in dialogues can perform two types of external actions: a. *communicative actions* intended to cause some cognitive effect in the recipient, and b. *non-communicative actions* to observe or change particular properties of the domain. Obviously, the two types of action can be considerably interrelated. In addition, we will include an action type that is neither communicative nor external, namely *inference* – i.e. the process of adjusting the cognitive states of participants solely based on their previous states. In short, the basic model includes perception, action, communication and thinking in an extremely rudimentary form.

The distinctive interaction channels are represented in the so-called triangle metaphor (Figure 1), where the corners represent the domain of discourse (or the external world)

and the two participants, and the arrows the flow of information between the participants and between the participants and the domain. A communicative act performed by participant x towards participant y is a flow of information from x to y; observation of the domain is a flow of information from the domain towards the observer and an action carried out in the domain is a flow of information from the actor to the domain. In practice, the channel between the two participants may cause messages to be delayed or disturbed by, for instance, noise. Also, the channel can be duplex, where both participants can speak at a time, or half-duplex, where only one participant can speak at the time. Here, we will consider the channel between the participants and between the participants and the domain of discourse as an ideal half-duplex channel, which means that no information is delayed or lost during transfer and that information can flow only in one direction at a time.

Discourse or 'semantic' information will be divided into two categories: a. perceivable facts (represented by single proposition letters p, q, ...) and b. inference relations. Two types of inference relations will be distinguished which enable the dialogue participants to reason about their beliefs (belief inference) and their commitments (commitment inference). The commitment inference enables the participants to develop sub-commitments; for instance, A wants to be outside and A is inside, then A has to open the door and, consequently, 'opening the door' becomes a commitment by A.

The two types of inferences will be denoted as follows: belief inferences will be of the form ('$p \rightarrow q$', ...) and commitment inferences of the form ('$p \times q \rightarrow r$', ...), connecting a simple proposition or a pair of simple propositions (the antecedent) with a simple proposition (the consequence), respectively. Intuitively, '$p \rightarrow q$' means that in all possible states, if p is true, then q is true; '$p \times q \rightarrow r$' means that if p is true in the current state (e.g., 'I am inside and the door is closed') and q is true in a committed state ('I want to be outside'), then r is necessarily true in some intermediate state ('I have to open the door').[2]

An important question for a dialogue generation model is *why* information flows in the first place. In other words, what is a participant's basic motivation to perform a communicative action? Psychological oriented theories about motivation, such as Maslow's hierarchy of needs [26], are based on assumptions of internal representations and processes such as beliefs, goal setting, expectancies and desires. We will avoid motivational concepts such as hunger, fear and sexuality, however, and borrow the concept 'homeostasis' from system theory, i.e. the process by which a system maintains a balanced state. We will assume that dialogue behaviour can be modelled as an abstract process of balancing an agent's internal *belief* and *commitment* state, bring-

[2]The commitment inference can be formalized in terms of possible-world semantics as follows. Let $M = (W, \pi, R)$ be a model that consists of a set W of states with typical elements u, v and w, a valuation function π which assigns a truth value $\pi_w(p)$ to every proposition letter p in each state w, and R is a reflexive, transitive and anti-symmetric accessibility relation. The semantics of the arrow in '$p \times q \rightarrow r$' is defined as follows. Suppose $u, w \in W$ then:

$$M, u, w \models p \times q \rightarrow r \Leftrightarrow \text{ if } \pi_u(p) = \mathit{true} \text{ and } \pi_w(q) = \mathit{true}$$
$$\text{then on every } R\text{-path from } u \text{ to } w : \exists v \text{ with } \pi_v(r) = \mathit{true}$$

In short: r is *instrumentally necessary* for q if p. In the pragmatic model, the expression '$p \times q \rightarrow r$' is used as follows: $B_x p \wedge C_x q \Rightarrow Add(C_x r)$, meaning that if x beliefs that p is true and x is committed to q, then x is also committed to r.

ing about that the agent believes everything the agent is committed to. Below we will distinguish various types of belief and commitment, but the basic motivation for action always comes from the imbalance between the two types of states. As the needs in Maslow's pyramid, commitments may have different priorities which follow from both the order and the type of commitments; these priorities determine the basic structure of the conversation.

4 Mental constructs and a balanced state

We will assume that the agent's cognitive state consists of a number of mental constructs and that each construct contains particular pieces of information with respect to the domain of discourse. Precisely which constructs have to be included depends on the phenomenon one wants to explain or, in a generation framework, on the rules that are needed to generate particular dialogue contributions. We motivated the use of two basic types of mental constructs in the previous section – beliefs and commitments – and assumed that the basic motivation for action is the imbalance between the belief and the commitment state of the participants. In other words, the imbalance will be considered as the driving force behind the dialogue generation process. We will now explain which mental constructs are included in our generation framework and how the generation of communicative acts can be embedded in the homeostatic process based on these mental constructs.

4.1 The agent's mental constructs

An agent x's cognitive state consists of various types of beliefs and commitments:

Beliefs

- Private belief of an agent x about the domain of discourse (B_x)
- Mutual beliefs about the domain (MB)
- Mutual beliefs about the commitments of the partner y (MBC_y)
- Information of which the partner y is ignorant (I_y) In particular, we discern two types: ignorance with respect to the beliefs of the partner y (IB_y) and with respect to the commitments of y (IC_y)

where mutual beliefs about the domain are considered as a subset of private beliefs.

Commitments

- Private commitments of x with respect to a particular state of the domain of discourse (C_x)
- Social commitments of x (S_x). In particular, we discern two types: with respect to the beliefs of the partner y (S_xB_y) and with respect to the y's commitments (S_xC_y)

Social commitments with respect to beliefs of the other are used to indicate that the partner has asked a question with respect to a particular piece of information about the discourse domain; social commitments with respect to the commitments of the other are used to express that the other has asked a question about his own commitments. In the latter case we may think of questions such as 'Should I take an umbrella to go to the supermarket' or 'Will I have pain when I go to the dentist?'. Intuitively, commitments fulfill two roles in this paper: First, they indicate a particular desire, i.e. a particular goal state the agent wants to be in. Second, sub-commitments indicate a particular unavoidable state, i.e. a state the agent must be in before the goal state can be achieved. Whether the intermediary state is desired or not, is of no importance in this paper.

4.2 Achieving a balanced situation

To avoid unnecessary complexity, we will make two important simplifications. First, during the dialogue, the participants have no access to a domain of discourse, i.e. they are unable to observe or manipulate particular aspects of the domain. In other words, information only flows between the two dialogue partners like, for instance, in a telephone dialogue. A second simplification is that the participants only hold positive information about the domain of discourse, i.e. negation is excluded. This implies that the two participants will never hold conflicting beliefs or commitments, and, since alleged inconsistencies will never arise, the agents will never argue about a specific statement. Information may be incorrect with respect to a particular instance of the domain of discourse, but the incorrectness will never be discovered, since the agents have no access to the domain.

A balanced situation can be achieved by updating the mental constructs of the agents. Let us therefore first define a balanced state of an agent.

An agent x has a *balanced state* iff

1. x has no commitments or
2. (a) all private commitments x are privately believed by x and
   (b) all social commitments of x about y's beliefs are mutually believed and
   (c) all social commitments of x about y's commitments are mutually believed to be y's commitments

In order to achieve a balanced situation, agents may either modify their belief state or their commitment state. In line with the communication model of Section 3, private beliefs can in principle be modified in three ways: a. by belief inference to make implicit beliefs explicit, b. by an appropriate communicative act from the dialogue partner, and c. by direct perception of the domain of discourse. In the second case, agents 'take over' the belief of the dialogue partner. Below, the third case will not be considered, since we assumed that the agents have no access to the domain of discourse. In the presented dialogue game, explicit beliefs can thus only be modified in two ways: via a reasoning mechanism for the belief states of the agents and by communication with the partner. Since we deal with an ideal communication channel and since we did not include negation, all manifested beliefs will be included in the

agents' mutual beliefs about the domain. In practice, two versions of mutual belief exist, x's version and y's version, but since we will assume that both versions contain the same information, we will speak of one version only.

An initial commitment state can in principle be adjusted in four ways: a. by commitment inference to make implicit private commitments explicit, b. if the agent receives particular information with respect to its private commitment state, c. by a question of the partner (inducing a social commitment) and d. if the agent concludes that he is unable to change the belief state in such a way that the situation can be balanced (c.f. [11]). In the fourth case the commitment will be cancelled.

5 The Dialogue Game

The dialogue game is divided into two parts (for a similar approach, see [27] or [3]): a. the game-board that contains information about the cognitive states and the communicative acts, and b. the dialogue rules that control the behaviour of the participants (generation rules) and that prescribe how the game-board changes (update rules). The game-board represents the participants' cognitive state and typically changes because of the participants' communicative actions.

'Moves' or information flows between the two participants are composed of two elements: a. plain information about the domain of discourse (the semantic content), and b. information about the way the different mental constructs should be updated (the communicative function). Every move is completely determined by the cognitive state of the participant who has the turn to act and by the rules for co-operative behaviour that will be presented below. Since the cognitive states are updated after every move, the next move is not only determined by the previous one, as would be the case in a dialogue grammar, but also by the context of the move. Each play is a sequence – not necessarily finite – of linearly or hierarchically alternating moves.

In the dialogue game and in line with the Gricean maxims [15], agents do not put forward information they do not believe or information they mutually believe. The distinction between private and mutual beliefs enables us to give concrete form to the maxim of quantity. If relevant, private beliefs can always be manifested, unless they are part of the mutual beliefs. Mutual beliefs give us a criterion to leave out particular information in the dialogue contribution (otherwise we would manifest information that the user already believes). In the dialogue game, agents also do not ask information they believe or they believe the other is ignorant about.

5.1 The agents' cognitive state

All constructs of an agent's cognitive state are modelled as sets of information items, except for the private and social commitments which are both modelled as a list. The two types of social commitments together form one list.

We adopt the following shorthand notations: we write $B_x p$ as a shorthand for $p \in B_x$ and write $\neg B_x p$ for $p \notin B_x$, and similarly for the other mental constructs.

Note that in writing expressions like $B_x p$ ('x believes that p'), B_x is not to be confused with a modal operator with corresponding semantics in terms of accessibility relations and possible worlds. Our focus is here on dialogue generation, so B_x is

instead modelled as a dynamic state with an *operational* semantics given by the rules of the dialogue. In particular, the rules make use of the fact that particular information p is present ($B_x p$), is absent ($\neg B_x p$), is added ($Add(B_x p)$) or is deleted ($Del(B_x p)$).

We use the notation $S_x\underline{p}$ or $C_x\underline{p}$ to denote that p is the commitment under discussion. A commitment list may be empty, indicated by $S_x\emptyset$ and $C_x\emptyset$.

5.1.1 Inference rules

We assume that the agents can reason about their beliefs and commitments by the following inference rules:

I1	$B_x p \wedge B_x(p \rightarrow q) \Rightarrow Add(B_x q)$
I2(a)	$B_x p \wedge C_x r \wedge MB(p \times r \rightarrow q) \Rightarrow Add(C_x q)$
I2(b)	$MBp \wedge MBC_x r \wedge MB(p \times r \rightarrow q) \Rightarrow Add(MBC_x q)$

So, if the belief base of agent x contains the information that p and that $p \rightarrow q$ then according to rule I1 it is updated with the information that q. Additionally, if x believes p, is committed to r and $p \times r \rightarrow q$ is part of the mutual beliefs then according to rule I2(a) the private commitment base is updated with q. Rule I2(b) is similar. To keep things simple we will in the dialogue model assume that all commitment inferences are part of the mutual beliefs.

The belief states are monotonic, i.e. everything that can be inferred from previous states, can also be inferred from new belief states. Information about commitments can be retracted after particular communicative acts, for instance, if an agent receives the answer 'Don't know' to his question 'whether p' then the private commitment p is dropped. We are not concerned with the full details of the update mechanism, but assume that the cognitive states are updated in line with the principles I1 and I2 and the update rules presented below.

In the dialogue model, we assume that after each round an agent's beliefs and commitments are closed by (successive) application of the inference rules I1 and I2.

Finally, in the dialogue model, we will use a function $link$ that gives us the set of all antecedents that are connected to a particular consequence of a belief inference. More precisely, link is defined in the following way:

$$link(x, q) \equiv \{p \mid B_x(p \rightarrow q)\}$$

For instance, if x believes that '$p \rightarrow q$' and believes that '$r \rightarrow q$', then $link(x, q) = \{p, r\}$. If there is no compound proposition with q as its consequence in belief state x, the set is empty ($\emptyset$).

5.2 Communicative acts

Agents manifest their beliefs and commitments by means of communicative acts or moves, such as statements and questions. The content of a move consists of a formula: Cp ('p is a commitment'), Bp ('p is believed') or $Cp \leftarrow Bq$ ('p is a commitment since q is believed'); the communicative function is tagged by one of the following markers ('?', '!', '$*$' and '♣'):

- Questions: $[Bp]^?$ and $[Cp]^?$

- Statements:$[Bp]^!$, $[Cp]^!$ and $[Cp \hookleftarrow Bq]^!$
- Ignorance: $[Bp]^*$ and $[Cp]^*$
- Closure of the dialogue: $[\,]^{\clubsuit}$

We use the notation '$x : m$' to denote that agent x is the performer of the move m.

5.3 Generation rules

The general dialogue mechanism is as follows.

1. An agent's first priority is to resolve any imbalance with respect to its social commitments. For instance, if it receives the question 'whether p' and p is part of private beliefs then it responds that p holds.
2. If there are no such imbalances, then the second priority is to ask questions that allow the derivation of sub-commitments of its commitments. For instance, if the agent believes $p \times r \rightarrow q$, is committed to p and not (yet) to the sub-commitment q then it asks whether p holds.
3. Otherwise the agent aims at resolving any imbalance between its private believes and commitments. For instance, if the agent is committed to r but does not (yet) know that p then it asks whether p holds.
4. If there are no imbalances, the agent closes the dialogue.

With respect to the generation of sub-commitments in step 3. we discern two options: we can either ask for the belief (p) or for the sub-commitment (q). For instance, if we are committed to prepare dinner we prefer asking for the belief 'Are our guests vegetarian?' to asking for the (many) sub-commitments it would yield ('Do I need to buy eggs?', 'Do I need to go to the greengrocer's?', 'Do I need to soak beans?' and so on). Conversely, if we are committed to drive home we prefer to ask for the sub-commitment 'Do I need to take the road E12?' to asking for of the (many) beliefs from which this particular commitment could be derived ('Is there a traffic jam on the E27?', 'Is the E38 still road-blocked?', 'Is the Prins Claus Bridge closed?' and so on). Hence, in the dialogue model we adopt the following (simple) strategy: given the rule $p \times r \rightarrow q$ an agent asks for the belief p (rather than for the sub-commitment q) if there is at least an additional rule $p \times r \rightarrow q'$. The agent asks for the the sub-commitment q (rather than for the belief p) if there is at least an additional rule $p' \times r \rightarrow q$. If neither or both of these conditions hold then the agent chooses randomly.[3]

Speech acts (or moves) are fully determined by the cognitive state of the participant who performs the move and by the rules that are applicable to this state. The double arrow '$\Rightarrow$' links the preconditions of the move to the move itself. The left side of the arrow is of type proposition and represents the preconditions in terms of the cognitive state of an agent; the right side is of type action and represents the generated move.

[3]For the purposes of this paper, we consider this simple preference ordering. More involved orderings could for instance take the number and / or priorities of beliefs and sub-commitments into account.

5.3.1 Questions

Since the agents have no access to the domain of discourse, the initial move can only be a question. There are two types of questions. The first is a question whether some proposition is believed to hold:

$$G0. \quad C_x p \wedge \neg B_x p \wedge S_y \emptyset \Rightarrow x : [Bp]^?$$

The first two preconditions of G0 indicate that the agent's state is out of balance with respect to the commitment p, the third condition indicates that there are no social commitments to be handled first. We assume that the imbalance with respect to p is to be resolved by communication (and not for instance by means of actions and/or observations in the world).

The second reason for asking whether some proposition is believed to hold is that it would yield the generation of a new private commitment (according to inference rule I2).

We first introduce the following short-hand notation Φ:

$$\Phi(x, p, r, q) \equiv MB(p \times r \rightarrow q) \wedge C_x r \wedge \neg C_x q$$

which expresses that the agent believes $p \times r \rightarrow q$, has r as it private commitment and does not have q as a private commitment (yet). The dialogue rules are then as follows:

$$\begin{array}{ll} G1. & \Phi(x, p, r, q) \wedge \neg IB_y p \wedge S_y \emptyset \Rightarrow x : [Bp]^? \\ G2. & \Phi(x, p, r, q) \wedge \neg IC_y q \wedge S_y \emptyset \Rightarrow x : [Cq]^? \end{array}$$

The first precondition indicates that a new sub-commitment q of r could be developed from the belief that p. The question whether this proposition p holds can be asked if the agent does not believe that the other is ignorant. Similar to rule G0 it is also required that there are no social commitments to be handled first.

In order to decide between the application of rule G1 and G2 we define the following strategy. Given x and r,

- prefer G1 if there exist at least two distinct instances of q
- prefer G2 if there exist at least two distinct instances of p
- if both or none are preferred then choose randomly

So for instance, if $\Phi(x, p, r, q)$ and $\Phi(x, p, r, q')$ hold then ask for the belief p, if $\Phi(x, p, r, q)$ and $\Phi(x, p', r, q)$ hold then ask for the commitment q and if all of these conditions hold then either ask for the belief p or for the commitment q.

5.3.2 Responses to questions

After the initiator has asked a question, his question becomes manifest as a social commitment of the follower. There are three possibilities for the next move:

a. The follower knows the answer and thus gives the answer. If applicable she even makes a more cooperative move by including a relevant explanation.

b. The follower does not know the answer directly, but concludes that there may be a way to find the answer and asks a counter-question.

c. The follower is ignorant and does not have a solution. She manifests her ignorance.

With respect to case a. an explanation is relevant if it allows for the generation of additional sub-commitments. For instance, given the question 'Do I need to come home?', adding an explanation ('because your dog is sick') to the answer 'Yes' is relevant if it would allow the other to infer one or more additional sub-commitments (such as 'I need to hurry.').

Generation rule G3 expresses that if x has the social commitment with respect to the belief q of y and q is believed by x, then x will answer that q is believed to hold:

$$G3. \quad S_x B_y \underline{q} \wedge B_x q \Rightarrow x : [Bq]^!$$

Additionally, if x has the social commitment with respect to the commitment q of y and the commitment q can be inferred from the belief that r, then if r is relevant then x will answer that q is a commitment because r is believed (G4), otherwise x will just answer that q is a commitment (G5).

$$G4. \quad S_x C_y \underline{q} \wedge \Psi(x, p, r, q) \wedge B_x p \wedge \mathit{relevant}(x, p, r, q) \Rightarrow x : [Cq \hookleftarrow Bp]^!$$

$$G5. \quad S_x C_y \underline{q} \wedge \Psi(x, p, r, q) \wedge B_x p \wedge \neg \mathit{relevant}(x, p, r, q) \Rightarrow x : [Cq]^!$$

where we use short-hand notation:

$$\Psi(x, p, r, q) \equiv MB(p \times r \rightarrow q) \wedge MBC_y r$$

to express that the agent believes $p \times r \rightarrow q$ and it is mutually believed that r is a private commitment of the other.

Additionally, the notion of relevance is formalised as follows:[4]

$$\mathit{relevant}(x, p, r, q) \equiv \exists q' \neq q \wedge \Psi(x, p, r, q') \wedge \neg MBC_y q'$$

which expresses that p (e.g, 'dog is sick') is relevant if it would allow the generation of an additional commitment q' ('need to hurry') in addition to q ('need to come home').

Additionally, if x does not know the answer, x may ask a counter-question. The counter-question can only be asked if the agent finds the antecedent of a linked proposition, and if he or she does not believe that the other is ignorant with respect to the linked proposition (G6 and G7):

$$G6. \quad S_x B_y \underline{q} \wedge \neg B_x q \wedge p \in \mathit{link}(x, q) \wedge \neg IB_y p \Rightarrow x : [Bp]^?$$

$$G7. \quad S_x C_y \underline{q} \wedge \Psi(x, s, r, q) \wedge \neg B_x s \wedge p \in \mathit{link}(x, s) \wedge \neg IB_y p \Rightarrow x : [Bp]^?$$

For reasons of legibility we have omitted from rule G7 the preconditions implying that neither rule G4 nor G5 is applicable.

[4]For the purposes of this paper we consider a simple notion of relevance.

If x does not know the answer and cannot ask a counter-question, x will manifest his or her ignorance (G8 and G9).

$$G8. \quad S_x B_y \underline{q} \wedge \neg B_x q \wedge \neg(\exists p : p \in link(x, q) \wedge \neg IB_y p) \Rightarrow x : [Bq]^*$$

$$G9. \quad S_x C_y \underline{q} \ \wedge not\_applicable(G7) \Rightarrow x : [Cq]^*$$

where for reasons of legibility we use the notation $not\_applicable(G7)$ to express that the preconditions of G7 do not hold.

Finally, in G10 a closing act is generated if the social commitment list is empty and if the situation is in balance:

$$G10. \quad S_x\emptyset \wedge \neg\exists q : (C_x q \wedge \neg B_x q) \Rightarrow x : [\,]^{\clubsuit}$$

To avoid an infinite sequence of closing acts, a meta-rule has been defined to close the dialogue:

Closing (CL)

Both dialogue partners stop generating communicative acts iff two successive closing acts are performed (i.e. the sequence $x : [\,]^{\clubsuit}$ and $y : [\,]^{\clubsuit}$).

5.4 The update of cognitive states

The update function yields a new cognitive state depending on the old state and the move just performed. To represent the consequences of a particular move, we use the notation '$\Rightarrow$'. The left side is of type action and represents the performed move; the right side represents the postconditions and denotes how the cognitive states should be updated. The relevant attitudes are preceded by Del or Add depending on whether information is to be added or removed. So, for instance, $Del(C_x\, q)$ means that q should be deleted from the private commitments list.

In update rule U1, it is expressed that if x utters a statement that q holds then q is added to the mutual beliefs.

$$U1. \quad x : [Bq]^{!} \Rightarrow Add(MBq)$$

Update rule U2 indicates that if x utters that q is a commitment then this becomes a mutual belief.

$$U2. \quad x : [Cq]^{!} \Rightarrow Add(MBC_y\, q)$$

Update rule U3 is similar to U2. In addition the explanation p is added to the mutual beliefs.

$$U3. \quad x : [Cq \Rightarrow Bp]^{!} \Rightarrow Add(MBp) \wedge Add(MBC_y\, q)$$

Rules U4 express that if x utters a question whether q holds, then y obtains the social commitment with respect to the belief q of x.

$$U4. \quad x : [Bq]^{?} \Rightarrow Add(S_y B_x\, q)$$

If x utters a question whether q is a commitment, then y obtains the social commitment with respect to the commitment q of x (U5):

$$U5. \quad x : [Cq]^{?} \Rightarrow Add(S_y C_x q)$$

Rule U6 and U7 express that if x indicates that he or she has no information about (the belief or commitment) q, q will be added to the beliefs of y about the ignorance of x and removed from the social commitments and, if present, from the private commitments of y:

$$U6. \quad x : [Bq]^{*} \Rightarrow Add(IB_x q) \wedge Del(S_x B_y q) \wedge Del(C_y q)$$

$$U7. \quad x : [Cq]^{*} \Rightarrow Add(IC_x q) \wedge Del(S_x C_y q)$$

The last rule, U8, expresses that cognitive states do not change after a closing act:

$$U8. \quad x : [\,]^{\clubsuit} \Rightarrow \otimes$$

Finally, we assume the following two belief state maintenance rules:

$$M1. \quad Add(MBp) \Rightarrow Add(B_x p) \wedge Add(B_y p) \wedge Del(IB_y p) \wedge Del(IB_x p)$$
$$M2. \quad Add(MBC_x p) \Rightarrow Add(C_x p) \wedge Del(IC_x p)$$

Rule M1 indicates that if p becomes mutual knowledge it is added to the private beliefs and if present removed from the beliefs about the other agent's ignorance. According to rule M2, if p is added to the mutual beliefs about the commitments of agent x it is added to the commitments of x as well and deleted from the ignorance of y about x's commitments.

6 A dialogue example

We turn now to an example where John and Mary play the co-operative dialogue game based on the previously introduced mental constructs, and the inference, generation, update and maintenance rules. First, we present an abstract version of the example, and after that, we will the example into a 'natural' language dialogue.

In Figure 2, we have depicted the dialogue transition table, i.e. the cognitive states of Mary and John, the communicative acts (MOVE) and, in addition, a reference to the applied update and generation rules. Empty states are indicated by '$\emptyset$'. In the example, we have left out Mary's commitment state (C_M), Mary's belief state about John's ignorance (I_J), the mutual beliefs about Mary's commitments (MBC_M) and John's social commitments with respect to the commitments of Mary ($S_J C_M$), since in this particular situation these states remain empty during the course of the dialogue. In the initial situation, the beliefs are as follows (because of space limitation we have not depicted them in the table). John and Mary mutually believe that:

1. if one is late and wants to go to the supermarket then one needs to hurry ('$l \times s \rightarrow h$')
2. if one is late and wants to go to the supermarket then one needs to go by bike ('$l \times s \rightarrow b$')

	Mary						John			
Nr.	B	SB_J	SC_J	MBC_J	MB	$MOVE$	C	B	SB_M	IC_M
G1 1.	$l, r, w \rightarrow uu$	$\emptyset$	$\emptyset$	s	$\emptyset$	$J : [Bl]?$	s	w	$\emptyset$	$\emptyset$
U4 G3 2.		$\underline{l}$				$M : [Bl]!$				
U1 M1 I2 G2 3.				h, b	l	$J : [Cd]?$	h, b	l		
U5 G2 4.			$\underline{d}$			$M : [Cd]^*$				
U7 G2 5.			$\emptyset$			$J : [Cj]?$				d
U5 G4 6.			$\underline{j}$			$M : [Cj \Leftarrow Br]!$				
U3 M1 I2 M2 G2 7.				j u	r	$J : [Ca]?$	u j	r		
U5 G7 8.			$\underline{y}$			$M : [Bw]?$				
U4 G3 9.						$J : [Bw]!$			$\underline{f}$	
U1 M1 I1 G5 10.	w uu				w	$M : [Ca]!$				
U2 M2 G10 11.				a		$J : [\,]^{\clubsuit}$	a			
U8 G10 12.						$M : [\,]^{\clubsuit}$				
U8 CL										

Figure 2: Dialogue between John and Mary about John's commitment to go to the supermarket. On the basis of Mary's responses various sub-commitments (e.g., going by bike, taking an umbrella) are generated.

3. if there is a road-block and one wants to go to by bike then one needs to take the detour ('$rb \times b \rightarrow d$')

4. if the bridge is closed and one wants to go to go by bike then one needs to take the detour ('$bc \times b \rightarrow d$')

5. if it is raining and one wants to go to by bike then one needs to take a jacket ('$r \times b \rightarrow j$')

6. if it is raining and one wants to go to by bike then one needs to take an umbrella ('$r \times b \rightarrow u$')

7. if it is cold and one wants to go to by bike then one needs to take a jacket ('$c \times b \rightarrow j$')

8. if all umbrella's are currently in use and one wants to take an umbrella one needs to fetch one from the attic ('$uu \times u \rightarrow a$')

9. if all umbrella's are broken and one wants to take an umbrella one needs to fetch one from the attic ('$ub \times u \rightarrow a$')

Mary believes that John is late (l), that it is raining (r) and that if William is at school (and thus has taken his umbrella) then all umbrella's are currently in use ($w \rightarrow uu$). Mary has no commitments. John believes that William is at school (w). It is mutually believed that John is committed to go to the supermarket. John is the initiator and starts with the initial question whether l (move 1). Mary is able to answer this question directly (move 2). From the answer it is concluded that John has to hurry (h) and has to go by bike (b). Based on his new commitment to go by bike John asks whether he has to take the detour (d) (move 3). Mary informs that she has no information about this (move 4). John asks whether he needs to take a jacket (j) (move 5). Mary answers that he indeed needs one because it is raining (move 6), from which John also infers that he should take an umbrella (u). On the basis of this new commitment, John asks whether he needs to fetch one from the attic (a) (move 7). Mary is unable to answer this question directly, but may find an answer if she has the information that William is at school (w). So, she asks for this information (move 8). John answers that this is indeed the case (move 9), from which Mary infers that John should fetch an umbrella from the attic. She is now able to answer his question (move 10). Finally, since the remaining imbalance between beliefs and commitments cannot be solved by communication (but instead have to be resolved by other means such as actions and observations), the dialogue is closed (move 11 and move 12).

Below the corresponding dialogue is presented in 'natural language':

0. *John: I want to go to the supermarket.*
   (not manifested in the transition table)
1. *John: Am I late?*
2. *Mary: Yes.*
3. *John: Do I need to take the detour then?*
4. *Mary: That I don't know.*
5. *John: And do I need to take my jacket?*
6. *Mary: Yes, because it is raining.*
7. *John: Do I need to fetch an umbrella from the attic then?*
8. *Mary: Is William at school?*
9. *John: Yes.*
10. *Mary: Well in that case you need to fetch one from the attic.*
11. *John: OK, thank you.*
12. *Mary: OK, no thanks.*

7 Discussion and future research

In this paper we presented a framework for the generation of coherent elementary conversational sequences at the speech act level. We were able to show by means of an explicit presentation of the transition tables that the structure and the coherence of conversational units are the result of an interplay between the dialogue and update rules and the initial cognitive states of the dialogue partners. Figure 2 carefully shows the pre- and post-conditions of every speech act and the change of the mental constructs during the dialogue as a basis for utterance production. Clearly, the dialogue is still unnatural and lacks many of the ingredients that we usually observe when we study the properties of natural language dialogue. Taking a more profound look, however, we notice the elementary structural phenomena discussed at the beginning of this paper. In Figure 2 we can observe the linear organisation of adjacency pairs, such as question-response (moves 1-2, and moves 3-4) and the closing of the dialogue (moves 11-12). We also observe the hierarchical organisation of insertion sequences, such as moves 8 and 9 between the question in 7 and its reply in 10. Depending on the initial states, the dialogue rules generate an arbitrary number of levels of sub-sequences and the final reply may be originated many turns away from the initial question (see also [7]). The dialogue coherence, although admittedly oversimplified, comes from both the background knowledge of the dialogue partners and the way the commitments states are processed.

In our approach we extended the work presented in [7] in several ways in order to generate richer dialogue structures. In contrast to [7], we did not take a rigid non-planning approach, but included a simple possibility to reason about commitments. In [7] a planning approach was rejected because of complexity problems and the presented communication model incorporated an extreme sensitivity to the local circumstances of the conversation. In this paper we still prefer a local solution, but admit that talking and reasoning about commitments (or goals) is a substantial aspect of conversation and therefore has to be included to generate more interesting speech act sequences. We therefore included a possibility to reason about the commitments of the participants in a very simple manner.

Another extension of the model described in [7] is the inclusion of the participants' mutual belief. Implicitly, and in line with other researchers (e.g. [10, 2, 37]), we assumed that successful communication also requires some degree of alignment or common ground and that the goal of grounding of information is a vital activity in cooperative communication. Prior to a conversation, participants not only have beliefs of a particular discourse domain, they also assume that there is some agreement about these beliefs and they augment manifested beliefs to the 'agreed beliefs' during the conversation. In practice it is hard (and, since dialogue participants have no direct access to their partner, even impossible) to decide whether mutual beliefs are really common, but our main point is that dialogue participants act as if these beliefs are common. In fact, the introduction of mutual beliefs enabled us to give concrete form to the Gricean maxim of quantity, since mutual beliefs give a criterion to leave out particular information in the dialogue move.

The approach presented in this paper is still rudimentary and extensions could be developed along different lines, such as an extension of the domain and communication language, the addition of roles played by the agents and the investigation of different communicative situations. It should be noted, however, that some extensions may have far-reaching consequences for different aspects of the game. For instance, including negation in the domain language seems another inevitable step towards a generalisation of the framework, since it introduces the possibility of modelling conflicting beliefs and the generation of argumentative dialogues. Nevertheless, a negated proposition cannot simply be added to a belief state of one of the participants, since it may result in unwanted inconsistencies. A solution is the introduction of a temporary state that represents the beliefs of an agent about the beliefs of his or her partner - 'A believes that B believes' – so that the different types of information can be carefully separated and inconsistencies can be avoided. It is unclear, however, how these conflicts will ever be resolved in the present game without other extensions, since both participants are considered 'equal' and there is no reason why they would prefer one proposition over the other. In other words, an agent can never accept a conflicting proposition stated by the other if the model does not contain a notion of 'expertise' or 'power'.

Another line of research would be a careful analysis of what actually happens in human dialogues. In order to determine what humans do in realistic conversational circumstances and to validate the model presented in this paper, the acquisition of empirical data is a necessary step in the research process. We will, therefore, collect empirical data from various conversations and hope that the analysis of the transcripts will lead to a further extension and refinement of the model, such as the inclusion of richer semantic descriptions of the domain information.

Some important simplifications were made in the game with respect to the underlying communication model. Utterances were always accepted and did not have to be checked for inconsistencies or other counter-evidence. Participants could never be misinformed and could not have weak evidence for a specific fact. In natural situations, however, where people have multimodal access to various aspects of the world, information channels may be disturbed and an agent's attention may be attracted by a variety of sources, including pointing acts of the dialogue partner. The triangle model described in Section 3 includes some of the necessary basic ingredients to describe these phenomena, but there are a number of important questions left. For instance, when do agents decide to observe the domain rather than infer the information from

their own belief state or ask a question to their partner? How does an agent's cognitive representation depend on whether information is observed or communicated and how do these representations influence the course of the dialogue? Partly, the answers depend on fundamental psychological issues, such as perceptual abilities, memory capacity and attention capabilities. In this paper, however, we abstracted from these matters, since including them dramatically increases the complexity of the model without supplying a substantial contribution to the explanation of the dialogue structure.

Acknowledgements

The authors would like to thank Lidwien van de Wijngaert for the statistical analysis of the experimental data. We also would like to thank the three anonymous jolli reviewers for their helpful comments on previous drafts of this paper.

References

[1] R.M.C. Ahn, R.J. Beun, T. Borghuis, H.C. Bunt, and C.W.A.M. van Overveld. The denk-architecture: A fundamental approach to user-interfaces. *Artificial Intelligence Review*, 8(9):431–445, 1995.

[2] J. Allwood, J. Nivre, and E. Ahlsen. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26, 1992.

[3] L. Amgoud, N. Maudet, and S. Parsons. Modelling dialogues using argumentation. In *Proceedings of the Fourth International Conference on MultiAgent Systems (ICMAS 2000)*, pages 31–38, Boston (MA), 2000.

[4] N. Asher and A. Lascarides. Questions in dialogue. *Linguistics and Philosophy*, 23(2):237–309, 1998.

[5] J.L. Austin. *How to do Things with Words*. Clarendon Press, Oxford, 1962.

[6] J.A. Bateman and K.J. Rondhuis. Coherence relations: Towards a general specification. *Discourse Processes*, 24:3–49, 1997.

[7] R.J. Beun. On the generation of coherent dialogue: A computational approach. *Pragmatics and Cognition*, 9(1):37–68, 2001.

[8] H.C. Bunt. Information dialogues as communicative action in relation to partner modelling and information processing. In Taylor et al. [36], pages 47–73.

[9] L. Carlson. *Dialogue Games. An Approach to Discourse Analysis*. D. Reidel Publishing Company, Dordrecht, 1985.

[10] H.H. Clark and C.R. Marshall. Definite reference and mutual knowledge. In A.K. Joshi, B.L. Webber, and I.A. Sag, editors, *Elements of Discourse Understanding*, pages 10–63. Cambridge University Press, Cambridge, 1981.

[11] P.R. Cohen and H.J. Levesque. Persistence, intention and commitment. In P.R. Cohen, J. Morgan, and M.E. Pollack, editors, *Intentions and Communication*, pages 33–69. MIT Press, Cambridge, Mass., 1990.

[12] A Gatt and K. van Deemter. Conceptual coherence in the generation of referring expressions. In *Proceedings of the Workshop on Coherence for Generation and Dialogue (ESSLLI 2006)*, pages 17–24, Malaga, 2006.

[13] M.A. Gernsbacher and T. Givón. *Coherence in Spontaneous Text*. John Benjamins Publishing Company, Amsterdam, 1995.

[14] T. Givón. Coherence in text vs. coherence in mind. In M.a. Gernsbacher and T. Givón, editors, *Coherence in Spontaneous Text*, pages 59–115. John Benjamins Publishing Company, Amsterdam, 1995.

[15] H.P. Grice. Logic and conversation. In P. Cole and J.L. Morgan, editors, *Speech Acts. Syntax and Semantics, Vol. 11*, pages 41–58. Academic Press, New York, 1975.

[16] B.J. Grosz and C.L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[17] C.L. Hamblin. *Falacies*. Methuen, London, 1970.

[18] J. Hobbs. Coherence and coreference. *Cognitive Science*, 3(1):67–90, 1979.

[19] J. Hulstijn. *Dialogue Models for Inquiry and Transaction*. PhD thesis, University of Twente, 2000.

[20] J. Hulstijn, F. Dignum, and M. Dastani. Coherence constraints for agent interaction. In R.M. van Eijk, M.-P. Huget, and F. Dignum, editors, *Agent Communication*, volume 3396 of *LNAI*, pages 134–152. Springer-Verlag, 2005.

[21] E. Hutchins. Metaphors for interface design. In Taylor et al. [36], pages 11–28.

[22] R. Kibble and R. Power. Optimizing referential coherence in text generation. *Computational Linguistics*, 30(4):401–416, 2004.

[23] S. Larsson and D.R. Traum. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, 6(3-4):323–340, 2000.

[24] S.C. Levinson. *Pragmatics*. Cambridge University Press, Cambridge, 1983.

[25] R.E. Longacre. *The Grammar of Discourse*. Plenum Press, New York, 1996.

[26] A.H. Maslow. *Motivation and Personality*. Harper and Row, New York, 1970.

[27] P. Piwek. *Logic, Information and Conversation*. PhD thesis, Eindhoven University of Technology, 1998.

[28] P. Piwek. Meaning and dialogue coherence. In *Proceedings of the Workshop on Coherence for Generation and Dialogue (ESSLLI 2006)*, pages 57–64, Malaga, 2006.

[29] L. Polanyi. A formal model of the structure of discourse. *Journal of Pragmatics*, 12:601–638, 1988.

[30] R. Power. The organisation of purposeful dialogues. *Linguistics*, 17:107–152, 1979.

[31] G. Redeker. Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics*, 14:367–381, 1990.

[32] G. Rickheit and H. Strohner. Towards a cognitive theory of linguistic coherence. *Theoretical Linguistics*, 18(2/3):209–237, 1992.

[33] T.J.M. Sanders and L.G.M. Noordman. The role of coherence relations and their linguistic markers in text processing. *Discourse Processes*, 29(1):37–60, 2000.

[34] T.J.M. Sanders, W. Spooren, and L.G.M. Noordman. Toward a taxonomy of coherence relations. *Discourse Processes*, 15(1):1–35, 1992.

[35] M.P. Singh. An ontology for commitments in multi-agent systems: toward a unification of normative concepts. *Artificial Intelligence and Law*, 7:97–113, 1999.

[36] M.M. Taylor, F. Néel, and D.G. Bouwhuis, editors. *The structure of multimodal dialogue*, Amsterdam, 1989. Elsevier Science Publishers.

[37] D. R. Traum. *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, University of Rochester, 1994. Technical Report 545.

AGENS SAPIENS

JOHN-JULES MEYER

Sapiens fortissime agit sed digne.

In this short article I will sketch the field of agent technology and in particular the work we have done in Utrecht in this area. The field emerged in the 90s as an interdisciplinary field in between artificial intelligence and main stream computer science, particularly the areas of distributed computing and software engineering. It also has many ties to the disciplines of philosophy, logic, cognitive science, social science and economy.

In the present paper I will mainly restrict myself to our own views on the field and sketch some of the topics that we have studied. It is not meant to be complete, but just to give an inkling of what this work is about, and especially what our role has been. It is also rather personal and subjective, viewed from my perspective, and it only treats matters I am/was involved in and which are/were really on my mind. My apologies for all people and work not mentioned. Moreover, I've tried to be as non-technical as possible. More objective and comprehensive overviews of the field of agent technology can be found in e.g. [51, 29].

1. Intelligent Agents

As stated above the field of agents emerged in the 90s. Based on philosophical ideas, from Aristotle to Dennett and Bratman, the concept of an agent as an autonomous entity, making its own decisions to perform actions, took shape. Admittedly, the philosophers mentioned primarily thought of human beings as decision makers, but it was realised soon by researchers in the field of Artificial Intelligence such as Pollack and Israel that these ideas could also be fruitfully employed to construct artificial agents in the sense of software entities that make decisions to perform actions in their environment, (more or less) independently from their user(s).

Properties of agents that were deemed to be crucial are reactiveness (does the agent respond adequately and timely to stimuli from the environment, including its user or other agents?), proactiveness (does the agent take possible future developments into consideration, does it work goal-directedly, pursuing and setting (sub)goals?), and sociality (does the agent coordinate its behaviour, either cooperatively or competitively, with other agents sharing the same environment?). In the 90s researchers started to propose both formal models and concrete architectures for this kind of systems, usually called autonomous / intelligent / rational agents, and multi-agent systems / agent societies for systems involving multiple agents.

Intelligent Systems Group, Universiteit Utrecht.

2. BDI Agents: Logics and Architectures

In order to achieve the desiderata for agent systems as mentioned above, many researchers turned to a mentalistic description of agents, based on folk psychology, employing such notions as beliefs, desires and intentions (BDI for short). The idea being that agents show autonomous (reactive, proactive and social) behaviour if this is based on a proper consideration of these attitudes when deliberating their next action. Following the philosophical treatment of Michael Bratman ([4]) researchers on the one hand started to make this more precise by means of logical specifications ([8, 36, 25, 26, 31]), and on the other hand devised more concrete architectures for agent-based systems, such as IRMA, PRS and InteRRap ([5, 36, 14, 32, 51]). In this period our group focused on the logic of agents by employing a logic of action (dynamic logic), called KARO, enriched with modal operators for BDI-like attitudes ([25]). In a sense this work was the natural continuation of our earlier work on logics for artificial intelligence, and in particular on the logic of knowledge and belief ([20, 30, 27]) and the (modal) logic of action (e.g. [6, 23]). The logic of knowledge, or epistemic logic, is about the logical properties of the knowledge of agents, such as that knowledge is true ($K\phi \rightarrow \phi$), and knowledge is known ($K\phi \rightarrow KK\phi$). Modal action logics such as dynamic logic enables one to reason about the effects of performing actions, including the execution of programs. In KARO, containing a combination of logics of action and knowledge and belief, we can express things like that the agent is aware of its commitments to perform actions (commitments are known, $Com\alpha \rightarrow KCom\alpha$), and that an agent that possibly intends to do an action α for the purpose of achieving a known goal ϕ, meaning that it knows that it is able to perform the action α and that performing α will result in the achievement of ϕ, may commit itself to such an action $(K(\langle\alpha\rangle\phi \wedge Able(\alpha) \wedge Goal\phi) \rightarrow \langle commit\alpha\rangle Com\alpha)$. Lately work on logics of agents has shifted more to social aspects, see below.

3. BDI+ Agents, Sapient Agents

Recently we have come to realize that we can draw further inspiration from cognitive science and cognitive models in particular. Since a couple of years we have been working on incorporating emotions into our agent models, with the aim of constructing agent systems that on the one hand are more efficacious (the use of emotions as heuristics in decision-making) and more believable on the other (in applications in which agents are interacting with users, see below), and we have obtained some first results here ([28, 42]. I refer to cognitive agents with mentalistic capabilities beyond pure BDI as 'BDI+' agents.

Furthermore, we explored the possibility to enhance the intelligence of agents even further, and coined a notion of a 'sapient agent' ([34, 33]). This denotes a general kind of agent that can handle different tasks simultaneously, situated in an environment, involved in sustained activity over longer periods of time, and for this purpose possessing accumulated knowledge and learning capabilities. Sapient agents are also presumed to be adaptive and have 'insight' and 'judgement' in the sense that they are able to use their mentalistic (BDI+) attitudes together with sensing, communication and learning capabilities, and the capability to reflect on and reason about these capabilities, to determine the 'right way to

go'. Of course, as this is a very ambitious idea, one needs to develop more concrete methods and techniques for realizing sapient agents. One of these techniques we have been focusing on in Utrecht that can be viewed as a first step towards the realization of sapient agents, is that of programming cognitive (BDI-like) agents by means of a concrete programming language as we will describe below. Other lines of research in this direction, increasing agents' capabilities include work on adjustable autonomy. Although there is little consensus in the literature on this topic (as with agent-hood itself), we have considered autonomy in the sense of having control over external influences and adjustable autonomy as dealing dynamically with this. In particular we have investigated in which way the mental state, and especially the beliefs, of an agent can be allowed to be influenced by other agents ([45, 46]). Other work is related to mobile services (see below), and investigates a more dynamic view on deliberation: in these applications it makes sense to interrupt the normal deliberation cycle such as the generation and execution of plans (since they are not useful at that moment), which may be resumed later once it becomes useful / relevant again [22].

4. Multi-Agent Systems

Most current research about agents is about multi-agent systems (MAS), in which multiple agents share a common environment and have to interact and coordinate with each other ([51]). So here we are concerned with social properties of agents: how do they communicate, coordinate, negotiate, cooperate or compete? In our group we have done several projects concerning aspects of multi-agent systems. We have looked at programming languages for agent communication ([13]), and especially we have studied how techniques from the area of concurrent and distributed computing can be fruitfully employed to obtain programming constructs for agent communication and negotiation together with sound semantics and verification methods. We have also considered how heterogeneous agents, that is, agents that have different ontologies (concepts; "speak different languages") could possibly communicate with each other in a meaningful way and make decisions on the information gained from their dialogues ([13, 11, 24]). In fact, with a company called Emotional Brain, we have investigated how, based on these ideas, to make practical systems (Heterogeneous MAS) for reasoning about complex interdisciplinary domains such as medicine including pharmacological and psychological aspects. Another, related, strand of research looks at how agents can reason with each other (and persuade each other) of certain things (to know or to do) by using formal argumentation in the style of [35, 47], where also agent-oriented concepts such as BDI, values and personalities play a role ([50]). For instance, from cognitive science it is known that it depends on the (human) agent's personality what arguments are convincing for him/her.

Other projects related to the issue of MAS that we have done or are currently doing, partly in cooperation with other universities and institutes, include: programming coordination of agents ([48, 3]), multi-agent planning ([49], the use of logic for the analysis of game-theoretical notions ([18]), the use of a combination of multi-agent systems and computational economy to repair problems with plans in air traffic management ([21]), and last but not least a method(ology) for designing multi-agent systems (OperA, [12]).

OperA is an elegant framework for MAS (or agent societies), comprising three models: an organisational model, in which the organisational structure of the society is described, consisting of roles and interactions as intended by the organisational stakeholders, a social model, in which agents are assigned/linked to roles via social contracts, and finally an interaction model, in which the possible interactions between agents are described. An important feature is that in this way the organisational specification is separated from the internals of the agents involved. Currently we are also looking at logical issues pertaining to MAS, in particular we are interested in logics that describe how agents should ideally act within an agent society where the interests of groups of agents (coalitions) is taken into account [7]: as is well-known from game theory the optimal way to act may depend on the coalition considered, with as special cases the individual agent on its own and the group of all agents.

4.1. **Normative Systems.** An especially interesting topic within MAS research concerns how to deal with the regulation of the (more or less) autonomous agents within such a MAS or agent society. So here the challenge is to find a balance between the autonomy of the individual agents and the desired overall behavior of the system. Here one sees solutions inspired by the human society: using norms one specifies how the agent should behave, and next one devises systems such as 'electronic institutions' to monitor, regiment or enforce the desired behavior upon the agents. In general one may call these systems *normative systems* or *normative MAS*. In our group we have conducted both theoretical and more practical research how such normative systems can be specified and realized (e.g. [1, 15]). Currently (e.g. [44]) we also are looking at how to integrate ways of programming these normative systems with programming individual agents, an important topic of our group during the last decade, to which we will turn now.

5. Agent-Oriented Programming

Since the second half of the 90s we have paid much attention to the question of how to program agent systems. The subfield of agent-oriented programming was initiated by Yoav Shoham with his proposal of the language AGENT0 [41], in which for the first time BDI-like concepts such as beliefs and commitments could be employed in the language for writing agent programs.

In Utrecht Koen Hindriks came up with the agent language 3APL (An Abstract Agent Programming Language), for which a rigorous and formal semantics was provided, which formed the basis of our implementation(s) of the language ([19, 39, 37]). This was an improvement over Shoham's work which lacked such a rigorous semantics. We deem semantics to be very important. It provides precise operational meaning of the concepts and operations / constructs used in a language. Especially in an area such as agent technology where notions are loosely based on human concepts and thus inherently vague, this is of the utmost importance!

3APL is a rule-based language with both features from imperative and logic programming, with as key construct a so-called 'practical reasoning' (or 'plan revision', PR) rule

of the form

$$\pi \leftarrow \varphi \mid \pi'$$

where π, π' stand for procedural goals (or plans) and φ for a belief. The interpretation of the rule is that if the current plan is π and φ follows from the belief base of the agent then π can be replaced by the plan π'.

Later we, and especially Birna van Riemsdijk and Mehdi Dastani, developed the ideas behind 3APL further, also extending the notions that are used in programming BDI-like cognitive agents. For example, also *declarative* goals ('goals-to-be', describing desirable 'states' rather than 'actions') were introduced, and much attention was paid to their semantics [10, 37, 38]). This led to the incorporation of rules of the form

$$\gamma \leftarrow \varphi \mid \pi$$

where γ is a declarative goal, φ is a belief and π is a plan, with as interpretation that if γ is a current declarative goal and φ is believed then the plan π can be generated and put into the plan base of the agent. Therefore this kind of rules is called 'planning goals' (or 'plan generation', PG) rules. Eventually this development led to the language 2APL (A Practical Agent Programming Language)[9], which we are currently developing further and using for various applications.

6. Applications

During the last years we have looked at various applications of agent technology in diverse domains. The main aim of this is to investigate the practicality and usefulness of this new technology in general, and our implementation of it through 3APL/2APL more in particular. Here we have focused mainly on applications that could be viewed as resorting under the areas of human media interaction and (to a lesser extent) logistics and computational economy.

As for the former we are conducting projects on the use of agent technology in modelling and generating/constructing video game characters ([40]), companion robots ([2, 43]), explainable AI [16, 17]. With respect to video game characters we are interested in modelling them with BDI or BDI+ agents, and also in capabilities to derive ('abduce') the mental attitudes such as beliefs, goals and intentions of virtual characters from their behaviours, so that characters equipped with these capabilities can predict the behaviour of other agents and anticipate on it in a believable manner. As to companion robots we are concerned with the reasoning and communication (dialogues) with a robot that is designed to help elderly people with kitchen tasks. Regarding explainable AI we are interested in designing and constructing virtual trainings for (para)military situations, in which BDI agents are employed that also can explain their actions and decisions to students/trainees. The projects on mental state abduction and explainable AI are being carried out under the umbrella of the GATE project (Game Research for Training and Entertainment).

As to applications in logistics and computational economy, we have researched multi-agent methods to deal with plan repair in the context of air traffic management, especially for the question how to devise an efficient but equitable (for the parties concerned) solution

to disturbances in the planning of gate assignments ([21]). Finally, as said before, we have also considered the application of agents for mobile services [22]. Mobile applications are interesting since they pose new challenges for agent technology. For instance, if one would like to find restaurants in the city where one is driving, an agent may plan a route to such a restaurant; however if for some reason we drive too far off from the city, on the highway, for example, this particular choice makes no sense anymore. Then it is fruitful to stop the planning and look at other possibilities/goals/plans. Returning to the same area later would make it sensible to resume the earlier deliberation on getting to a nearby restaurant in the city again. This calls for a reconsideration of the standard deliberation cycle in agent systems programming ([22]).

7. Conclusion

In this paper I have sketched some of the main lines of our research in agent technology, that developed from theoretical/logical foundations via the development of an agent programming language to several applications. It is tempting to speculate on the question where agent technology is heading for. On the one hand it is clear that because of its cognition-inspired models it will never become a standard for generic programming, on the other hand it is clear that a lot of 'intelligent' systems that are currently emerging can be dealt with by agent technology in some form or another.

8. Postscript for Wiebe

I've written this article on the occasion of 25 years of computer science in Utrecht. Here I'd like to add some words especially dedicated to Wiebe. Wiebe, as you can see from the article (and the selected 'wiebliography' below) we have done rather a lot in Amsterdam and Utrecht over the years. As my first PhD student in the direction of Logics for AI, you were right at the center and origin of it all. I really think that without you we would not have come that far. We were working already on agents and particularly agent logics without even knowing it. Your own excellent PhD work was seminal for our group and gave birth to work continued by our students, in the first instance by our first joint student Bernd van Linder (who has proved, btw, that doing possible world semantics can also lead you to the heart of very real worlds like the financial one!).

Wiebe, I treasure the memories of our long-standing working relation in Amsterdam and Utrecht. We got along extremely well. At your PhD party you said that this was perhaps due to the fact that I was a very young professor and you were an very old PhD student. ;-) Well, at least it made the difference in age rather small, and in fact you are exactly as old as my younger brother (only two weeks difference). Although we do not write many joint papers anymore, we have, somewhat paradoxically, kept a close friendship over a distance. I'm very proud of what you've accomplished in Liverpool during the last 6.5 years and hope that we'll remain colleagues and friends for a long time to come!

Congratulations on your 50th birthday!

9. Selected Annotated Joint Wiebliography

In this section I have chosen 11 out of more than 50(!) joint publications.

(1) W. van der Hoek & J.-J.Ch. Meyer, Possible Logics of Belief, Logique & Analyse 127-128, 1989, pp.177-194.

This was our first paper together, in which we explored our new territory of doxastic logic.

(2) W. van der Hoek & J.-J. Ch. Meyer, Making Some Issues of Implicit Knowledge Explicit, Int. J. of Foundations of Computer Science 3(2), 1992, pp. 193-223.

We started this because I was faced with a problem in deontic logic concerning intersection of accessibility relations. Here your great technical skills became already apparent!

(3) W. van der Hoek, M. van Hulst & J.-J. Ch. Meyer, Towards an Epistemic Approach to Reasoning about Concurrent Programs, Proc. REX Workshop Beekbergen 1992 (J.W. de Bakker, W.P. de Roever & G. Rozenberg, eds.), LNCS 666, Springer, Berlijn, 1993, pp. 261-287.

This was the result of your excursion to Nijmegen, my second affiliation at that time, where Marten van Hulst was a PhD student of mine.

(4) W. van der Hoek, B. van Linder & J.-J. Ch. Meyer, A Logic of Capabilities. Proc. 3rd Int. Symp. on the Logical Foundations of Comp. Sc. (LFCS'94), (A. Nerode & Yu.V. Matiyasevich, eds.), LNCS 813, Springer-Verlag, Berlin, 1994, pp. 366-378.

Our first paper with Bernd van Linder, which constitutes the birth of the KARO formalism.

(5) J.-J. Ch. Meyer & W. van der Hoek, Epistemic Logic for AI and Computer Science, Cambridge Tracts in Theoretical Computer Science 41, Cambridge University Press, 1995.

The book that we wrote together, based on my lectures in Amsterdam. It took something like 7 years to complete, from initial lecture notes to printed book. I believe the elaboration of the exercises alone already took us more than a year...

(6) J.-J. Ch. Meyer, W. van der Hoek & B. van Linder, A Logical Approach to the Dynamics of Commitments, AI Journal 113, 1999, pp. 1-40.

Our AI Journal publication on KARO, particularly the 'motivational' part. It almost got rejected since the editor misinterpreted the reviews in the first instance. Fortunately we had the audacity to protest. ;-)

(7) K.V. Hindriks, F.S. de Boer, W. van der Hoek & J.-J. Ch. Meyer, Agent Programming in 3APL, in Int. J. of Autonomous Agents and Multi-Agent Systems 2(4), 1999, pp. 357-401.

Our journal paper with Koen Hindriks on 3APL, which would influence our work in Utrecht on agent programming for many years.

(8) F.S. de Boer, R.M. van Eijk, W. van der Hoek & J.-J. Ch. Meyer, A Fully-Abstract Model for the Exchange of Information in Multi-Agent Systems, Theoretical Computer Science 290, 2003, pp. 1753-1773.

One of the nice papers on agent communication we wrote with Frank de Boer and Rogier van Eijk.

(9) W. de Vries, F.S. de Boer, K.V. Hindriks, W. van der Hoek &. J.-J. Ch. Meyer, A Programming Language for Coordinating Group Actions, in: Proc. of the Second International Workshop of Central and Eastern Europe on Multi-Agent Systems (CEEMAS01) (B. Dunin-Keplicz & E. Nawarecki, eds.), 2001, pp. 297–304.

One of the things we did with Wieke de Vries when she returned to Utrecht as a kind of prodigal daughter, after her bold expedition to Amsterdam. ;-)

(10) B.P. Harrenstein, W. van der Hoek, J.-J. Ch. Meyer & C. Witteveen, A Modal Characterization of Nash Equilibrium, Fundamenta Informaticae 57(2-4), 2003, pp. 281–321.

I remember that when giving a talk about this joint work with Paul Harrenstein and Cees Witteveen at a workshop in Poland, Witold Lukascewicz said to me jokingly: "What *can't* you guys do with modal logic...?"

(11) M.B. van Riemsdijk, W. van der Hoek & J.-J. Ch. Meyer, Agent Programming in Dribble: from Beliefs to Goals Using Plans, in: Proc. 2nd Int. J. Conf, on Autonomous Agents and Multiagent Systems (AAMAS03)(J.S. Rosenschein, T. Sandholm, M. Wooldridge & M. Yokoo, eds.), Melbourne Australia, ACM Press, New York, 2003, pp. 393–400.

This was the wonderful result of the Master's thesis work by Birna van Riemsdijk, who later did a PhD with me.

References

[1] H. Aldewereld, Autonomy vs. Conformity: An Institutional Perspective on Norms and Protocols, PhD Thesis, UU, Utrecht, 2007.

[2] R.J. Beun, R.M. van Eijk, J.-J. Ch. Meyer & N.L. Vergunst., A Computational Approach to the Interpretation of Indirect Speech Acts, in Proc. International Conference on Multidisciplinary Information Sciences and Technologies, (V.P. Guerrero-Bote, ed.), Open Institute of Knowledge, Mrida, 2006, pp. 311-315.

[3] F.S. de Boer, C. Pierik, R.M. van Eijk & J.-J. Ch. Meyer, Coordinating Agents in OO, in: Objects, Agents, and Features (M. Ryan, J.-J. Ch. Meyer & H.-D. Ehrich, eds., LNCS 2975, Springer, Berlin, 2004, pp. 8-25.

[4] M.E. Bratman, *Intentions, Plans, and Practical Reason*, Harvard University Press, Massachusetts, 1987.

[5] M.E. Bratman, D.J. Israel & M.L. Pollack, Plans and Resource-Bounded Practical Reasoning, Computational Intelligence 4, 1988, pp. 349-355.

[6] J. Broersen, Modal Action Logics for Reasoning about Reactive Systems; PhD Thesis, VUA, Amsterdam, 2003.

[7] J. Broersen, R. Mastop, J.-J. Ch. Meyer & P. Turrini, A Deontic Logic for Socially Optimal Norms, in: Deontic Logic in Computer Science (Proc. DEON 2008) (R. van der Meyden & L. van der Torre, eds.), Luxembourg, LNAI 5076, Springer, Berlin/Heidelberg, 2008. pp/ 218-23

[8] P.R. Cohen & H.J. Levesque, Intention is Choice with Commitment, *Artificial Intelligence* 42(3), 1990, pp. 213–261.

[9] M. Dastani & J.-J. Ch. Meyer, A Practical Agent Programming Language, in Pre- Proc. AAMAS07 Workshop on Programming Multi-Agent Systems (ProMAS2007) (M. Dastani, A. El Fallah Seghrouchni, A. Ricci & M. Winikoff, eds.), Honolulu, Hawaii, 2007, pp. 72-87.

[10] M. Dastani, M.B. van Riemsdijk, F. Dignum & J.-J. Ch. Meyer, A Programming Language for Cognitive Agents: Goal-Directed 3APL, in: Programming Multi-Agent Systems (Proc. ProMAS 2003) (M. Dastani, J. Dix, & A. El Fallah-Seghrouchni, eds.), LNAI 3067, Springer, Berlin, 2004, pp. 111-130.
[11] J. van Diggelen, Achieving Semantic Interoperability in Multi Agent Systems: A Dialogue-Based Approach, PhD Thesis, UU, Utrecht, 2007.
[12] V. Dignum, A Model for Organisational Interaction: Based on Agents, Founded in Logic, PhD Thesis, UU, Utrecht, 2004.
[13] R. van Eijk, Programming Languages for Agent Communication, PhD Thesis, UU, Utrecht, 2000.
[14] M.P. Georgeff & A.L. Lansky, Reactive Reasoning and Planning, in: Proceedings of the 6th National Conference on Artificial Intelligence (AAAI-87), Seattle, WA, 1987, pp. 677-682.
[15] D. Grossi, Designing Invisible Handcuffs: Formal Investigations in Institutions and Organizations for Multi-agent Systems, PhD Thesis, UU, Utrecht, 2007.
[16] M. Harbers, K. van den Bosch, F. Dignum & J.-J. Ch. Meyer, A Cognitive Model for the Generation and Explanation of Behavior in Virtual Training [Systems], in Proc. ECAI 2008 Workshop on Explanation-aware Computing (ExaCt 2008), (T.R. Roth-Berghofer, S. Schulz, D. Bahls & D.B. Leake, eds.), University of Patras, Patras, Greece, 2008, pp: 99-110.
[17] M. Harbers, F. Dignum, J.-J. Ch. Meyer & K. van den Bosch, Explaining Simulations through Self Explaining Agents, in: Proc. Workshop Epistemological Perspectives on Simulation (EPOS), (N. David, J.C. Caldas & H. Coelho, eds.), Lisbon, Portugal, ISCTE, 2008, pp: 85-100.
[18] B.P. Harrenstein, Logic in Conflict: Logical Explorations in Strategic Equilibrium, PhD Thesis, UU, Utrecht, 2004.
[19] K.V. Hindriks, Agent Programming Languages: Programming with Mental Models, PhD Thesis, UU, Utrecht, 2001.
[20] W. van der Hoek, Modalities for Reasoning about Knowledge and Quantities, PhD Thesis, VUA, Amsterdam, 1992.
[21] G. Jonker, Effcient and Equitable Exchange in Air Traffc Management Plan Repair Using Spender-Signed Currency, PhD Thesis, UU, Utrecht, 2008.
[22] F. Koch, J.-J. Ch. Meyer, F. Dignum & I. Rahwan, Programming Deliberative Agents for Mobile Services: The 3APL-M Platform, n: Programming Multi-Agent Systems: Third International Workshop (ProMAS 2005) (R.H. Bordini, M. Dastani, J. Dix & A. El Fallah-Seghrouchni , eds.), LNAI 3862, Springer, Berlin/Heidelberg, 2006, pp. 222-235.
[23] M. Kracht, J.-J. Ch. Meyer & K. Segerberg, The Logic of Action, The Stanford Encyclopedia of Philosophy (2009 Edition), Edward N. Zalta (ed.), to appear.
[24] H.-J. Lebbink, Dialogue and Decision Games for Information Exchanging Agents, PhD Thesis, UU, Utrecht, 2006.
[25] B. van Linder, Modal Logics for Rational Agents, PhD Thesis, UU, Utrecht, 1996.
[26] J.-J. Ch. Meyer, Intelligent Agents: Issues and Logics, in: Logics for Emerging Applications of Databases (J. Chomicki, R. van der Meyden & G. Saake, eds.), Springer, Berlin, 2004, pp. 131-165.
[27] J.-J. Ch. Meyer, Modal Epistemic and Doxastic Logic, in: Handbook of Philosophical Logic (2nd edition) (D. Gabbay & F. Guenthner, eds.) Vol. 10, Kluwer, Dordrecht, 2003, pp. 1-38.
[28] J.-J. Ch. Meyer, Reasoning about Emotional Agents, *Int. J. of Intelligent Systems* 21 (6), 2006, pp. 601-619.
[29] J.-J. Ch. Meyer, Agent Technology, in: Encyclopedia of Computer Science and Engineering (B.W. Wah, ed.), Wiley, to appear.
[30] J.-J. Ch. Meyer & W. van der Hoek, *Epistemic Logic for AI and Computer Science*, Cambridge University Press, Cambridge, UK, 1995.
[31] J.-J. Ch. Meyer & F. Veltman, Intelligent Agents and Common Sense Reasoning, Chapter 18 of: P. Blackburn, J.F.A.K. van Benthem & F. Wolter (eds.), Handbook of Modal Logic, Elsevier, 2007, pp. 991-1029.

[32] J. Müller, A Cooperation Model for Autonomous Agents, in: Intelligent Agents, III (J.P Müller, M. Wooldridge & N.R. Jennings, eds.), Lecture Notes in Artificial Intelligence 1193, Springer, Berlin, 1997, pp. 245-260.
[33] M. van Otterlo, The Logic of Adaptive Behavior: Knowledge Representation and Algorithms for the Markov Decision Processes Framework in First-Order Domains, PhD Thesis, Twente University, 2008.
[34] M. van Otterlo, M. Wiering, M. Dastani & J.-J. Ch. Meyer, A Characterization of Sapient Agents, in Proc. 2005 IEEE International Conference on Integration of Knowledge Intensive Multi-Agent Systems (KIMAS05,) Workshop on Sapient Agents (R.V. Mayorga & L.I. Perlovsky, eds.). Boston MA, USA, April 2005
[35] H. Prakken, Logical Tools for Modelling Legal Argument, PhD thesis, VU Amsterdam, 1993.
[36] A.S. Rao & M.P. Georgeff, Modeling rational agents within a BDI-architecture, in *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR'91)* (J. Allen, R. Fikes & E. Sandewall, eds.), Morgan Kaufmann, 1991, pp. 473–484.
[37] M.B. van Riemsdijk, Cognitive Agent Programming: A Semantic Approach, PhD Thesis, UU, Utrecht, 2006.
[38] M.B. van Riemsdijk, M. Dastani & J.-J. Ch. Meyer, Goals in Conflict: Semantic Foundations of Goals in Agent Programming, JAAMAS, to appear.
[39] M.B. van Riemsdijk, J.-J. Ch. Meyer & F.S. de Boer, Semantics of Plan Revision in Intelligent Agents, Theoretical Computer Science 351, 2006, pp. 240-257.
[40] M.P. Sindlar, M.M. Dastani, F. Dignum & J.-J. Ch. Meyer, Mental State Abduction of BDI-Based Agents, in: Declarative Agent Languages and Technologies VI: 6th Int. Workshop (DALT 2008) (M. M. Baldoni, S. Tran Cao, M.B. van Riemsdijk & M. Winikoff, eds.), LNAI 5397, Springer, Berlin, 2008, pp. 162-179.
[41] Y. Shoham, Agent-Oriented Programming, *Artificial Intelligence* 60(1), 1993, pp. 51–92.
[42] B. Steunebrink, M. Dastani & J.-J. Ch. Meyer, A Logic of Emotions for Intelligent Agents, in *Proc. AAAI-07* (R.C. Holte & A.E. Howe, eds.), Vancouver, Canada, AAAI Press, 2007, pp. 142-.147.
[43] B.R. Steunebrink, N.L. Vergunst, Chr.P. Mol, F.P.M. Dignum, M. Dastani & J.-J. Ch. Meyer, A Generic Architecture for a Companion Robot, in: Proc. 5th Int. Conf. on Informatics in Control, Automation and Robotics (ICINCO'08) Vol. 2, (J. Filipe, J.A. Cetto & J.-L. Ferrier, eds.), Funchal, Madeira, Portugal, 2008., pp. 315-321.
[44] N.A.M. Tinnemeier, M. Dastani & J.-J. Ch. Meyer, Orwells Nightmare for Agents? Programming Normative Multi-Agent Organisations, in Preproc. AAMAS08 Workshop on Programming Multi-Agent Systems (ProMAS08) (K. Hindriks, A. Pokahr & S. Sardina, eds.), Estoril, Portugal, 2008, pp. 43-58.
[45] B. van der Vecht, A.P. Meyer, R.M. Neef, F. Dignum & J.-J. Ch. Meyer, Influence- Based Autonomy Levels in Agent Decision-Making, in: Coordination, Organizations, Institutions, and Norms in Agent Systems II (Proc. COIN2006) (P. Noriega, J. Vzquez-Salceda, G. Boella, O. Boissier, V. Dignum, N. Fornara & E. Matson, eds.), LNAI 4386, Springer, 2007, pp. 322-337.
[46] B. van der Vecht, F. Dignum, J.-J. Ch. Meyer & M. Neef,, A Dynamic Coordination Mechanism Using Adjustable Autonomy, in: Coordination, Organizations, Institutions, and Norms in Agent Systems III (Proc. COIN 2007) (J. Sichman, P. Noriega, J. Padget & S. Ossowski , eds.), LNCS 4870, Springer, 2008, pp. 83-96.
[47] G.A.W. Vreeswijk, Studies in Defeasible Argumentation, PhD thesis, VU Amsterdam, 1993.
[48] W. de Vries, Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems, PhD Thesis, UU, Utrecht, 2002
[49] M. de Weerdt, Plan Merging in Multi-Agent Systems, PhD Thesis, Delft Univ. of Technology, 2003.
[50] T.L. van der Weide, F. Dignum, J.-J. Ch. Meyer, H. Prakken & G.A.W. Vreeswijk, Personality-Based Practical Reasoning, in Preproc. AAMAS08 Workshop on Argumentation in Multi-Agent Systems (ArgMAS 2008) (I. Rahwan & P. Moraitis, eds.), Estoril, Portugal, 2008, IFAAMAS, pp. 76-93
[51] M.J. Wooldridge. An Introduction to MultiAgent Systems. John Wiley & Sons. Chichester, 2002.

Concurrently decomposable constraint systems

Wiebe van der Hoek
University of Liverpool

Brammert Ottens
Ecole Polytechnique Federale de Lausanne

Nico Roos
Maastricht University

Cees Witteveen
Delft University of Technology

Abstract

Preface

What should you submit as a contribution to a workshop in honor of the 50th birthday of a best friend? Instead of submitting an already published paper, or composing a story about all the things we have done together in the past, I decided to submit something we could do together in the future. So, here is an unfinished paper to contribute to the Wiebe fest. Originally, the contents of this paper were planned to be written together with Wiebe and two other colleagues. Although we planned to submit it to a suitable conference at the end of 2008, it did not work out, since the four of us had other, more urgent, papers to work on and the ideas put forward in this paper were (are) still in their infancy. We made some progress in the very last weeks of 2008, but at at that time we all felt that the paper was not ready for submission. Meanwhile, I've been looking again at some problems and now I feel that at least we have some results that are worthwhile to work on, although the number of problems we still have to solve outweighs the number of problems we think we have solved. Anyway, this paper still needs quite a lot of extensions and improvements, but I'm confident that it can be finished in the coming months. Therefore, Wiebe, instead of repeating what we have accomplished in the past, let this paper in concept be a sign of our cooperation in the past, present and the future. Maybe a finished and polished version can be submitted to the Wiebe 60 Fest ...

1 Introduction

The problem we want to discuss is very simple to state: Let Σ be a set of formulae (or constraints) over a partitioned set $X = \{X_i\}_{i=1}^n$ of variables. Each block X_i of variables is controlled by an actor A_i who, independently from the other actors, tries to find a satisfying assignment σ_i for its subset $\Sigma_i \subseteq \Sigma$ of formulas over X_i. Suppose that these local assignments σ_i are simply composed to a global assignment σ for X. Can we guarantee σ to be a satisfying assignment for Σ, if we don't have any control over the choice of the locally satisfying assignments σ_i?

Let's start with a simple example to illustrate this problem. Suppose that Alice and Bob plan a party. Alice would like to invite Charles (c) or Diane (d) or both, while Bob independently from Alice wants to make an (inclusive) choice between Fred (f), Gerald (g) and Harald (h). It is also known, however, that inviting both Charles and Gerald will result in a really disastrous party, hence, we would not like to have them both invited. So, let $\Sigma = \{c \vee d, f \vee g \vee h, \neg(c \wedge g)\}$ and let $X = \{\{c, d\}, \{f, g, h\}\}$.

Now Alice, controlling $X_1 = \{c, d\}$ chooses a satisfying assignment for $\Sigma_1 = \{c \vee d\}$, while Bob, controlling $X_2 = \{f, g, h\}$, takes a satisfying assignment for $\Sigma_2 = \{f \vee g \vee h\}$. In this case, it is easy to see that we can't guarantee the existence of a globally satisfying assignment σ. For example, Alice might choose $\sigma_1 = \{c = 1,\ d = 0\}$ while Bob might take $\sigma_2 = \{f = 0,\ g = 1, h = 1\}$, but then the constraint $\neg(c \wedge g)$ is violated by the simple composition $\sigma = \{c = 1,\ d = 0, f = 0,\ g = 1, h = 1\}$ of these assignments, implying that the party is over.

Some questions can be raised almost immediately. For example, it is clear that the composition of satisfying local assignments cannot always be guaranteed to be a satisfying global assignment, but how difficult is it to *detect* exactly when such a guarantee can be given or not? What is the connection between this problem and deciding (in)consistency of the set of formulae Σ? Does it help if we already know some assignments satisfying Σ? Note that these questions are all pertaining to the role of the available *information* about the constraint system in answering this decision problem.

There is however, another, more constructive, way of looking to this problem. Suppose that we have a no-instance of the problem, that is an instance for which such a global satisfying assignment cannot always be guaranteed, could we, by *minimally changing* the instance, turn it into a yes-instance? For example, if in the example above we could force Bob to choose between Fred and Harald, the existence of a globally satisfying assignment can be guaranteed. But of course, this also raises a lot of questions, like:

What is minimality in this respect? What are the allowable changes? How difficult is it to find such changes?

In this (preliminary version of the) paper, we will only touch upon some of these questions. First, we will briefly discuss some related work and provide some general motivation for this research subject. Then we provide some formalisation by introducing a general framework for distributed constraint satisfaction problems and we define our problem in a more precise way. Next, we discuss the complexity of some associated decision problems and the problems associated with minimally changing the problem such that a global satisfying assignment can be guaranteed.

2 Motivation and Background

In constraint satisfaction, decomposition is a common technique to split a problem in a number of parts in such a way that the global solution can be efficiently assembled from the solutions of the parts. Most of these decomposition techniques that have been applied are *structurally* motivated, that is the decomposition is performed by analyzing the structure of the set C of constraints. The goal of such decomposition techniques is to split the original problem into a set of problems that can be solved in an easier way. The sets variables associated with resulting subproblems do not need to be completely disjunct, but do need to cover the global set X of variables. In general, the subproblems are easy to solve by guaranteeing that the resulting subproblems are *acyclic*: it is well-known that acyclic constraint solving problems can be solved efficiently, i.e. in polynomial time [1].

There is an extensive set of literature [2, 4, 5, 6, 8, 10] on this constraint decomposition problem with quite a number of different approaches to achieve suitable decompositions like e.g. bi-connected components, (hyper)tree decomposition, hinge decomposition, query decomposition, tree clustering methods, and so on. A common aspect of all these approaches, however, is that *(i)* the structure of the problem dictates the way in which the subproblems are generated and *(ii)* the subproblems generated do not have to be completely independently solvable, that is, in general, the decomposition will not allow the problems to be *concurrently* solvable.

The problem we are interested in differs in some respects from the problem solved by these structural decomposition methods. First of all, while in structural decomposition methods the partitioning (or covering) of the variables is a *result* of the decomposition and depends upon the structure of the constraint problem, we are interested in decomposition methods that

take a *given* partitioning of the variables into account. This is more in line with applications where a general (constraint) problem has to be solved by different parties that, each using their own approach, do want to solve their own part of the problem. This own part is determined by the capacities of these parties and not vice-versa.

Secondly, we require a *complete decomposition* of the original problem instance, that is, we would like to find a set of subproblems that can be solved *concurrently* and *independently* to obtain a complete solution to the original instance. This, in particular, makes our approach relevant to those problems where there are actors who want to solve their part of the constraint satisfaction problem in a completely *autonomous* way, without using communication (or without being able to communicate) and revision or backtracking necessitated by incompatible partial solutions proposed by other players.

In this paper, first of all, we address some decision problems associated with this concurrent decomposition approach. Then we want to enhance its applicability in the following way: if we detect that a system $\mathcal{S}$, whose set of variables X is partitioned, does not allow for a decomposition in independently solvable subsystems, we *minimally change* the constraint system $\mathcal{S}$ to a system $\mathcal{S}'$ in such a way that *(i)* every solution to $\mathcal{S}'$ is a solution to $\mathcal{S}$ and *(ii)* $\mathcal{S}'$ can be decomposed into independently solvable subsystems whose solutions always can be always reassembled to give a total solution to $\mathcal{S}$.

Before we state these problems in a more precise way, we first introduce some notational conventions and a general framework for discussing distributed constraint systems.

3 Preliminaries

We consider (abstract) constraint systems $\mathcal{S} = (X, D, C)$ where X is a (finite) set of variables X, D is a set of (value) domains D^i for every variable $x_i \in X$ and C is a set of constraints on X.

We assume constraints $c \in C$ to be specified as formulas over some language. A solution s of the system is an assignment $s = \{x_i := d_i\}_{i=1}^n$ of all variables in X such that each $c \in C$ is satisfied. A *partial solution* is just an assignment to a subset $X' \subseteq X$ of the variables. We will assume that we have a constant d_i in the language for every domain element d_i, and we often will identify d_i with d_i and think of an assignent $s = \{x_i := d_i\}_{i=1}^n$ as a *formula* $\bigwedge_{i=1}^n (x_i = d_i)$ or a set of formulas $\{(x_i = d_i)\}_{i=1}^n$. To preserve

generality, we don't feel the need to specify the set of allowable operators used in the constraints $c \in C$ and their interpretation.

By $Sol(\mathcal{S})$ we denote the set of solutions s, i.e., satisfying assignments, to a constraint system $\mathcal{S}$. The system $\mathcal{S}$ is called *consistent* if $Sol(\mathcal{S}) \neq \emptyset$. For every $c \in C$, let $Var(c)$ denote the set of variables mentioned in c. For a set of constraints C, we put $Var(C) = \bigcup_{c \in C} Var(c)$. Given $\mathcal{S} = (X, D, C)$ we obviously require $Var(C) \subseteq X$. Similarly, if D is a set of value domains D^i for variables $x_i \in X$ and we have $X' \subseteq X$, then $D_{X'}$ is the set of value domains $D^i_{X'}$ for variables $x_i \in X'$ with the obvious conditon that for all $x_i \in X'$, $D^i = D^i_{X'}$. Given a set of constraints C and a set of variables X' we let $C_{X'}$ denote the subset $\{c \in C \mid Var(c) \subseteq X'\}$. Furthermore, if X_1 and X_2 are subsets of X and s_1 is an assignment of values to variables in X_1 and s_2 an assignment of values to variables in X_2, then the *composition* $s = s_1 \sqcup s_2$ denotes an assignment of values to variables in $X_1 \cup X_2$. In particular, this assignment is well-defined if $\{X_1, X_2\}$ is a partitioning of X.

3.1 Simple constraint systems

Given a constraint system $\mathcal{S} = (X, D, C)$ we often want to concentrate on the constraints relevant to a subset X' of the variables X. Selecting such a subset of variables and the constraints associated with it will induce just another constraint system, being a subsystem of the original system:

Definition 1 *Let $\mathcal{S}_1 = (X_1, D_1, C_1)$ and $\mathcal{S}_2 = (X_2, D_2, C_2)$ be two constraint systems. Then we say that $\mathcal{S}_1$ is a subsystem of $\mathcal{S}_2$, written $\mathcal{S}_1 \sqsubseteq \mathcal{S}_2$, if the following holds:*

1. $X_1 \subseteq X_2$
2. $D_1 = D_{2_{X_1}}$
3. $C_1 = C_{2_{X_1}}$

Let $s = \{x_i := d_i\}_{i=1}^n$ be a solution for $\mathcal{S}_2 \sqsupseteq \mathcal{S}_1$. Then $s_{\mathcal{S}_1} = s_{(X_1, D_1, C_1)}$ is the assignment $\{x_i := d_i \mid x_i := d_i \in s, x_i \in X_1\}$.

It seems reasonable to assume that if a global constraint system $\mathcal{S} = (X, D, C)$ is consistent, any subsystem $\mathcal{S}' = (X', D_{X'}, C_{X'})$ with $X' \subseteq X$ derived from it, is also consistent.[1]

[1] But note that if we use an underlying nonmonotonic logic for the satisfaction relation this does not necessarily holds.

Note that, by definition of a constraint system $\mathcal{S} = (X, D, C)$, we have $Var(C) \subseteq X$. Now, given a constraint system $\mathcal{S}_2 = (X_2, D_2, C_2)$, there are at least three natural ways to obtain a subsystem $\mathcal{S}_1 = (X_1, D_1, C_1)$ from it:

1. Fix a set $X_1 \subseteq X_2$ and from that, derive D_1 and C_1 using Definition 1. In this case, we will write $\mathcal{S}_1 = \mathcal{S}_{2_{X_1}}$.

2. Fix a subset set D_1 of value domains from D_1 and find as set X_2 such that $D_1 = D_{2_{X_2}}$. This can always be done by removing from X_2 those variables that have a value in a doman in D_2 but not in D_1. The set of constraints C_1 is then also directly obtained: $C_1 = C_{2_{X_1}}$. We write: $\mathcal{S}_1 = \mathcal{S}_{2_{D_1}}$.

3. Fix a subset C_1 of the constraints of C_2 and find as set X_2 such that $C_1 = C_{2_{C_2}}$. This can always be done by removing from X_2 those variables that do occur in C_2 but not in C_1. The set of domain values D_1 is then also directly obtained: $D_1 = D_{2_{X_1}}$. We write: $\mathcal{S}_1 = \mathcal{S}_{2_{C_1}}$.

With respect to the subsystem relation $\sqsubseteq$, we assume our constraint systems to satisfy the following *preservation* property:

Preservation

Let $(X_1, D_1, C_1) = \mathcal{S}_1 \sqsubseteq \mathcal{S}_2$. Then $s \in Sol(\mathcal{S}_2)$ implies $s_{(X_1,D_1,C_1)} \in Sol(\mathcal{S}_1)$.

Constraint systems that satisfy Preservation will be called *Simple Constraint Systems*.

3.2 Distributed Constraint Systems

Usually, constraint systems $\mathcal{S}$ are distributed, that is, there is a set of actors A_i, each being able to make assignments or adding relations for/to only a subset X_i of variables and these agents are collectively responsible for producing a global solution for $\mathcal{S}$. More specifically, if $\mathcal{S} = (X, D, C)$ is a constraint system and $X_i \subseteq X$ is the subset of variables controlled by agent A_i then $\mathcal{S}_i = (X_i, D_{X_i}, C_{X_i})$ is the subsystem that has to be solved by agent A_i, where D_{X_i} is the set of domains for the variables in X_i and C_{X_i} is as defined above.

If $X = X_1 \cup X_2 \cup \ldots \cup X_n$ and $\bigcup_{i=1}^{n} X_i = X$, while for $1 \leq i \neq j = n$, $X_i \cap X_j = \emptyset$, the collection $\{X_i\}_{i=1}^{n}$ constitutes a *partitioning* of X, and each X_i is called a *block* of X. If $\{X_i\}_{i=1}^{n}$ is a partitioning of X, we let $\mathcal{S} = (\{X_i\}_{i=1}^{n}, D, C)$ denote a *distributed constraint* system.

4 The concurrent decomposition problem

In general, the *distributed constraint solving problem* can be simply stated as follows:

> Given a distributed constraint system $\mathcal{S} = (\{X_i\}_{i=1}^n, D, C)$ is it always possible to find a solution $s \in Sol(\mathcal{S})$ using solutions $s_i \in Sol(\mathcal{S}_i)$ for its induced subsystems $\mathcal{S}_i = (X_i, D_{X_i}, C_{X_i})$?

As we observed, while there are quite a few proposals for solving distributed systems, they almost all come down to some (distributed) backtracking process needed to resolve conflicts between partial solutions. Basically, what we are interested in are *backtracking-free* concurrent solutions. That is, we would like to investigate the following *concurrent decomposion* problem:

> Given a distributed constraint system $\mathcal{S} = (\{X_i\}_{i=1}^n, D, C)$, is it true that the composition $s = s_1 \sqcup s_2 \sqcup \ldots \sqcup s_n$ of arbitrary solutions $s_i \in Sol(\mathcal{S}_i)$, where $\mathcal{S}_i = (X_i, D_{X_i}, C_{X_i})$, is always a solution for the total system $\mathcal{S}$?

If the answer is yes, we say that a distributed constraint system is *concurrently decomposable*. More exactly we can define this property as follows:

Definition 2 (Concurrent decomposition) *A consistent distributed constraint system $\mathcal{S} = (\{X_i\}_{i=1}^n, D, C)$ is concurrently decomposable if,*

1. *for every $i = 1, \ldots, n$ there exists a consistent partial constraint system $\mathcal{S}_i = (X_i, D_{X_i}, C_{X_i})$ restricted to X_i, and*
2. *$Sol(\mathcal{S}_1) \sqcup \ldots \sqcup Sol(\mathcal{S}_n) \subseteq Sol(\mathcal{S})$, that is, for every $(s_1, \ldots, s_n) \in Sol(\mathcal{S}_1) \times \ldots \times Sol(\mathcal{S}_n)$ it holds that $s = s_1 \sqcup s_2 \ldots \sqcup s_n$ is well-defined and $s \in Sol(\mathcal{S})$.*

We note that most constraint systems $\mathcal{S}$ will not allow us to simply decompose $\mathcal{S}$ into partial constraint systems $\mathcal{S}_i$ derived from $\mathcal{S}$, determine the solutions s_i to the partial systems and then just merge or compose these (partial) solutions to obtain the solution to the original system.

Example 1 *Take a simple constraint system $\mathcal{S} = (X, D, C)$ where $X = \{x_1, x_2\}$ and is partitioned into $X_1 = \{x_1\}$ and $X_2 = \{x_2\}$, and $D^1 = D^2 = \mathbb{N}$. Let $C = \{x_1 \neq x_2\} \cup \{n_i < x_i \leq m_i : i = 1, 2\}$ for some given numbers $n_1 + n_2 + 5 < m_1 + m_2$. It is easy to see that the partial solutions s_1 for $\mathcal{S}_{\{x_1\}}$ and s_2 for $\mathcal{S}_{\{x_2\}}$ cannot always be joined to a global solution, since x_1 might be given the same value as x_2.*

Example 2 *Take a meeting scheduler, which has the aim to schedule two different meetings at a University in one and the same week. Meeting m_1 should be among faculty members of a specific Department, its Head, and its Dean, while meeting m_2 involves the Head of Department, the Dean and the Vice Chancellor. Moreover, we should keep in mind that no person can attend different meetings at the same time. Again, the partial solutions cannot always be joined to obtain a global solution.*

Note that in both cases splitting the problem into several parts means that some constraints c such as $x_1 \neq x_2$ (Example 1) and the constraint that two meetings cannot overlap if there is a person that should attend both of them (Example 2) are not taken into account while solving the partial constraint systems individually. These constraints are the so-called *inter-block* constraints. Therefore, concurrent decomposability should also be viewed upon as a specification of a special relation between the set of intra-block constraints C_{X_i} and this set of inter-block constraints.

It is not difficult to show that, indeed, if the sets C_{X_i} of constraints covered by the partition blocks X_i together imply *all* constraints in C, that is also the inter-block constraints $c \in C$ such that $Var(c)$ is not contained in a single partition block, then the constraint system is concurrently decomposable:

Proposition 1 *Let $\mathcal{S} = (\{X_i\}_{i=1}^n, D, C)$ be a consistent distributed constraint system and for $i = 1, \ldots, n$, let $\mathcal{S}_i = (X_i, D_{X_i}, C_{X_i})$. Then $Sol(\mathcal{S}_1) \times \ldots \times Sol(\mathcal{S}_n) \subseteq Sol(\mathcal{S})$ iff $\bigcup_{i=1}^n C_{X_i} \models C$.*

Proof. [Sketch] Assume that $Sol(\mathcal{S}_1) \times \ldots \times Sol(\mathcal{S}_n) \subseteq Sol(\mathcal{S})$. Take an arbitrary s satisfying $\bigcup_{i=1}^n C_{X_i}$. Then s can be written as $s = s_1 \sqcup s_2 \sqcup \ldots \sqcup s_n$ where each s_i satisfies C_{X_i} and therefore, $s_i \in Sol(\mathcal{S}_i)$. By assumption, $s \in Sol(\mathcal{S})$. Therefore, s satisfies C.
Conversely, assume $\bigcup_{i=1}^n C_{X_i} \models C$. Then every solution s satisfying $\bigcup_{i=1}^n C_{X_i}$ will satisfy C. Each such a solution s can be written as $s = s_1 \sqcup s_2 \sqcup \ldots \sqcup s_n$ where each s_i satisfies C_{X_i}. Hence, $Sol(\mathcal{S}_1) \times \ldots \times Sol(\mathcal{S}_n) \subseteq Sol(\mathcal{S})$. □

The last proposition almost immediately suggest that the problem whether a given distributed constraint system is concurrently decomposable or not is computationally closely related to deciding propositional logical consequence, which is co-NP complete. Indeed, as the next proposition shows, this is the case, even in the most simple distributed cases:

Proposition 2 *Let $\mathcal{S} = (\{X_i\}_{i=1}^n, D, C)$ be a distributed constraint system. The problem to decide whether $\mathcal{S}$ is concurrently decomposable is coNP-complete.*

Proof. Membership of coNP is easy: just guess a set of solutions $\{s_i\}_{i=1}^n$, where each s_i is a solution guessed for subsystem $\mathcal{S}_i$. Now check for each $i = 1, 2, \ldots, n$ whether $s_i \in Sol(\mathcal{S}_i)$ and then check the compatibility of these solutions. The violation of one or more inter-block constraints c can be easily verified using the composed global solution s.

Completeness follows using the following reduction from the coNP-complete Logical Consequence problem (Given a set of variables U, a set of clauses C over U and a clause c, is c implied by C?): Given an instance (U, C, c) of this problem, we consider the distributed constraint system $\mathcal{S} = (U \cup \{x\}, C \cup \{c'\} \cup \{\neg x\}, \{\{0,1\}^i\}_{i=1}^{|U|+1})$, where the set of variables is partitioned in the set U and the set $\{x\}$, and $c' = c \cup \{x\}$ is the clause c extended with the new atom $x \notin U$. Clearly, using this partitioning, $\mathcal{S}$ is decomposed into two systems $\mathcal{S}_U = (U, C, \{\{0,1\}^i\}_{i=1}^{|U|})$, while $\mathcal{S}_{\{x\}} = (\{x\}, \{\neg x\}, \{\{0,1\}\})$.

Now, let τ be any assignment verifying C, but falsifying c. Then τ is a solution for $\mathcal{S}_U$ and together with the only possible solution $\{x = 0\}$ for $\mathcal{S}_{\{x\}}$ it constitutes an assignment $\tau \sqcup \{x = 0\}$ that also falsifies $c' = c \cup \{x\}$. Hence $\tau \sqcup \{x = 0\}$ is a certificate for non-decomposability. For the converse, assume that there are local assignments τ_1 and τ_2 such that τ_1 is a solution for $\mathcal{S}_U$ and τ_2 is a solution for $\mathcal{S}_{\{x\}}$, while $\tau_1 \sqcup \tau_2$ does not satisfy $\mathcal{S}$. Then, clearly τ_1 satisfies C, but cannot satisfy c. Hence, τ_1 is a certificate for showing that C does not imply c. □

Note that this proof shows that this problem is already co-NP-complete for the simplest possible distributed case where a partition contains only 2 blocks.

Remark 1 It is well-known that for general constraint systems finding a solution is NP-hard [3]. We therefore might ask whether having additional information about a satisfying assignment would help us in solving the decomposition problem. This turns out not be the case:

Proposition 3 *Let $\mathcal{S} = (X, D, C)$ be a* consistent *constraint system, let $\{X_i\}_{i=1}^n$ a partitioning for X and $\sigma \in Sol(\mathcal{S})$ a satisfying assignment. Then the problem to decide whether $\mathcal{S}$ is concurrently decomposable is co-NP complete.*

Proof. Take a formula $\phi(x_1, x_2, \ldots x_n)$ over some alphabet $X = \{x_1, x_2, \ldots, x_n\}$. Without loss of generality we may assume that $n > 1$. Consider the constraint system $\mathcal{S} = (X \cup \{y\}, D, C)$ where $C = \{\phi(x_1, x_2, \ldots x_n) \vee y, y \vee \neg y\}$, D is a set of $\{0, 1\}$ domains and the partitioning of X is $X = \{\{x_i\}_{i=1}^n, \{y\}\}$. Let σ be an arbitrary assignment where $y = 1$. $\mathcal{S}$ is consistent and is concurrently decomposable exactly iff $\phi(x_1, x_2, \ldots x_n)$ is a tautology, the latter being a co-NP complete problem. □

It is easy to see that the same proof can be used to show that the availability of a partial solution that can be extended to a complete solution will not alleviate the difficulty of the decomposition problem.

Finally, note that if we have given *all* satisfying truth assignments σ to the constraint solving problem, the concurrent composability problem can be seen to be polynomially solvable, but of course, this comes at the price of an exponential blow-up: we might be forced to take into account exponentially many assignments.

5 Minimal Change and Concurrent Decomposability

As we have seen, the problem whether a distributed constraint problem is concurrently decomposable is an intractable problem (unless P=NP). But what happens if we could *change* a particular instance in such a way that it would become a yes-instance of the concurrent decomposition problem? Let us consider an example we discussed before:

Example 3 *Take the constraint system $\mathcal{S} = (X, D, C)$ where $X = \{x_1, x_2\}$ and is partitioned into $X_1 = \{x_1\}$ and $X_2 = \{x_2\}$, and $D^1 = D^2 = \mathbb{N}$. Let $C = \{x_1 \neq x_2\} \cup \{n_i < x_i \leq m_i\}$ for some given numbers $n_1 + n_2 + 5 < m_1 + m_2$. The system is not concurrently decomposable, but if we add the constraints 'x_1 is odd' and 'x_2 is even' to the set of constraints, joining individual solutions will always deliver a global solution.*

A first obvious restriction on the set of allowable changes of a given distributed constraint system $\mathcal{S} = (\{X_i\}_{i=1}^n, D, C)$ would be that the set of solutions of $\mathcal{S}$ is preserved: For every resulting system $\mathcal{S}'$ it must hold that $Sol(\mathcal{S}') \subseteq Sol(\mathcal{S})$.

Of course, one would immediately ask whether such allowable changes are always possible. The following proposition shows that every consistent

distributed constraint system can be turned into a concurrently decomposable one, without violating the solution preservation condition:

Proposition 4 *Given a consistent distributed constraint system* $\mathcal{S} = (\{X_i\}_{i=1}^n, D, C)$*, there always exists a concurrently decomposable distributed system* $\mathcal{S}' = (\{X_i\}_{i=1}^n, D, C')$ *such that* $Sol(\mathcal{S}') \subseteq Sol(\mathcal{S})$*.*

Proof. Since $\mathcal{S}$ is consistent, there exists a solution $s = \{x_i = d_i\}_{i=1}^n \in Sol(\mathcal{S})$. We show that the system $\mathcal{S}' = (\{X_i\}_{i=1}^n, D, C \cup s)$ is concurrently decomposable and satisfies the solution preservation condition. For an arbitrary i, take the subsystem $\mathcal{S}_i' = (X_i, D_{X_i}, C_{X_i} \cup \{s_{X_i}\})$. By Preservation, $s_{X_i} \in Sol(\mathcal{S}_i')$ and every solution $s' \neq s_{X_i}$ will violate at least one constraint $x_j = d_j$ occurring in s_{X_i}. Hence, for every $X_i \in \{X_i\}_{i=1}^n$, $Sol(\mathcal{S}_i') = \{s_{X_i}\}$. Likewise, we have $Sol(\mathcal{S}') = \{s\}$. Therefore, $\mathcal{S}'$ is concurrently decomposable and $Sol(\mathcal{S}') \subseteq Sol(\mathcal{S})$. □

Such solutions, however, are not always wanted, since they add quite a lot of additional constraints and seriously affect the set of solutions of the original system. In general, instead of adding an arbitrary set of constraints, we would like to apply the idea of minimal change: how could we *minimally* change the original system such that it becomes concurrently decomposable.

Applying this idea of minimal change, we could follow two different approaches:

1. *maximize* the set of solutions $Sol(\mathcal{S}')$ such that the difference $Sol(\mathcal{S}) - Sol(\mathcal{S}')$ is minimized;

2. minimize the amount of *constraint change* necessary to obtain the resulting system $\mathcal{S}'$.

We could view the first approach as a *semantically* inspired approach, and the second as a *syntactical* approach. While the latter ensures that the syntactical difference between the two constraint systems is minimized, the first approach does not care which syntactical changes have to be applied but takes care for minimizing the loss of information associated with the transition to a decomposable system.

Here the bad news is: both the syntactical and the semantical approach give rise to intractable problems. To start with the latter approach, let $\mathcal{S} = (\{X_i\}_{i=1}^n, D, C)$ be an instance of the distributed constraint problem. Following the semantical approach, we would like to obtain a distributed system $\mathcal{S}$ such that

1. $\mathcal{S}' = (\{X_i\}_{i=1}^n, D, C')$;
2. $Sol(\mathcal{S}') \subseteq Sol(\mathcal{S})$ and $Sol(\mathcal{S}) - Sol(\mathcal{S}')$ is minimal
3. $\mathcal{S}'$ is fully decomposable, i.e., $Sol(\mathcal{S}') = Sol(\mathcal{S}'_1) \times \ldots \times Sol(\mathcal{S}'_n)$

This problem, however, can be easily shown to be intractable, even if the set of solutions to the original system is of polynomial size and can be obtained in polynomial time:

Proposition 5 *Let $\mathcal{S} = (\{X_i\}_{i=1}^n, D, C) = (X, D, C)$ be a distributed constraint system. The problem to find a set of decomposed subsystems $\{\mathcal{S}'_i = (X_i, D_i, C'_i)\}_{i=1}^n$ such that* (i) $\Pi_{i=1}^n Sol(\mathcal{S}'_i) \subseteq Sol(\mathcal{S})$ *and* (ii) $\Pi_{i=1}^n Sol(\mathcal{S}'_i)$ *is a cardinality maximal subset of $Sol(\mathcal{S})$ is NP-hard.*

Proof. (Sketch) We use a reduction from the Max-Clique problem. Let $G = (V, E)$ be an instance of the Max-Clique problem. We create an instance of concurrent decomposition problem as follows: Let $\mathcal{S} = (\{X_i\}_{i=1}^n, D, C)$ be a distributed constraint system where $X = \{x_1, x_2\}$ partitioned as $X = \{x_1\} \cup \{x_2\}$ with domains $D_1 = D_2 = V$ and let C contain the constraint $r(x_1, x_2) = (x_1 = x_2) \vee (\{x_1, x_2\} \in E)$. Finding two subsystems $\mathcal{S}_1 = (\{x_1\}, V, C_1)$ and $\mathcal{S}_2 = (\{x_2\}, V, C_2)$ such that $x_1 \in C_1$, $x_2 \in C_2$ would imply $r(x_1, x_2)$. Note that C_1 and C_2 are unary relations defining subsets of V and maximizing $C_1 \times C_2$ under the constraint $C_1 \times C_2 \subseteq r$ comes down to finding a maximal clique in G. □

Taking the *syntactical* approach, let us define a distributed constraint system $\mathcal{S} = (\{X_i\}_{i=1}^n, D, C)$ to be *k-decomposable* if a partial assignment to k variables already suffices to decompose $\mathcal{S}$ in independently solvable subsystems. Here, we assume that after adding the k assignments to the variables, the original system of constraints is simplified by taking these assignments into account, i.e. replacing the variables by their values.

Note that a 0-decomposable distributed constraint system is just a concurrently decomposable system. Also note that a consistent distributed constraint system is always n-decomposable: just use a solution $s \in Sol(\mathcal{S})$ and add s to the set of constraints C.

The general problem to decide whether or not a system is k-decomposable for some $k \geq 0$ turns out to be harder than just checking whether the system is concurrently decomposable:

Proposition 6 *Let $\mathcal{S} = (X, D, C)$ be a constraint system, $\{X_i\}_{i=1}^n$ a partitioning of X and k a positive integer. The problem to decide whether $\mathcal{S}$ is k-decomposable is Σ_2^p-complete.*

Proof. To show that the problem is in Σ_2^p, given a constraint system $\mathcal{S}$ and a partitioning $\{X_i\}_{i=1}^n$ for X, guess a partial solution σ for k-variables in X and add the constraints $x = \sigma(x)$ for all $x \in dom(\sigma)$ to C. Then, use a co-NP-oracle to check 0-decomposability of the resulting constraint problem $\mathcal{S}'$.

To show that the problem is Σ_2^p-hard, we take the Σ_2^p-complete SUCCINCT SET COVER problem [9]: Given a collection $T = \{\phi_1, \phi_2, \ldots, \phi_m\}$ of 3-DNF formulae on a set Σ of variables and a positive integer k, is there a subset T' of T with $|T'| = k$ such that $\vdash \bigvee_{\phi \in T'} \phi$? The reduction from this problem is as follows: Let $(\Sigma, T = \{\phi_1, \phi_2, \ldots, \phi_m\}, k)$ be an instance of SUCCINCT SET COVER. Construct a distributed constraint system $\mathcal{S} = (\{X_i\}_{i=1}^n, D, C)$ where

1. $\{X_i\}_{i=1}^n = \{\Sigma, \{x_1, x_2, \ldots, x_m\}\}$, where for $j = 1, \ldots n$, $x_j \notin \Sigma$,
2. C contains two constraints $\Sigma_{\phi_i \in T}(\phi_i' + x_i) \geq 1$ and $\Sigma_{i=1}^m x_i = m - k$, where each ϕ_i' is obtained from ϕ_i by replacing $\vee$ by $+$, and $\wedge$ by $+$,
3. D a set of $\{0,1\}$-domains for each of the variables occurring in $\Sigma \cup \{x_1, \ldots x_m\}$.

Suppose there is a subset $T' \subseteq T$ of size k such that $\vdash \bigvee_{\phi \in T'} \phi$. Then, for every $\phi_i \notin T'$, add a constraint $x_i = 1$ to C. Now consider the subsystem $\mathcal{S}_\Sigma$. Due to the presence of the variables x_i, the set of constraints C_Σ is empty. Hence, any assignment s_1 to the variables in Σ can be proposed as a solution $s \in Sol(\mathcal{S}_\Sigma)$. Each such a solution s_1 will satisfy the constraint $\Sigma_{\phi_i \in T}(\phi_i' + x_i) \geq 1$ since $\vdash \bigvee_{\phi \in T'} \phi$, implies that $\Sigma_{\phi_i \in T'}(\phi_i' + x_i) \geq 1$ for every assignment s_1 to variables occurring in T' and every assignment to the variables x_i and $\Sigma_{\phi_i \in T - T'}(\phi_i' + 1) \geq 1$ is also satisfied by every such an assignment s_1.

Considering the subsystem $\mathcal{S}_{\{x_1, x_2, \ldots, x_m\}}$, we observe that $C_{\{x_1, x_2, \ldots, x_m\}} = \{\Sigma_{i=1}^m x_i = m - k\} \cup \{x_i = 1 : \phi_i \in T - T'\}$. Since $|T - T'| = m - k$, exactly one assignment will satisfy these constraints: the assignment s_2 that assigns 0 to all variables x_i such that $\phi_i \in T'$ and 1 to all variables x_i such that $\phi_i \notin T'$. Hence, every combination $s_1 \sqcup s_2$ will satisfy C.

The converse is proven along the same lines. □

Note that the proof of this proposition again show that Σ_2^p-completeness already holds for the simplest distributed case where we have a partition into two blocks.

6 Discussion

As announced in the beginning, this is an incomplete first draft of a paper investigating the problem of concurrent decomposability in distributed constraint problems. There still remains quite a lot to investigate. Let us mention a few problems that definitely have to be addressed. First of all, we have to investigate whether some of the complexity results obtained in the general case still do hold if we restrict the class of allowable constraints. For example, what are the most simple distributed constraint systems where decomposition is tractable? What are the distinguishing features of such classes? Next, we have to investigate what happens if in the case of minimal change we do not constrain the objects that can be added to partial assignments i.e., unary constraints, but also allow refinements of general constraints to be added. How does this influence the complexity of the k-decomposability and other syntactical notions of minimal change? For example, it can be shown that Simple Temporal Networks are minimally decomposable in polynomial time [7], if we are allowed to refine a certain subset of constraints and look at minimal refinements as a criterion for minimal change. Finally, we should concentrate on the construction of suitable heuristics that should provide additional constraints in order to make a given distributed constraint system decomposable.

Quite a lot of problems, but hopefully we also have quite a lot of time after Wiebe's fest!

References

[1] Catriel Beeri, Ronald Fagin, David Maier, and Mihalis Yannakakis. On the desirability of acyclic database schemes. *Journal of the ACM*, 30(3):479–513, 1983.

[2] David A. Cohen, Marc Gyssens, and Peter Jeavons. A unifying theory of structural decompostions for the constraint satisfaction problems. In *Complexity of Constraints*. Dagstuhl Seminar Proceedings 06401, 2006.

[3] R. Dechter. *Constraint Processing*. Morgan Kaufmann Publishers, 2003.

[4] Rina Dechter and Judea Pearl. Tree clustering for constraint networks. *Artif. Intell.*, 38(3):353–366, 1989.

[5] Georg Gottlob, Nicola Leone, and Francesco Scarcello. A comparison of structural csp decomposition methods. *Artificial Intelligence*, 124:2000, 1999.

[6] Chih-Wei Hsu, Benjamin W. Wah, Ruoyun Huang, and Yixin Chen. Constraint partitioning for solving planning problems with trajectory constraints and goal preferences. In *IJCAI*, pages 1924–1929, 2007.

[7] Luke Hunsberger. Algorithms for a temporal decoupling problem in multi-agent planning. In *In Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI-2002*, 2002.

[8] Wady Naanaa. A domain decomposition algorithm for constraint satisfaction. *J. Exp. Algorithmics*, 13:1.13–1.23, 2009.

[9] Marcus Schaefer and Christopher Umans. Completeness in the polynomial-time hierarchy: A compendium. *SIGACT News*, 33(3):32–49, September 2002.

[10] Benjamin W. Wah and Yixin Chen. Constraint partitioning in penalty formulations for solving temporal planning problems. *Artificial Intelligence*, 170(3):187–231, March 2006.

Expectations of Agents

Koen V. Hindriks
Delft University of Technology, Delft, The Netherlands
k.v.hindriks@tudelft.nl

February 21, 2009

Abstract

Expectations, among other things, are beliefs about the future. Expectations play a role in decision making and may influence the choice of action of an agent. Expectations may facilitate an agent's decision-making as it may make it easier to select an action to perform next. Vice versa, expectations may depend again on actions that the agent needs to perform itself in order to make these expectations come true. Expectations thus should not lead an agent to conclude it does not have to perform the actions required to make expectations come true. In Artificial Intelligence, the latter problem has been labelled the *Little Nel* problem. Expectations may also make an action seem less attractive, and make an agent reconsider performing the action. In this paper, we propose a model of how to balance these different roles of expectations in decision making. We present a formal model of action selection in the agent programming language GOAL as a formal idealized model of how expectations may be operative in an agent's decision making. We only treat the simple case where a single agent is the sole factor of change in the environment.

1 Introduction

Expectations, among other things, are beliefs about the future. Wiebe may not have foreseen that a workshop was being organized for his 50th birthday and therefore will have had quite different expectations related to March 16th 2009.

There has not been paid a lot of attention to the role of expectations in decision making within the agent community. Some work on expectations aiming to describe the "cognitive anatomy" of expectations has been reported in [1, 9], but this work has not focussed on the role of expectations in decision making in particular.

Wiebe, together with Wojciech and Mike, introduced a notion of *weak belief* in [13] that seems closely related to the concept of expectation. "Weak" belief is based on the actions an agent intends to perform and is *optimistic* in the sense that it assumes that the agents intentions are bound to succeed. As a result, an agent may believe that the (necessary) effects of all its intended actions will eventually occur. Correspondingly, in [13] an axiom of the form $I\chi \rightarrow WB\chi$ is introduced that expresses that intention to realize χ implies a weak belief in χ for linear temporal formulae χ. The axiom is to represent a very important feature of agents, namely, that the future is under control of the agent. In contrast, the notion of "strong" belief is *pessimistic* and only gives

rise to beliefs about the future that are independent from actions the agent would have to successfully perform. Strong belief thus is a stronger notion than weak belief and implies it. Although the notion of weak belief seems related to that of expectation, the latter notion is not discussed in [13].

We believe the distinction between two different forms of belief is interesting and useful, and we will use a similar strategy that distinguishes between two types of beliefs about the future. The first type of beliefs are those that are considered to be inevitable by the agent given the information available to it. These beliefs about the future concern properties of the world that the agent believes are not within its control, i.e. independent of any action it may perform. Here we will not be particularly concerned with how an agent comes to believe that certain things are inevitable, nor how an agent may justify having such beliefs.[1] Such beliefs about the future that are considered to be independent of an agent's own actions are simply called beliefs here, and relate to the notion of "strong" belief introduced in [13]. The second type of beliefs about the future are dependent on the agent's own actions and are called *expectations* here. It should be noted that the notion of expectation discussed here thus is more restrictive than the common sense concept of expectation. We focus in particular on the notion of expectation that depends on the successful performance of actions by the agent that has the expectation.

The paper is organized as follows. Section 2 discusses the role of expectations in decision making informally by means of a simple example. We will present a model of expectations in the context of the agent programming language GOAL. GOAL is an languages for programming rational agents, which derive their choice of action from their beliefs and goals. As it is natural to consider the role of expectations in the choice of action that an agent makes, introducing the concept of expectation into an agent programming language seems a useful extension in itself. Moreover, GOAL provides a formal semantic framework and by extending it we thus are able to provide a formal model of expectations. Section 3 presents a model of expectations as an extension of the GOAL programming framework. Finally, Section 4 concludes the paper.

2 The Role of Expectations in Decision Making

Here we informally discuss the role of expectations in decision making by means of a simple example.

Suppose an agent believes it will rain today and has a(n achievement) goal to be at work. Also suppose that the agent's only means to get to work is a bicycle. Initially, the agent does not expect to get wet. Only after deciding to go to work by bicycling the agent will expect to get wet and form the corresponding belief. As a result, because the agent will only consider bicycling to be an option if the agent will not get wet, the choice to bicylce will be reconsidered. Given the expectation to get wet, however, the agent may decide to wear an umbrella which will prevent the agent from getting

[1] See [7] for one approach. Moreover, beliefs about the future need not be believed to be inevitable, e.g. an agent may believe that it is likely that tomorrow will be a sunny day. To simplify matters, we do not consider notions related to the plausibility or probability of an event happening in the future. In this paper, an agent may be wrong and have beliefs about the future that do not correspond with the way the world will turn out to be.

wet in the first place. This decision, in turn, will result in *removing* the expectation to get wet; moreover, the agent may *add* the expectation that it will not get wet. As a consequence, the agent may now reconsider going to work by bicycling again.

Using the action rules available in the GOAL programming language, this informal example may be modelled by the following rules:

$$\textbf{if } \textbf{A-Goal}(at(work)) \wedge \neg \mathbf{B}(\Diamond wet) \textbf{ then do}(bicycle).$$
$$\textbf{if } \mathbf{B}(\Diamond wet) \vee \mathbf{B}(\Box \neg wet) \textbf{ then do}(wearUmbrella).$$

Initially, the first rule will fire and the agent will select the action $bicycle$. However, after computing expectations given this choice the agent believes that it will get wet. It will reconsider the choice to bicycle (the first rule does not fire anymore because of the second conjunct in the condition) and will decide to wear an umbrella, given the second rule. Upon recomputing expectations given the choice to wear an umbrella the agent will expect it to not get wet. As the second disjunct of the second rule represents this expectation, the rule will fire also given the presence of this expectation. Therefore, the choice to wear an umbrella is stable (the choice is not reconsidered given the new expectation) as are the expectations the agent has. As a consequence, moreover, the choice to bicycle is available again because the agent does not expect to get wet anymore. As bicycling while wearing an umbrella will not result in an expectation to get wet, the end result of performing the action $wearUmbrella$ and thereafter performing the action $bicycle$ is stable and will not be reconsidered by the agent again.[2]

As the example illustrates, expectations may lead an agent to reconsider its choices and may result in a decision to not perform an action that the agent initally committed to perform. Expectations play a role in maintaining the rational balance between beliefs, goals and action choices. It is only rational for an agent to stick to its decision to perform an action if the expectations the agent has about the effect of performing the action will be met. Reconsidering the performance of an action may introduce yet again new expectations that may also lead an agent to choose to perform other actions. The choose to perform these actions may result in changes to the agent's expectations again, however, simply because the world will look different when other actions are performed. Only when a stable state is reached in this process of decision making, reconsideration and updating of expectations an agent may be said to have reached a (final) decision to perform an action. The main issue therefore that we will consider in the remainder is how we can formally model this balance of updating of expectations and decision making, which both involves projecting the future, and how we can compute a stable state (if it exists).

[2]The subtlety here resides in the fact that we added the disjunct $\mathbf{B}(\Box \neg wet)$ to the second rule to avoid Little Nel like problems where the choice to perform an action is dropped again as the reason for selecting the action is no longer operative, as in this case the agent will no longer expect to get wet. The basic idea to avoid the Little Nel problem is to add the *expected* result of an action as a condition to the action rule that is required to choose the action. The fixed-point semantics introduced below then will automatically yield the desired result obtained by the stepwise reasoning in the text.

3 Formal Model of Expectations in GOAL

The reasons for choosing GOAL to formally model expectations are twofold. First, GOAL provides a programming framework with a rule-based decision mechanism. We need a decision-making mechanism to model the formation of expectations based on the performance of own actions and the possible reconsideration of the choice of action based on such expectations again. Second, GOAL agents derive their choice of action from their beliefs and goals. In [7], we showed how to incorporate temporal formulae in the belief and goal bases of GOAL agents. We can build on this work and extend the semantics of [7] with expectations.

Section 3.1 presents the GOAL programming framework where temporal formulae are allowed in agent's belief and goal bases as in [7]. The main difference with [7] is that we do not account for the bounded rationality of agents in this paper which motivated the work of [7]; we are also not concerned here with the origin of beliefs about the future and whether such beliefs are justified. Instead we focus here on introducing a model for the role of expectations in decision making. Section 3.2 extends this framework and introduces a formal model of expectations and their role in decision making.

3.1 Temporal GOAL

GOAL agents maintain a mental state of declarative *beliefs* and *goals* and derive their choice of action from their beliefs and goals. A GOAL agent program defines the *initial beliefs and goals* of an agent, *specifies the preconditions and effects of the actions* available to the agent, and contains a set of *action rules* to select actions for execution at runtime. Action rules define a *strategy or policy* of the agent for acting. The beliefs and goals of an agent are dynamic and change over time. The action specifications and action rules are static.

Here we use linear temporal logic (LTL) to represent beliefs and goals of agents, as in [6, 7]. One advantage of using temporal logic is that it facilitates establishing a connection between the agent programming language GOAL and agent logics, as done in [6], which in turn paves the way for verification of agent programs. Obviously, the use of temporal logic increases the expressive power compared to a framework that only allows Boolean operators as in [3]. In particular, it allows to express expectations, i.e. beliefs about the future. A base language $\mathcal{L}_0$ of classical propositional formulae over a set of atoms At, with typical element ϕ, is assumed that includes $\top, \bot \in \mathcal{L}_0$, denoting respectively the true and false sentence.

Definition 3.1 *(Linear Temporal Logic)*
The language of linear temporal logic $\mathcal{L}_{LTL}$, with typical element χ, is defined by:

$$\begin{array}{lcl} \phi & ::= & \text{any element from } \mathcal{L}_0 \\ \chi & ::= & \phi \mid \neg\chi \mid \chi \wedge \chi \mid \bigcirc\chi \mid \chi \ \mathbf{until}\ \chi \end{array}$$

The eventuality operator $\Diamond\chi$ is introduced as an abbreviation for $\top\ \mathbf{until}\ \chi$ and its dual $\Box\chi$ is defined as $\neg\Diamond\neg\chi$. We will also use $\chi\ \mathbf{before}\ \chi'$, which is defined as $\neg(\neg\chi\ \mathbf{until}\ \chi')$.

The semantics of LTL formulae is defined as usual on traces. It will be convenient here to define a trace as an infinite sequence of *states* instead of valuations. A state *state* simply is a subset of objective formulae $\mathcal{L}_0$. An objective formula ϕ then is evaluated on a trace by evaluating ϕ on the corresponding state, i.e. if t is a trace then ϕ is satisfied on t in state i, denoted by $t, i \models_{LTL} \phi$, iff $t_i \models \phi$ where t_i denotes the ith state in trace t. Temporal operators are evaluated as usual. For example, $\bigcirc\chi$ is satisfied on t in state i, i.e. $t, i \models_{LTL} \bigcirc\chi$, iff $t, i+1 \models_{LTL} \chi$. See, e.g., [4] for further details.

Rational agents need to maintain a rational balance among their beliefs and goals [2]. That is, the beliefs and goals of a rational agent need to be reasonable and ideally are justified in some way. Allowing temporal formulae as beliefs and goals raises particular issues related to maintaining such a balance. In particular, arguing that imposing particular constraints on the relation between beliefs and goals is reasonable, is much harder in a temporal setting and has been the subject of much debate. In this paper, we follow the approach proposed in [2] when it comes to defining *rationality constraints* on beliefs and goals and their relation. That is, we provide primitives for the representation of an agent's beliefs and goals, imposing only a few basic constraints at this level on the relation between beliefs and goals. Using these primitives, other variants of these mental attitudes can be defined that can be used for rational action selection.

Mental State A GOAL agent maintains a *mental state* that consists of a belief base, typically denoted by Σ, and a goal base, typically denoted by Γ, which represent the agent's current beliefs and goals. Here we assume $\Sigma \subseteq LTL$ and $\Gamma \subseteq LTL$, and require both the belief base and goal base each to be consistent. Following [2], we maintain that a rational agent should not want to change the inevitable. This is called *realism* in [2]. Informally, things that the agent believes will happen inevitably are represented by an agent's beliefs about the future, whereas a goal expresses something that an agent wants to achieve at some moment in time in the future. That is, the goals of an agent should determine a condition that is more specific than what is believed to be inevitable; since the more specific entails the less specific, an agent's goals should entail its beliefs. Goals are thus more specific than beliefs, in the sense that they add desired properties that can be influenced or controlled by the agent to the inevitable beliefs.

The following definition formally defines mental states and the accompanying rationality constraints.

Definition 3.2 *(Mental States)*
A mental state of a GOAL agent, typically denoted by m, is a pair $\langle \Sigma, \Gamma \rangle$ with $\Sigma \subseteq \mathcal{L}_{LTL}$ the belief base, and $\Gamma \subseteq \mathcal{L}_{LTL}$ the goal base. Additionally, mental states need to satisfy the following *rationality constraints*:

(i)	The belief base is consistent:	$\Sigma \not\models_{LTL} \bot$,
(ii)	The goal base is consistent:	$\Gamma \not\models_{LTL} \bot$,
(iii)	Goals refine (inevitable) beliefs:	$\Gamma \models_{LTL} \Sigma$.

Note that it follows from this definition that the belief base and goal base are also mutually consistent, i.e., $\Sigma \cup \Gamma \not\models \bot$, which means that the agent cannot have something

as a goal that is never realizable according to its beliefs.

It should be noted that the beliefs in the belief base Σ of an agent are those beliefs about the future that are independent of the performance of actions by that agent. In the terminology of [13], these beliefs are "strong" beliefs. An example is the belief $\mathbf{B}(\Box raining)$ that it will be raining, no matter what the agent will do. In this sense, the belief $\Box raining$ is inevitable as the agent has no control over the atom $raining$. In Section 3.2 another semantic component will be added to represent the expectations of agents.

As, intuitively, the future beliefs of an agent represent all the possible futures that an agent considers conceivable, it seems also reasonable to require the stronger $\Gamma \models \Sigma$, given that a goal base is consistent. This constraint expresses that the agent's goals should aim at realizing a subset of the timelines considered conceivable by the agent.[3]

Another motivation for introducing the constraint $\Gamma \models \Sigma$ relates to the interaction of disjunctive goals and beliefs. For example, a rational agent may be expected to derive the goal $\Diamond p$ from a goal $\Diamond p \vee \Diamond q$ and the belief that $\Box \neg q$. This follows immediately given the constraint.

Even though the "goals" in an agent's goal base deviate from the common sense notion of a goal in that they entail the inevitable, it should be noted that the concept of goal used here is a primitive notion and concepts more closely related to intuition can be defined (see below). It is off course true that an agent should not invest any of its own time and resources into goals that have been or will be achieved no matter what the agent will do, but we will be able to express such a constraint using the belief operator $\mathbf{B}$ introduced below by means of $\neg \mathbf{B} \varphi$.

Moreover, even though it seems reasonable to also require that (*) $\forall \chi \in LTL : \Sigma \models_{LTL} \chi \Rightarrow \Gamma \not\models_{LTL} \chi$, i.e. an agent does not have goals that it believes are inevitable, we argue that this requirement would be too strong. The reason is that this constraint prohibits an agent to believe that *part of one of its goal* has been realized. For example, consider the goal $p \wedge q$ **before** r; as it may be reasonable for an agent to first achieve p and thereafter q, in such a scenario it would be expected that the agent comes to believe p **before** r before it achieves $p \wedge q$ **before** r. Since we have that $p \wedge q$ **before** r implies p **before** r, however, by constraint (*) the agent would not be allowed to believe p **before** r, or would be required to update its goal $p \wedge q$ **before** r to q **before** r somehow. Intuitively, the latter is not what is desired since what is intended here is that the agent aims to achieve $p \wedge q$ simultaneously before r. As it is important in the agent's decision-making to be able to take into account (i.e. believe) that part of its goal has been achieved in order to focus its attention on what is left to be done, we conclude (*) is not a desirable property of a rational agent.

Mental State Conditions A GOAL agent needs the means to inspect its beliefs and goals in order to derive its choice of action from these. To do so, so-called *mental state conditions* are introduced to reason about the agent's beliefs and goals. The language $\mathcal{L}_m$ of mental state conditions extends $\mathcal{L}_{LTL}$ with a belief $\mathbf{B}$ and (primitive) goal $\mathbf{G}$

[3] We would like to note that there is a direct correspondence between the constraint $\Gamma \models \Sigma$ and the constraint $G \subseteq B$ introduced in [2] with G and B modal accessibility relations respectively modeling goals and beliefs (see also [5]). Moreover, our notion of goal as defined here is a primitive notion similar to the GOAL operator in [2]. That is, it introduces a basic motivational operator that facilitates the definition of notions of goals that are more closely related to common sense notions of goals.

operator, which can be used to express conditions on the mental state of an agent. That is, the set of mental state conditions consists of Boolean combinations of mental atoms of the form $\mathbf{B}\chi$ and $\mathbf{G}\chi$ with $\chi \in \mathcal{L}_{LTL}$.

Definition 3.3 *(Mental State Conditions: Syntax)*
The language $\mathcal{L}_m$, with typical element ψ, of *mental state conditions* is defined by:

$$\begin{array}{lcl} \chi & ::= & \text{any element in } \mathcal{L}_{LTL} \\ \psi & ::= & \mathbf{B}\chi \mid \mathbf{G}\chi \mid \neg\psi \mid \psi \wedge \psi \end{array}$$

Note that it is not allowed to nest the operators $\mathbf{B}$ and $\mathbf{G}$, nor to use temporal operators outside the scope of these operators. The semantics of mental state conditions is defined with respect to mental states.

Definition 3.4 *(Mental State Conditions: Semantics)*
Let $\langle \Sigma, \Gamma \rangle$ be a mental state. The semantics of mental state conditions is defined by:

$$\begin{array}{lll} \langle \Sigma, \Gamma \rangle \models_m \mathbf{B}\chi & \text{iff} & \Sigma \models_{LTL} \chi, \\ \langle \Sigma, \Gamma \rangle \models_m \mathbf{G}\chi & \text{iff} & \Gamma \models_{LTL} \chi, \\ \langle \Sigma, \Gamma \rangle \models_m \neg\psi & \text{iff} & \langle \Sigma, \Gamma \rangle \not\models_m \psi, \\ \langle \Sigma, \Gamma \rangle \models_m \psi \wedge \psi' & \text{iff} & \langle \Sigma, \Gamma \rangle \models_m \psi \text{ and } \langle \Sigma, \Gamma \rangle \models_m \psi'. \end{array}$$

Using the belief and primitive goal modalities $\mathbf{B}$ and $\mathbf{G}$ it is possible to *define* several common sense notions of goals. First, we define an operator $\mathbf{Goal}\chi$ by $\mathbf{G}\chi \wedge \neg\mathbf{B}\chi$, i.e., $\mathbf{Goal}\chi$ holds if χ follows from the agent's goal base, and is not believed to occur inevitably. The operator $\mathbf{Goal}\chi$ corresponds more closely to the intuitive notion of a goal as being something that the agent should put effort into bringing about. Using this operator, we can make several additional classifications of types of goals that an agent may be said to have. For example, χ is said to be an *achievement goal* whenever $\mathbf{Goal}\Diamond\chi$, and we write $\mathbf{A\text{-}Goal}\chi$.[4] Similarly, *maintenance goals* may be defined as $\mathbf{Goal}\Box\chi$.

Note that if an agent has an achievement goal $\mathbf{A\text{-}Goal}\chi$ and believes that χ always implies χ', i.e. $\mathbf{B}\Box(\chi \rightarrow \chi')$, although we have $\mathbf{G}\Diamond\chi'$, it does not follow that $\mathbf{A\text{-}Goal}\chi'$ since the agent may for instance believe that χ'. Achievement goals are particular instances of *deadline goals* of the form $\mathbf{Goal}(\chi_1 \textbf{ before } \chi_2)$ with $\chi_2 = \bot$. An agent is said to have a *bounded maintenance goal* when $\mathbf{Goal}(\chi_1 \textbf{ until } \chi_2)$ holds.

We list some of the properties of these operators, see also [7].

[4] $\mathbf{A\text{-}Goal}\chi$ thus is defined as $\mathbf{G}\Diamond\chi \wedge \neg\mathbf{B}\Diamond\chi$. We agree with [5] that it is more natural to have $\neg\mathbf{B}\chi$ than $\mathbf{B}\neg\chi$ as second conjunct, but differ in that we believe this condition should be weakened to $\neg\mathbf{B}\Diamond\chi$ to exclude the possibility that the agent believes χ inevitably will occur. We agree with [2] that it should be allowed that an achievement goal refers to the agent's environment and it is more natural to have $\mathbf{G}\Diamond\chi$ instead of $\mathbf{GB}\Diamond\chi$ in the first conjunct. Finally, it should be noted that the goal operator $\mathbf{G}$ used in the presentation of GOAL in [3] is different from the goal operator $\mathbf{G}$ introduced here; the operator $\mathbf{G}$ in [3] is best read as an achievement goal operator, an interpretation formally justified in [6].

Proposition 3.5 The following formulae are valid on mental states:

1. $\neg\mathbf{B}\bot \wedge \neg\mathbf{G}\bot$
2. $\mathbf{B}(\chi_1 \rightarrow \chi_2) \rightarrow (\mathbf{B}\chi_1 \rightarrow \mathbf{B}\chi_2)$
3. $\mathbf{B}\chi \rightarrow \mathbf{G}\chi$
4. $(\mathbf{B}(\chi_1 \textbf{ before } \chi_2) \wedge \mathbf{B}\Diamond\chi_2) \rightarrow \mathbf{B}\Diamond\chi_1$
5. $(\mathbf{G}(\chi_1 \textbf{ before } \chi_2) \wedge \mathbf{B}\Diamond\chi_2) \rightarrow \mathbf{G}\Diamond\chi_1$
6. $\mathbf{Goal}\chi \leftrightarrow (\mathbf{G}\chi \wedge \neg\mathbf{B}\chi \wedge \neg\mathbf{B}\neg\chi)$

Item 1 expresses that both beliefs and goals are consistent. Item 2 says that $\mathbf{B}$ is a normal modal operator. Item 3 implies realism (cf. [2]). Item 5 (4 is similar) expresses the following rationality with respect to goals: if the agent has a goal that χ_1 will happen before χ_2, and it indeed believes that χ_2 will sometime occur, then it has a goal that χ_1 will sometime occur. Item 6 explains why $\mathbf{Goal}$ can be considered to model goals that the agent is willing to act upon: any χ for which $\mathbf{Goal}\chi$ holds is in the agent's goal base, and not believed to be guaranteed or impossible.

Some other desirable properties follow rather straightforwardly from those above, e.g., we have that if an agent has a goal, it does not believe that the opposite is inevitable, i.e., $\mathbf{G}\chi \rightarrow \neg\mathbf{B}\neg\chi$, or, equivalently, $\mathbf{B}\neg\chi \rightarrow \neg\mathbf{G}\chi$.

The principle $\mathbf{B}\chi \rightarrow \mathbf{G}\chi$ adopted here was first proposed in [2], where it is called *realism*. In [12] the principle $\mathbf{G}\alpha \rightarrow \mathbf{B}\alpha$, called *strong realism*, was proposed, where α is assumed to express a future possibility. As the latter seems close to the "converse" of the former, we believe this has given rise to various misunderstandings. The main issue here seems to center around the acceptance of beliefs about the inevitable future as goals. The operator GOAL in [2] does entail such beliefs, which are excluded again in their defined notion of achievement goals (compare our definition above). [12] ensure their primitive goal operator GOAL does not entail such beliefs, ensuring in that way that an agent's goals do not entail things the agent believes will occur inevitably. In addition, using the machinery of CTL instead of LTL, they require an agent to *explicitly* believe in the *possibility* of realizing a goal α, that is, $\mathbf{B}E\alpha$ should hold. We conclude that, as long as there is no particular interest to have agents *explicitly* represent their belief that it is possible to achieve a goal, the differences between [2] and [12] are not so much conceptual but are more technical in nature. But also see the discussion above about the constraint $\Gamma \models \Sigma$.

Action Rules and The Semantics of Actions In order to simplify the technical presentation here, we assume a *transition function* $\mathcal{T}$ that maps an action $\mathtt{a}$ and a mental state m to a new mental state $\mathcal{T}(\mathtt{a}, m) = m'$, representing both the preconditions and effects of action execution. Note that we require action execution to lead to a mental state again, thus enforcing the agent to maintain the rationality constraints of Definition 3.2 at all times. We say that an action is *enabled* in a mental state m, denoted $m \models \mathbf{enabled}(\mathtt{a})$, if we have that $\mathcal{T}(\mathtt{a}, m)$ is defined.[5]

From the actions that are enabled in a mental state, a GOAL agent has to make a choice as to which actions it will actually perform. The basic mechanism available in GOAL that allows an agent to make this choice, is a *rule-based action selection mechanism* using so-called *action rules*. Action rules have the form $\textbf{if } \psi \textbf{ then do}(\mathtt{a})$ and are used to specify that action $\mathtt{a}$ may be selected by the agent for execution if mental state

[5] See [7] for an approach using LTL to specify preconditions and effects of actions, based on [10].

condition ψ holds; if that is the case we say that action $\mathtt{a}$ is *applicable*. If the preconditions of an applicable action also hold, i.e. the action is enabled, we say that the action is an *option*. We introduce a special predicate $\mathbf{option}(\mathtt{a})$ and write $m \models \mathbf{option}(\mathtt{a})$ to denote that $\mathtt{a}$ is an option. Formally, if $\mathbf{if}\ \psi_1\ \mathbf{then\ do}(\mathtt{a}), \ldots, \mathbf{if}\ \psi_n\ \mathbf{then\ do}(\mathtt{a})$ are all the action rules for action $\mathtt{a}$, then we have $m \models \mathbf{option}(\mathtt{a})$ iff $m \models (\psi_1 \vee \ldots \vee \psi_n) \wedge \mathbf{enabled}(\mathtt{a})$.

Action rules allow agents to derive their choice of action from their beliefs and goals in the *current mental state*. Using these rules, the agent selects an action from the set of actions that are options in the state. For example, the rule $\mathbf{if\ A\text{-}Goal}(atWork) \wedge \neg\mathbf{B}(\Diamond wet)\ \mathbf{then\ do}(bicycle)$ may be used to specify that if the agent has a goal to be at work and does not believe it will get wet, it can select the action of bicycling.[6]

The semantics of action selection and execution are formally specified by means of an operational semantics [11]. A GOAL agent non-deterministically selects a single action for execution in each state. This is formally defined in the following transition rule, which describes how an agent moves from one mental state to another.

Definition 3.6 *(Action Rule Semantics)*
Let $m = \langle \Sigma, \Gamma \rangle$ be a mental state. The labelled transition relation $\longrightarrow$ is the smallest relation induced by the following transition rule.

$$\frac{m \models \mathbf{option}(\mathtt{a})}{m \xrightarrow{\mathtt{a}} \mathcal{T}(\mathtt{a}, m)}$$

The action semantics of GOAL induces a set of possible *computations*. We define a computation as a sequence of mental states and actions, such that each mental state can be obtained from the previous by applying the transition rule of Definition 3.6. As GOAL agents are non-deterministic, the semantics of a GOAL agent is defined as the *set* of possible computations of the GOAL agent, where all computations start in the initial mental state of the agent.

Definition 3.7 *(Meaning of a* GOAL *Agent)*
A *computation* c is an infinite sequence $m_0, \mathtt{a}_0, m_1, \mathtt{a}_1, \ldots$ of mental states m_i and actions $\mathtt{a}_i$ such that $m_i \xrightarrow{\mathtt{a}_i} m_{i+1}$, or for all $\mathtt{a}$: $m_i \not\xrightarrow{\mathtt{a}}$ and $m_j = m_i$ for all $j > i$ and $a_j = \mathbf{skip}$ for all $j \geq i$.

We write c_i^m to denote the mental state at point i in c and c_i^a to denote the action performed at point i in c. The meaning $\mathcal{M}_{Agt}$ of a GOAL agent named Agt with initial mental state m_0 is the set of all computations starting in that state.

The purpose of the rule-based action selection mechanism is to allow the agent programmer to provide the agent with a means to choose actions for execution from an available set of executable actions. That is, rather than leaving it completely up to the agent to choose which actions to execute, the rules can be used to reduce the options an agent has, i.e., the rules specify when it may make sense to execute an action.

[6] In our example, the belief to get wet is an expectation derived from the actions chosen by the agent. Syntactically, however, we will not make a distinction between beliefs about the future that are independent from the agent's own action or not. That is, throughout we simply write $\mathbf{B}(\chi)$ for both types of beliefs χ. In Section 3.2 a semantics will be introduced for expectations which will be added to the mental state of an agent.

3.2 Expectations

The idea is to add expectations given the action choices of an agent and then using these expectations to reconsider the choices again until a stable state is reached. Both expectations as action choices need to be stable in this final state. That is, the expectations in that state should not give reason to reconsider the action choices in that state nor should the action choices provide reason to revise the expectations the agent has in that state. It thus is natural to define a semantics of expectations as a fixed-point construction, starting with the "strong" beliefs as a basis. The beliefs in the belief base Σ are used to bootstrap the process.

Expectations are derived from action choices. The meaning of a GOAL agent, the set of computations $\mathcal{M}_{Agt}$, is used as a starting point here. Technically, the expectations of an agent at a time point i in a computation c may be computed using the possible continuations of the initial computation up to point i. The possible continuations of a computation c from time point i on is denoted by $cont(c, i)$. These continuations depend only on the mental state at point i in c, as GOAL is a state-based formalism; we therefore also say that $cont(c, i)$ denotes the continuations of mental state c_i^m and also write $cont(m)$ where m is a mental state. The function $cont(c, i)$ is defined by:

$$cont(c, i) \quad ::= \quad \{c' \mid c' \text{ is a computation starting in mental state } c_i^m\}$$

As expectations depend on actions choosen, and vice versa, both need to be fixed simultaneously in order to reach a stable state in a decision process that takes expectations into account. It therefore is useful to introduce a function $cont(c, i, A)$ where A is a set of actions. We will also use $cont(m, A)$ below for the same reasons mentioned above. $cont(c, i, A)$ is defined as:

$$cont(c, i, A) \quad ::= \quad \{c' \mid c' \text{ is a computation starting in mental state } c_i^m \text{ with } c_0'^a \in A\}$$

In order to compute expectations, i.e. beliefs induced by performing actions chosen by the agent, we need to be able to derive which LTL formulae would be believed by the agent given that these actions are performed. We need to derive expectations from a computation, but this is not exactly what we need. We cannot simply evaluate LTL formulae on GOAL computations. We therefore extract LTL traces from computations that may be used to this end.

Definition 3.8 (*Mapping Computations to LTL Traces*)
Let c be a computation $m_0, a_0, m_1, a_1, \ldots$ where $m_i = \langle \Sigma_i, \Gamma_i \rangle$. The trace derived from c, denoted $trace(c)$, is a sequence $s_0, s_1, \ldots$ where $s_i = \{\phi \in \mathcal{L}_0 \mid \Sigma_i \models \phi\}$. The function $trace$ is lifted to sets of computations in the obvious way.

The traces resulting from the application of $trace$ to computations are LTL traces, where each state *state*$_i$ consists of the set of objective formulae that represent the current state of affairs. The idea is to derive the expectations that an agent has in a mental state from the LTL traces obtained by applying $trace$ to the continuations of that mental state. The set of expectations $E(m)$ that an agent has given a mental state m can then be defined by:

$$E(m) \quad ::= \quad \{\chi \in \mathcal{L}_{LTL} \mid \forall t \in trace(cont(m)) : t, 0 \models \chi\}$$

The next step is to add these expectations to the beliefs in the mental state again. Adding the expectations to the mental state will allow the agent to reconsider action choices it has made based upon the contents of the mental state without these expectations. That is, in the bicycling example, after adding the expectation that the agent will get wet, the agent may reconsider choosing the action of bicycling and select an alternative action, e.g. to wear an umbrella, instead. The idea is to update a mental state $m = \langle \Sigma, \Gamma \rangle$ simply by adding $E(m)$ to the belief base Σ, i.e. we get $m_E = \langle \Sigma \cup E(m), \Gamma \cup E(m) \rangle$.[7] We introduce a function $expect$ that updates a mental state $m = \langle \Sigma, \Gamma \rangle$ with the expectations induced by that mental state, i.e. $expect(\langle \Sigma, \Gamma \rangle) = \langle \Sigma \cup E(m), \Gamma \cup E(m) \rangle$.

This new mental state, obtained by adding expectations, then may be used to reconsider choices of actions, i.e. $m_E = expect(m)$ is used to recompute the actions the agent will perform. It is obvious that any changes to the choice of action may change the expectations again, and we cannot simply stop the decision and expectation process after a single step but need to continue this process until a stable state will be reached. That is, we need to compute $expect^{n+1}(m) = expect(expect^n(m))$, where $expect^1(m) = expect(m)$, until $expect^{n+1}(m) = expect^n(m)$ for some n. Such a stable state does not have to exist, but if it does, it is reached after a finite number of steps n. We therefore introduce a fixed point operator F to formalize this process, i.e. we define:

$$F(m) \quad ::= \quad \begin{cases} expect^{n+1}(m) & \text{for the least } n > 0 \text{ s.t. } expect^{n+1}(m) = expect^n, \\ & \text{if such an } n \text{ exists,} \\ \text{undefined} & \text{otherwise.} \end{cases}$$

Note that without the existence of an "expectation fixed point", i.e. the case that $F(m)$ is undefined for mental state m, an agent will not reach a stable state where decisions and expectations are fixed, and no decision is reached at all.[8]

Using the operator F, we can extend the reconsideration process using expectations from mental states to compute complete computations that take expectations into account. The idea is simply to first compute a fixed point for the initial mental state, given this fixed point compute a fixed point for the second state, and repeat this process for later points.

That is, given an initial mental state m_0 of an agent, we want to obtain computations that take expectations into account. To this end, we introduce an operator G and

[7] Note that we also add the expectations to the goal base in order to obtain a mental state again, i.e. to ensure that the rationality constraint $\Gamma_E \models \Sigma_E$ of Definition 3.2 is satisfied in the state m_E, which is trivially satisfied by taking $\Gamma_E = \Gamma \cup E(m)$ and $\Sigma_E = \Sigma \cup E(m)$ (as we already must have $\Gamma \models \Sigma$). Note that by adding expectations in this way we do not introduce new achievement goals χ, as such goals require that the agent does not believe χ. We should also make sure that adding expectations do not make the belief or goal base inconsistent; we simplify and do not provide a detailed account here, but stipulate that if inconsistency results $expect(m)$ is undefined.

[8] It may be beneficial to alternatively set $F(m) = m$ if no fixed point exists, although this is not completely clear. In that case, a decision mechanism is needed to establish the non-existence of a fixed point.

use it to produce computations c as follows:[9]

$$\begin{array}{rcl} G(m) & ::= & \{F(m), \mathtt{a}, c \mid F(m) \xrightarrow{\mathtt{a}} m', c \in G(m')\} \cup \\ & & \{F(m), \mathbf{skip}, F(m), \mathbf{skip}, \ldots \mid F(m) \not\xrightarrow{\mathtt{a}}\} \end{array}$$

The meaning of a GOAL agent with initial state m_0 that takes expectations into account in its action selection or decision making process may now be defined as $G(m_0)$, if a least fixed point exists, and as undefined otherwise.

```
;;
main example
{
    :beliefs{
      □(raining).
      ¬ umbrella.
      at(home).
      at(home) → ¬ at(work). at(work) → ¬ at(home).
      distance(work, 20).
      outside ∧ raining ∧ ¬ umbrella → wet.
      umbrella ∨ at(home) ∨ at(work) → ¬ wet.
    }
    :goals{
      ◇(at(work)).
      ...
    }
    :program{
      if A-Goal(at(work)) ∧ ¬B(◇(wet)) then do(bicycle).
      if B(◇(wet)) ∨ B(□(¬ wet)) then do(wearUmbrella).
    }
    :action-spec{
      bicycle {
          :pre{ distance(work, X), X > 0 }
          :post{ outside, ¬ distance(work, X) ∧ distance(work, X − 1) }
      bicycle {
          :pre{ distance(work, 0) }
          :post{ ¬ outside ∧ at(work) }
      wearUmbrella {
          :pre{ ¬ umbrella }
          :post{ umbrella }
      }
    }
}
```

Table 1: GOAL Agent Program Example

Remark 3.1 The semantics of expectations and their role in decision making defined here may provide a setup that is a simplification of what is actually needed. The point is that the semantics now fixes the initial action first and assumes this action will never need to be reconsidered again based upon reconsiderations of choices to perform actions at a later time, but we do not discuss such complications here any further.

Table 1 is a GOAL program that implements the example discussed in Section 2 in more detail. The dots ... in the goal base indicate the missing formulae needed to ensure the initial mental state satisifies the constraint $\Gamma \models_{LTL} \Sigma$, where Γ represents the content of the initial goal base and Σ represents the content of the initial belief base.

[9] $G(m)$ is the least fixed point.

We have taken the liberty to use some variables here to be able to concisely represent the agent program. The action $bicycle$ is specified twice in the action specification section to represent the different effects of performing the action given different distances to work.

It is not difficult to verify that in the initial mental state m_0 of the example agent the only option is to perform $bicycle$, until the distance to work is 0; in the latter case, the agent has achieved its goal to be at work and $\mathbf{A\text{-}Goal}(at(work))$ will no longer be true. The bicycling action also inserts $outside$ into the belief base of the agent.[10] Using the semantics for expectations introduced above, we then are able to derive the expectation that the agent will get wet sometime, assuming that $\Box(raining)$ will progress to the next state as $\Box(raining)$ (cf. [7]). That is, we have $E(m_0) \models \Diamond(wet)$ since we have $outside \wedge \Box(raining) \wedge \neg umbrella$ in the next state. Using the $expect$ function, we then obtain that a mental state where $\Diamond(wet)$ is believed. This expectation leads the agent to reconsider the action $bicycle$. It is easy to see that the action $bicycle$ is no longer an option in this state, as the corresponding action rule requires $\neg\mathbf{B}(\Diamond(wet))$. Instead, the action $wearUmbrella$ now becomes an option. As a result, $umbrella$ is added to the mental state which implies $\neg wet$ in future states (as $umbrella$ is assumed to persist). Using the function E again, we can compute a new expectation that $\Box\neg wet$. This state is stable as the action $wearUmbrella$ does not need to be reconsidered again. In subsequent states, the agent then can perform $bicycle$ to arrive at work.

4 Conclusion

The formal model of expectations introduced here provides a model that is able to reproduce the natural interplay between expectations and action choices. That is, our model explains the role of expectations in decision making, as illustrated by the simple example presented in Section 2.

Various extensions of the work presented remain future work. In particular, it will be interesting to investigate the combination of the action theory integrated into GOAL in [7] with the model of expectations introduced here. The semantics for expectations is an extension of the semantics for GOAL. However, this does not mean the semantics proposed is computational. More work is needed to investigate its computational properties. Other extensions that seem interesting involve introducing notions of likelihood or probability into the model.

I would like to conclude with some more personal remarks. As a PhD in Utrecht, I have very much enjoyed working with Wiebe. Wiebe has an eye for detail that has always struck me and that, to my advantage, has helped improve my PhD work. After I returned to Academia, Wiebe had moved to Liverpool but I am very glad we managed to join forces again after some time. This work would not have been written without various interesting discussions with Wiebe, and I expect to have many more fruitful discussions with Wiebe in the future. Congratulations with your 50th birthday!

[10] We assume such basic facts are persistent, i.e. a fact ϕ will only be removed when an action is performed with a postcondition $\neg\phi$, i.e. we assume a STRIPS-style semantics of actions here, see [8]).

References

[1] Cristiano Castelfranchi and Emiliano Lorini. Cognitive anatomy and functions of expectations. In *Proceedings of the IJCAI'03 Workshop on Cognitive Modeling of Agents and Multi-Agent Interactions*, 2003.

[2] Philip R. Cohen and Hector J. Levesque. Intention Is Choice with Commitment. *Artificial Intelligence*, 42:213–261, 1990.

[3] Frank de Boer, Koen V. Hindriks, Wiebe van der Hoek, and John-Jules Ch. Meyer. A Verification Framework for Agent Programming with Declarative Goals. *Journal of Applied Logic*, 5(2):277–302, 2007.

[4] E. Allen Emerson. Temporal and Modal Mogic. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science*, volume B. North-Holland Publishing Company, Amsterdam, 1990.

[5] Andreas Herzig and Dominique Longin. C&L Intention Revisited. In *Proc. of the 9th Int. Conference Principles of Knowledge Representation and Reasoning (KR'04)*, pages 527–535, 2004.

[6] Koen V. Hindriks and Wiebe van der Hoek. Goal agents instantiate intention logic. In *JELIA'08*. 2003, 2008.

[7] Koen V. Hindriks, M. Birna van Riemsdijk, and Wiebe van der Hoek. Agent programming with temporally extended goals. In *Proc. of the Eighth International Conference on Autonomous Agents and Multiagent Systems (AAMAS'09)*, 2009.

[8] V. Lifschitz. On the semantics of STRIPS. In M.P. Georgeff and A.L. Lansky, editors, *Reasoning about Actions and Plans*, pages 1–9. Morgan Kaufman, 1986.

[9] Emiliano Lorini and Rino Falcone. Modeling expectations in cognitive agents. In *AAAI'05 Fall Symposium - From Reactive to Anticipatory Cognitive Embodied Systems*, 2005.

[10] Marta Cialdea Mayer, Carla Limongelli, Andrea Orlandini, and Valentina Poggioni. Linear temporal logic as an executable semantics for planning languages. *Journal of Logic, Lang and Information*, 16, 2007.

[11] Gordon D. Plotkin. A Structural Approach to Operational Semantics. Technical Report DAIMI FN-19, University of Aarhus, 1981.

[12] Anand S. Rao and Michael P. Georgeff. Intentions and Rational Commitment. Technical report, Australian Artificial Intelligence Institute, 1993.

[13] Wiebe van der Hoek, Wojciech Jamroga, and Michael Wooldridge. Towards a Theory of Intention Revision. *Synthese*, 155:265–290, 2007.

Cheers Wiebe!