T·L·G
Texts in Logic and Games
Volume 4

New Perspectives on Games and Interaction

EDITED BY
KRZYSZTOF R. APT AND ROBERT VAN ROOIJ

AMSTERDAM UNIVERSITY PRESS

NEW PERSPECTIVES ON GAMES AND INTERACTION

T·L·G
Texts in Logic and Games
Volume 4

New Perspectives on Games and Interaction

EDITED BY
KRZYSZTOF R. APT
ROBERT VAN ROOIJ

Texts in Logic and Games
Volume 4

AMSTERDAM UNIVERSITY PRESS

Cover design: Maedium, Utrecht

ISBN 978 90 8964 057 4
e-ISBN 978 90 4850 642 2
NUR 918

Table of Contents

Preface

In the period February 5-7, 2007 we organized in the historie building of the Dutch Royal Academy of Sciences (KN AW) in Amsterdam the Academy Colloquium titled "New Perspectives on Games and Interaction" . The program consisted of 14 invited lectures, each followed by a commentary, and 8 contributed talks.

Traditionally, logic and linguistics have been studied from a statie and non-interactive point of view emphasizing the structure of proofs and meanings. **In** computer science, dynamic processing of information has always played a major role, but from a non-interactive machine perspective. More recently, the dynamic and interactive aspects of logical reasoning, communication, and information processing have been much more central in the three above-mentioned disciplines.

Interaction is also of crucial importance in economics. Mathematical game theory, as launched by Von Neumarm and Morgenstern in 1944 in their seminal book, followed by the contributions of Nash, has become a standard tooi in economics for the study and description of various economie processes, including competition, cooperation, collusion, strategie behaviour and bargaining.

These different uses of games in logic, computer science, linguistics and economics have largely developed in isolation. The purpose of the workshop was to bring together the researchers in these areas to encourage interactions between these disciplines, to clarify their uses of the concepts of game theory and to identify promising new directions.

This volume consists of the contributions written by the speakers. It testifies to the growing importance of game theory as a tooi to capture the concepts of strategy, interaction, argumentation, communication, cooperation and competition. We hope that the reader will find in the papers presented in this volume useful evidence for the richness of game theory and for its impressive and growing scope of use.

We take this opportunity to thank Benedikt Löwe and Johan van Benthem for their cooperation in the preparations of the scientific programme of the Colloquium.

Amsterdam K.R.A. R.A.M.v.R.

The Logic of Conditional Doxastic Actions

Alexandru Baltag!
Sonja Smets[2,3]

[1] Computing Laboratory
Oxford University
Oxford OX1 3QD, United Kingdom

[2] Center for Logic and Philosophy of Scîence
Vrije Universiteit Brussel
Brussels, 1050, Belgium

[3] lEG Research Group on the Philosophy of lnformation
Oxford University
Oxford OX1 3QD, United Kingdom
ba1tag@corn1ab.ox.ac.uk, sonsrnets@vub.ac.be

Abstract

We present a logic of conditional doxastic actions, obtained by incorporating ideas from belief revision theory into the usual dynamic logic of epistemic actions. We do this by reformulating the setting of *Action-Priority Update* (see Baltag and Smets, 2008) in terrns of *eonditional doxastie models*, and using this to derive *general reduction laws for conditional beliefs* after arbitrary actions.

1 Introduction

This work is part of the on-going trend (see Aueher, 2003; van Benthem, 2004; van Ditmarseh, 2005; Baltag and Sadrzadeh, 2006; Baltag and Smets, 2006a,b,e, 2007a,b, 2008) towards ineorporating belief revision meehanisms within the Dynamie-Epistemie Logie (DEL) approach to information update. As sueh, this paper ean be eonsidered a sequel to our recent work (Baltag and Smets, 2008), and it is based on a revised and improved version of our older unpublished paper (Baltag and Smets, 2006e), presented at the 2006 ESSLLI Workshop on "Rationality and Knowledge".

We assume the general distinetion, made by van Ditmarseh (2005), Baltag and Smets (2006a) and van Benthem (2004), between *"dynamie" and "statie" belief revision.* In this sense, classieal AGM theory in (Alehourrón et al., 1985) and (Gärdenfors, 1988) (and embodied in our setting by the *eonditional belief* operators $B_a^P Q$) is "statie", eapturing *the aqent's ehanging beliefs about an unehanging uiorld.* As sueh, "statie" belief revision eannot be self-referential: statieally-revised beliefs eannot refer to themselves, but

Krzysztof R. Apt, Robert van Rooij (eds.). *New Perspectives on Games and Interaction.* Texts in Logic and Games 4, Amsterdam University Press 2008, pp. 9-31.

only to the original, unrevised beliefs. **In** contrast, "dynamic" belief revision deals with the agent's revised beliefs about the world *as it* is *afier revision* (including the revised beliefs themselves).

In (Baltag and Smets, 2006a), we proposed *two equivalent semantie settings* for "statie" belief revision *(eonditional doxastie models* and *epistemie plausibility modeis),* and proved them to be equivalent with eaeh ot her and with a multi-agent epistemie version of the AGM belief revision theory. We argued that these settings provided the "right" qualitative semanties for multi-agent belief revision, forrning the basis of a *eonditional doxastie logie* (CDL, for short), that eaptured the main "laws" of statie belief revision using *eonditional-belief* operators $B_a^P Q$ and *knowledge* operators *KaP*. The later correspond to the standard S5-notion of "knowiedge" (partition-based and fuliy introspeetive), that is eommonly used in Computer Seienee and Game Theory. **In** the same paper, we went beyond statie revision, using CDL to explore a restrieted notion of "dynamic" belief revision, by modeling and axiomatizing multi-agent belief updates indueed by *publie and private annouueements.*

In subsequent papers, eulminating in (Baltag and Smets, 2008), we extended this logie with a "safe belief' modality *DaP,* eapturing a form of "weak (non-introspeetive) knowiedge" , first introdueed by Stalnaker in his modal formalization (Stalnaker, 1996, 2006) of Lehrer's *defeasibility analysis of knowledge* (Lehrer, 1990; Lehrer and Paxson, 1969). We went on to deal with "dynamic" multi-agent belief revision, by developing a notion of *doxastie actions",* general enough to cover most examples of muiti-agent eommunieation actions eneountered in the literature, but also ftexible enough Lo *implement various "belief-revision policies" in a unified setting.* Foliowing Aueher (2003) and van Ditmarseh (2005), we represented doxastie actions using *(epistemic) action plausibility models.* The underlying idea, originating in (Baltag and Moss, 2004) and (Baltag et al., 1998), was to use *the same type of formalism that was used to model "statie" beliefs: epistemiejdoxastie actions should be modeled in essentially the same way as epistemiejdoxastie states.* The main differenee between our proposal and the proposals of Aueher (2003), van Ditmarseh (2005) and Baltag et al. (1998) lies in *our different notion of "update product"* of a state model with an action model: our "Aetion-Priority Update" was based on taking the *anti-lexieogmphie order* on the Cartesian product of the state model with the action model. This is a *naturel genemlization of the AGM-type belief revision* to complex multi-agent belief-ehanging aetions: foliowing the AGM tradition, it gives *priority to ineoming information* (i.e., to "aetions" in our sense). **In** the same paper (Baltag and Smets, 2008), we eompletely axiomatized the generallogie of dynamie belief revision, using Reduetion Axioms

1 Or "doxastic events", in the terminology of van Benthem (2004).

for knowledge and safe belief aftel' arbitrary doxastic actions.

In this paper, we go further to look at representations of doxastic actions in terrns of our other (equivalent) semantic setting for belief revision mentioned above (conditional doxastic modeis). We look in detail at an equivalent statement for the (same) notion of Action-Priority Update in terrns of conditional doxastic actions. This is in itself a non-trivial, rather intricate exercise, which as a side benefit gives us Reduction Axioms for conditional belief aftel' arbitrary actions. Though more complex than the Reduction Axioms for knowledge and safe belief in (Baltag and Smets, 2008) (and in principle derivable from these"), the axioms of the resulting Logic of Conditional Doxastic Actions are of more direct relevanee to belief revision and belief update, and are immediately applicable to deriving reduction laws for interesting special cases, such as the ones considered by van Benthem (2004).

In its spirit, our approach is closer to the one taken by .J. van Benthem and his collaborators (van Benthem, 2004; van Benthem and Roy, 2005; van Benthem and Liu, 2004) (based on *qualitative logies of eonditional belief, "preference" modalities and oarious [orms of "belief upgrade"*), rather than to the approaches of a more "quantitative" ftavor due to Aucher (2003) and van Ditmarsch (2005) (based on formalizing Spohn's ordinal degrees of beliefs (1988) as "graded belief' operators, and proposing various quantitative arithmetic formulas for updates). As aresuit, the "reduction axioms" by van Benthem (2004) (for "hard" public announcements, lexicographic upgrades and conservative upgrades) can be recovered as special cases of our main reduction axiom for conditional beliefs aftel' an arbitrary action.

Our conditional belief modalit.ies and our condit.ional doxast.ic rnodels can also be seen in the context of the wide logical-philosophical literature on notions of *eonditional* (see, e.g., Adams, 1965; Stalnaker, 1968; Ramsey, 1931; Lewis, 1973; Bennett, 2003). One can of course look at conditional belief operators as non-classical (and non-monotonic!) implications. Our approach can thus be compared with other attempts of using doxastic conditionals to deal with belief revision, (see, e.g., Gärdenfors, 1986; Ramsey, 1931; Grove, 1988; Rott, 1989; Fuhrmann and Levi, 1994; Ryan and Schobbens, 1997; Halpern, 2003; Friedmann and Halpern, 1994). As shown in (Baltag and Smets, 2006a), our operators avoid the known paradoxes'' arising from such mixtures of conditional and belief revision, by *failing to satisfy the so-ealled Ramsey test.*

2 Together with the axioms of the logic of knowledge and safe belief, and with the definition in (Baltag and Smets, 2008) of conditional belief in terms of knowledge and safe belief.

3 See e.g., (Stalnaker, 1968), (Gärdenfors, 1988) and (Rott, 1989).

2 Preliminaries: Epistemie plausibility models and conditional doxastic models

In this section, we review some basic notions and results from (Baltag and Smets,2ÜÜ6a).

Plausibility frames. An *epistemic plausibility frame* is a structure S = $(S, <\!\!«, \leq_a)_{a\in\mathcal{A}}$, consisting of a set S, endowed with a family of equivalence relations $<\!\!«$, called *epistemic indistinguishability relaiions*, and a family of "well-preorders" $\leq_a$, called *plausibility relations.* Here, a "well-preorder" is just a preorder" such that *every non-empty subset has minimal elements.*[5] Using the notation $\mathrm{Min}_{\leq} T := \{t \in T : t \leq ti \text{ for all } ti \in T\}$ for the set of minimal elements of T, the last condition says that: for every $T \subseteq S$, if $T \neq 0$ then $\mathrm{Min}_{\leq} T \neq o$.

Plausibility frames for only one agent and without the epistemic relations have been used as models for conditionals and belief revision by Grove (1988), Gärdenfors (1986, 1988), Segerberg (1998), etc. Observe that the conditions on the preorder $\leq_a$ are (equivalent to) Grove's conditions for the (relational version of) his models (Grove, 1988). The standard formulation of Grove modeIs (in terms of a "system of spheres", weakening the similar notion in (Lewis, 1973)) was proved by Grove (1988) to be equivalent to the above relational forrnulation."

Given a plausibility frame S, an *S-pmposition* is any subset $P \subseteq S$. We say that *the state* s *satisfies the pmposition P* if s ∈ *P*. Observe that a plausibility frame is just a special case of a *Kripke frame.* So, as is standard for Kripke frames, we can define an *epistemic plausibility model* to be an epistemic plausibility frame S together with a valuation map $\| \cdot \| : \Phi \to P(S)$, mapping every element of a given set Φ of "atomie sentences" into S-propositions.

Notation: strict plausibility, doxastic indistinguishability. As with any preorder, the ("non-strict") plausibility relation $\leq_a$ above has a "strict" (i.e., asymmetrie) version $<a$, as weil as a cortesponding equivalence relat ion $\simeq_a$, called "doxastic indistinguishability":

$$s <a\ t \text{ iff } s \leq_a \text{tand } t \not\leq_a s$$
$$s \simeq_a t \text{ iff } s \leq_a \text{tand } t \leq_a s$$

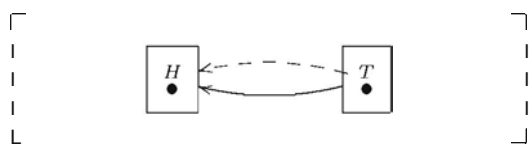
4 I.e., a reftexive and transitive relation.

5 Observe that the existence of minimal elements implies, by itself, that the relation $\leq_a$ is both *reflexive* (i.e., s $\leq_a$ s for all s ∈ *S)* and *connected* (i.e., either s $\leq_a$ *tor* t $\leq_a$ s, for all s, $t \in SJ$, i.e., elements that are below all the others. Note also that, when the set *Sisfinite,* a well-preorder is nothing but a connected preorder.

6 A more *concrete example* ofplausibility frames was given by Spohn (1988), in terms of ordinal plausibility maps assigning ordinals *d(s)* ("the degree of plausibility" of s) to each state s ∈ *S*. In our epistemic multi-agent context, this would endow each agent *a* with an ordinal plausibility map $d_a : S \to Ord$.

Interpretation. The elements of S will be interpreted as the *possible states* of a system (or "possible worlds"). The atomie sentences $p \in \Phi$ represent *"oniic" (non-doxastic) facts* about the world, that might hold or not in a given state, while the valuation tells us which facts hold at which worlds. The equivalence relations $\sim_a$ capture *the aqerü's knowledge about the actual state of the system* (intuitively based on the agent's *{pariial] observaiions of this state):* two states s, *t* are *indistinguishable fOT agent a* if $s \sim_a t$. **In** other words, when the actual state of the system is s, then agent *a* knows only the state's equivalence class $s(a) := \{t \in S : s \sim_a t\}$. Finally, the plausibility relations $\leq_a$ capture *the aqent's conditional beliefs about (virtual) states of the system:* given the information that some possible state of the system is either s or *t,* agent *a* will believe the state to be s iff $s <_a t$; will believe the state to be *t* iff $t <_a s$; otherwise (if $s \simeq_a t$), the agent will consider the two alternatives as equally plausible.

Example 1. The father informs the two children (Alice and Bob) that he has put a coin Iying face up on the table in front of them. At first, the face is covered (so the children cannot see it). Based on previous experience, (it is common knowledge that) the children believe that the upper face is (very likely to be) Heads: say, they know that the father has astrong preferenee for Heads. And in fact, they're right: the coin lies Heads up. Next, the fat her shows the face of the coin to Alice, in the presence of Bob but in such a way that Bob cannot see the face (though of course he can see that Alice sees the face). The plausibility model S for this situation is:



Here, we left the father out of the picture (sinee he only plays the role of God or Nature, not the role of an uncertain agent). The node on the left, labeled with *H,* represents the actual state of the system (in which the coin lies Heads up), while the node on the right represents the other possible state (in which the coin is Tails up). We use continuous arrows to encode Alice's beliefs and use continuous squares to eneode her knowiedge, while using dashed arrows and dashed squares for Bob. More precisely: the squares represent the agents' information cells, i.e., the equivalence classes $s(a) := \{t \in S : s \sim_a t\}$ of indistinguishable states (for each agent *a).* Observe that Alice's information cells (the continuous squares) are singletons: in every case, she *knows* the state of the system; Bob's information cell is one big dashed square comprising both states: he doesn't know which state is the realone, so he cannot distinguish between them. The arrows represent the

plausibility relaiions for the two agents; since these are always refiexive, we choose to *skip all the loops* for convenience. Both arrows point to the node on the left: *a priori* (i.e., before making any observation of the real state), both agents believe that it is likely that the coin lies Heads up.

Conditional doxastic frames. A plausibility frame is in fact nothing but a way to encode al! the agents' possible *conditional beliefs.* To see this, consider the fol!owing equivalent notion, introduced in (Baltag and Smets, 2üü6a): A *conditional doxastic frame (CD-frame, [or short}* $S = (S, \{\bullet_a^P\}_{a\in\mathcal{A}, P}$ *c;s)* consists of a set of states S, together with a family of *conditional (doxastic) eppearanee* maps, one for each agent a and each possible condition $P \subseteq S$. These are required to satisfy the fol!owing conditions:

1. if $S \in P$ then $s_a^P \neq \emptyset$;
2. if $P \cap s_a^Q \neq \emptyset$ then $s_a^P \neq \emptyset$;
3. if $t \in s_a^P$ then $s_a^Q = t_a^Q$;
4. $s_a^P < P$;
5. $s_a^{P\cap Q} = s_a^P \cap Q$, if $s_a^P \cap Q \neq \emptyset$.

A *conditional doxastic model* (CDM, for short) is a Kripke model whose underlying frame is a CD-frame. The conditional appearance s_a^P captures *the way a state sappears to an agent a, given some additional (plausible, but not necessarilsj tTUthful) information P.* More precisely: whenever s is the current state of the world, then aftel' receiving new information P, agent a wil! come to believe that any of the states $s_i \in s_a^P$ might have been the cutrent state of the world (as it was before receiving information P).

Using conditional doxastic appearance, the *knowledge s(a) possessed by agent a about state* s (i.e., the *epistemic eppearanee* of s) can be defined as the *union of all conditional doxastic appeamnces.* **In** other words, *something* is *known ijJ it* is *believed in any conditions: s(a) := UQc;s* s_a^Q. Using this, we can see that the first condition above in the definition of conditional doxastic frames captures the *truthfulness of knowiedge.* Condition 2 states the *success of belief reoision, when consistent with knowiedge: if something* is *not known to be false, then it can be consistently eniertained as a hypothesis.* Condition 3 expresses *full introspeetion of (conditional) beliefs:* agents *know* their own conditional beliefs, so they cannot revise their beliefs about them. Condition 4 says *hypotheses are hypothetically believed:* when making a hypothesis, that hypothesis is taken to be true. Condition 5 describes *minimality of revision:* when faced with new information Q, agents keep as much as possible of their previous (conditional) beliefs s_a^P.

To recover the usual, unconditional beliefs, we put *sa* := s_a^S. **In** other words: *unconditional ("default") beliefs ar-e beliefs conditionalized by tr-ivially irue conditions.*

For any agent *a* and any S-proposition *P,* we can define a *conditional belief operator* B_a^P : P(S) → P(S) S-propositions, as the Galois dual of conditional doxastic appearance:

$$B_a^P Q := \{s \in S : s_a^P \subseteq Q\}$$

We read this as saying that *agent a believes Q given P.* More precisely, this says that: *if the agent would learn P, then (ajter- leaming) he would come to believe that Q was the case in the curreni state [bejore the leaming).* The *usual (unconditional) belief operator* can be obtained as a special case, by conditionalizing with the trivially true proposition *S:* *BaQ* := $B_a^S Q$. The *knowledge operator* can similarly be defined as the Galois dual of epistemic appearance:

$$KaP := \{s \in S : s(a) < P\}.$$

As a consequence of the above postulates, we have the following:

$$K_a P = \bigcap_{Q \subseteq S} BQP = B_a^{\neg P} \varnothing = B_a^{\neg P} P$$

Equivalence between plausibility models and conditional doxastic modeis. *Any plausibility model gives rise to a* CDM, *in a canonical way,* by putting

$$s_a^P := \mathrm{Min}_{\leq_a} \{t \in P : t \sim s\}$$

where Min., $T := \{t \in T : t \leq_a ti$ for all $ti \in T\}$ is the set of $\leq_a$-minimal elements in *T.* We call this *the canonical* CDM associated to the plausibility mode!. The converse is given by a:

Theorem 2.1 (Representation Theorem). Every CDM is the canonieal CDM of some plausibility model."

The advantage of the CDM formulation is that it leads naturally to a complete axiomatization of a *logic of conditional beliefs,* which was introduced in (Baltag and Smets, 2ÜÜ6a) under the name of "Conditional Doxastic Logic" (CDL)8: the semantical postulates that define CDM's can be immediately converted into modal axioms governing conditional belief.

---

7 This result can be seen as an analogue in our semantic context of Gä.rdenfors' representation theorem (Gä.rdenfors, 1986), representing the AGM revision operator in terrns of the minimal valuat.ions for some total preorder on valuations.

8 COL is an extension of the well-known logic KL of "knowledge and belief'; see e.g., (Meyer and van der Hoek, 1995, p. 94), for a complete proof system for KL.

Conditional Doxastic Logic (CDL). The syntax of CDL (without common knowledge and common belief operators)? is:

$$\varphi := p \mid \neg\varphi \mid \varphi \wedge \varphi \mid B_a^{\varphi}\varphi$$

while the semantics is given by the obvious compositional clauses for the interpretation map $\|\cdot\|_S : \text{CDL} \to P(S)$ in a CDM (and so, in particular, in a plausibility model) S. In this logic, the *knowledge modality* can be defined as an abbreviation, putting $K_a\varphi := B_a^{\neg\varphi}\bot$ (where $\bot = p \wedge \neg p$ is an inconsistent sentence), or equivalently $K_a\varphi := B_a^{\neg\varphi}\varphi$. This way of defining knowledge in terms of doxastic conditionals can be traced back to Stalnaker (1968). It is easy to see that this agrees semantically with the previous definition of the semantic knowledge operator (as the Galois dual of epistemic appearance): $\|K_a\varphi\|_S = K_a\|\varphi\|_S$.

Doxastic propositions. A *doxastic pmposition* is a map P assigning to each plausibility model (or conditional doxastic model) S some S-proposition, i.e., a set of states $P_S \subseteq S$. The interpretation map for the logic CDL can thus be thought of as associating to each sentence φ of CDL a doxastic proposition $\|\varphi\|$. We denote by *Prop* the family of all doxastic propositions. All the above operators (Boolean operators as weil as doxastic and epistemic modalities) on S-propositions induce corresponding operators on doxastic propositions, defined pointwise: e.g., for any doxastic proposition P, one can define the proposition *KaP*, by putting $(K_aP)_S := K_aP_S$, for all models S.

Theorem 2.2 (Baltag and Smets 2006a). A complete proof system for CDL can be obtained from any complete axiomatization of propositional logic by adding the following:

Necessitation Rule:	From $\vdash p$ infer $\vdash B_a^{\psi}p$
Normality:	$\vdash B_a^{\vartheta}(\varphi \to \psi) \to (B_a^{\vartheta}p \to B_a^{\vartheta}\psi)$
Truthfulness of Knowiedge:	$\vdash K_a p \to p$
Persistenee of Knowiedge:	$\vdash K_a\varphi \to B_a^{\psi}p$
Pull Introspection:	$\vdash B_a^{\psi}\varphi \to K_aB_a^{\psi}\varphi$
	$\vdash \neg B_a^{\psi}\varphi \to K_a\neg B_a^{\psi}\varphi$
Hypotheses are (hypothetically) accepted:	$\vdash B_a^{\varphi}p$
Minimality of revision:	$\vdash \neg B_a^{\varphi}\neg\psi \to (B_a^{\varphi\wedge\psi}\vartheta \leftrightarrow B_a^{\varphi}(\psi \to \vartheta))$

Proo]. The proof is essentially the same as of (Board, 2002). It is easy to see that the proof system above is equivalent to Board's strongest logic of (Board, 2002) (the one that includes axiom for full introspection), and that our modeIs are equivalent to the "full introspective" version of the semantics of (Board, 2002). Q.E.D.

9 In (Baltag and Smets, 2006a), we present and axiomatize a logic that includes conditional common knowledge and conditional common true belief.

3 Action plausibility models and product update

The belief revision encoded in the models above is of a *static,* purely *hypothetical,* nature. Indeed, the revision operators cannot alter the models in any way: all the possibilities are already there, so both the unconditional and the revised, conditional beliefs *refer to the same uiorld and the same moment in time.* In contrast, a *belief update* in our sense is adynamic form of belief revision, meant to capture the actual change of beliefs induced by learning (or by other forms of epistemicjdoxastic actions).10 As already noticed before, by e.g., Gerbrandy (1999) and Baltag et al. (1998), the original model does not usually include enough states to capture all the epistemic possibilities that arise in this way. So we now introduce "revisions" that change the original plausibility model.

To do this, we adapt an idea coming from (Baltag et al., 1998) and developed in full formal detail in (Baltag and Moss, 2004). There, the idea was that *epistemic actions should be modeled in essentially the same way as epistemic states,* and this common setting was taken to be given by *epistemic Kripke models.* Since we now enriched our modeIs for states to deal with conditional beliefs, it is natural to follow (Baltag and Moss, 2004) into extending the similarity between actions and states to this conditional setting, thus obtaining *action plausibility models.*

An *action plausibility model* is just an epistemic plausibility frame $\mathbf{\Sigma} = (\Sigma, <\!\!\cdot, \leq_a)_{a\in\mathcal{A}}$, together with a *precorulition map* $\mathrm{pre} : \Sigma \to FTOp$ associating to each element of Σ some doxastic proposition $\mathrm{pre}(\sigma)$. As in (Baltag and Moss, 2004), we call the elements of Σ *(basic) epistemic actions,* and we call $\mathrm{pre}(\sigma)$ *the precorulition of action* σ.

Interpretation: Beliefs about changes encode changes of beliefs. The name "doxastic actions" might be a bit misleading; the elements of a plausibility model are not intended to represent "reai" actions in all their complexity, but only the *doxastic changes* induced by these actions: each of the nodes of the graph represents a *specific kind of change of beliefs (of all the agents).* As in (Baltag and Moss, 2004), we only deal here with pure "belief changes", i.e., actions that do not change the "ontic" facts of the world, but only the agents' beliefs.l' iVIoreover, we think of these as *deierminisiic* changes: there is at most one output of applying an action to a state.[12] Intuitively, the precondition defines the *domain of applicability* of

10 But observe the difference between our notion of belief update (originating in dynamic-epistemic logic) and the similar (and vaguer) notion in (Katsuno and Mendelzon, 1992).

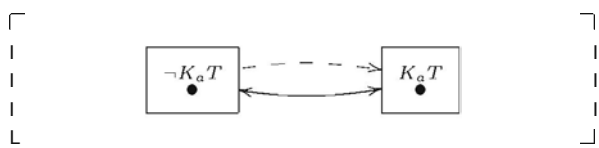
11 We stress this is a minor restrietion, and it is very easy to extend this setting to "ontic" actions. The only reasen we stick with this restrietion is that it simplifies the definitions, and that it is general enough to apply to all the actions we are interested here, and in particular to all *communication actions.*

12 As in (Baltag and Moss, 2004), we will be able to represent non-deterministic actions as sums (unions) of deterministic ones.

σ: this action can be executed on a state *s* iff *s* satisfies its precondition. The plausibility pre-orderings $\leq_a$ give the agent's beliefs about which actions are more plausible than others. But this should be interpreted as *beliefs about changes,* that *ericode changes of beliefs.* **In** this sense, we use such "beliefs about actions" as a way to represent doxastic changes: the information about how the agent changes her beliefs is captured by our action plausibility relations. So we read $\sigma <a\ \sigma'$ as saying that: if agent *a* is given the information that some (virtual) action is *either* σ *or* σ' *(without being able to know which),* then she believes that σ is the one actually happening.

Example 2: Successful lying. The action of "public successful lying" can be described as follows: given a doxastic proposition P, the model consists of two actions Lie, **P**and True., P, the first being the action in which agent *a* publicly lies that (she knows) P (while in fact she doesn't know it), and the second being the action in which *a* makes a truthful public announcement that (she knows) P. The preconditions are pret l.ie, P) = $\neg K_a$**P** and pre/'Irue, P) = *KaP*. Agent *a's* equivalence relation is simply the *identity:* she knows whether she's lying or not. The other agents' equivalenee relation is the *total relaiion:* they cannot know if *a* is lying or not. Let us assume that *a's* plausibility preorder is also the *total relaiion:* this would express the fact that agent *a* is *noi decided to a!ways tie; a priori,* she considers equally plausible that, in any arbitrarily given situation, she will lie or not. But the plausibility relations should refiect the fact that we are modeling a "typically successful lying": by default, in such an action, the hearer trusts the speaker, so he is inclined to believe the lie. Hence, the relation for any hearer $b \neq a$ should *make it more p!ausib!e to him that a* is *telling the truili rather than !ying:* True, P <*s* Lie, P.

As *aspecific examp!e,* consider the scenario in Example 1, and assume now that Alice tells Bob (aftel' seeing that the coin was lying Heads up): "I saw the face, so now I know: The coin is lying Tails up". Assume that Bob trusts Alice completely, so he believes that she is telling the truth. We can model this action using the following action model $\mathbf{\Sigma}$:



This model has two actions: the one on the left is the real action that is taking place (in which Alice's sentence is a *lie:* in fact, *she doesnit know* the coin is Tails up), while the one on the right is the other possible action (in which Alice is telling the truth: she *does know* the coin is Tails up). We labeled this node with their preconditions, $\neg K_a T$ for the lying action and $K_a T$ for the truth-telling action. **In** each case, Alice *knows* what action she

is doing, so her information cells (the continuous squares) are singletons; while Bob is uncertain, so the dashed square includes both actions. As before, we use arrows for plausibility relations, skipping all the loops. As assumed above, Alice is not decided to always lie about this; so, *a priori*, she finds her lying in any such given case to be equally plausible as her telling the truth: this is refiected by the fact that the continuous arrow is bidirectional. **In** contrast, (Bob's) dashed arrow points only to the node on the right: he really believes Alice!

The product update of two plausibility modeis. We are ready now to define the *updated (state) plausibility model*, representing the way some action, from an action plausibility model $\mathbf{\Sigma} = (\Sigma, \leftarrow, \leq_a, \text{pre})$aEA, will act on an input-state, from an initially given (state) plausibility model $S = (S, \sim_a, \leq_a, V)$aEA' We denote this updated model by $S \otimes \mathbf{\Sigma}$, and we call it the *update product* of the two modeis. Its states are elements (s, σ) of the Cartesian product $S \times \Sigma$. More specifically, the set of states of $S \otimes \mathbf{\Sigma}$ is

$$S \otimes \Sigma := \{(s, \sigma) : s \in \text{pre}(\sigma)_S\}$$

The valuation is given by the original input-state model: for all $(s, \sigma) \in S \otimes \Sigma$, we put $(s, \sigma) \models p$ iff $s \models p$. As epistemic uncertainty relations, we take the *product* of the two epistemic uncertainty relationsl": for $(s, \sigma), (s', \sigma') \in S \otimes \Sigma$,

$$(s, \sigma) \sim_a (s', \sigma') \text{ iff } \sigma \sim_a \sigma', s \sim_a s'$$

Finally, we define the plausibility relation as the *anti-lexicogmphic preorder relaiion on pairs* (s, σ), i.e.:

$$(s, \sigma) \leq_a (s', \sigma') \text{ iff either } \sigma <_a \sigma' \text{ or } \sigma \simeq_a \sigma', s \leq_a s'.$$

In (Baltag and Smets, 2008), we called this type of product operation the *Action-Prioritu Update*, with a term due to .1. van Benthem (personal communication).

Interpretation. To explain this definition, reeall first that we only deal with *pure "belief changes"*, not affecting the "facts" : this explains our "conservative" valuation. Second, the product construction on the epistemic indistinguishability relation $\sim_a$ is the same as in (Baltag and Moss, 2004): if two indistinguishable actions are successfully applied to two indistinguishable input-states, then their output-states are indistinguishable. Third, the anti-lexicographic preorder gives "priority" to the *action* plausibility relation; this is not an arbitrary choice, but is motivated by our above-mentioned interpretation of "actions" as specific types of *belief changes*.

13 Observe that this is precisely the uncertainty relation ofthe epistemic update product, as defined in (Baltag and Moss, 2004).

The action plausibility relation captures what agents *realls] believe* is *going on at the moment;* while the input-state plausibility relations only capture *past beliefs.* The doxastic action is the one that "changes" the initial doxastic state, and not vice-versa. If the "believed action" α requires the agent to revise some past beliefs, then so be it: this is the whole point of believing α, namely to use it to revise or update one's past beliefs. For example, in a successful lying, the action plausibility relation makes the hearer believe that the speaker is telling the truth; so she'll accept this message (unless contradicted by her knowiedge), and change her past beliefs appropriately: this is what makes the lying "successful". Giving priority to action plausibility does not in any way mean that the agent's belief in actions is "stronger" than her belief in states; it just captures the fact that, at the time of updating with a given action, *the belief about the action* is *what* is *actual, what* is *present,* is *the curreni belief about what* is *going on, while the beliefs about the input-states are in the past.*[14] The belief update *induced by a given action* is nothing but *an update with the (presently) believed action.*

In other words, the anti-lexicographic product update refiects our Motto above: *beliefs about changes* (as formalized in the action plausibility relations) *are nothing but ways to ericode changes of belief* (i.e., ways to change the original plausibility order on states). This simply expresses our *particular inierpretation* of the (strong) plausibility ordering on actions, and is thus a matter of *convention:* we decided to introduce the order on actions to eneode corresponding *changes of order* on states. *The product update* is *a consequence of this convention:* it just says that a strong plausibility order $\alpha <_a \beta$ on actions corresponds indeed to a change of ordering, (from whatever Lhe ordering was) between Lhe original input-statee $s, t,$ Lo Lhe order $(s, \alpha) <_a (t, \beta)$ between output-states; while equally plausible actions $\alpha \simeq_a \beta$ wilileave the initialordering unchanged: $(s, \alpha) \leq_a (t, \beta)$ iff $s \leq_a t$. So the product update is just a formalization of *our inierpretation* of action plausibility modeis, and thus it doesn't impose any further limitation to our setting.

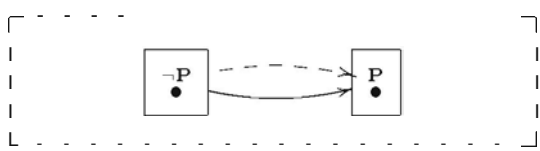
Example 3: By computing the update product of the plausibility model S in Example 1 with the action model Σ in Example 2, we obtain the following plausibility model:

14 Of course, *at a later moment,* the above-mentioned belief about action *(now* belonging to the past) might be itself revised. But this is another, *future update.*

This correctly describes the effect of an action of "successful lying": Bob's plausible ordering is reversed, since he believes Alice, and so now he believes that the coin is lying face up. In contrast, Alice's initial plausibility relation is unaffected (since the two actions were equally plausible, i.e., doxastically equivalent, for her); so, she should keep her *a priori* belief that the coin is Heads up; of course, in this case the last point is not so relevant, since Alice *knows* the state of the coin (as witnessed by the fact that the continuous squares consist of single states).

Example 4: "Hard" **public** announcements. A *iruihful public annouueement* lP *of some "hard fact"* P is not really about belief revision, but about the learning of *certified irue information:* it establishes *common knowledge* that P was the case. This is the action described by van Benthem (2004) as (public) "belief change under hard facts". As an operation on modeis, this is described by van Benthem (2004) as taking any state model S and *deleting all the non-P states, while keeping the same indistinguishability and plausibility relaiions between the surounru; states.* In our setting, the corresponding action model consists of only one node, labeled with P. It is easy to see that the above operation on models can be exactly "simulated" by taking the anti-lexicographic product update with this one-node action model.

Example 5: "Lexicographic upgrade" as a "soft" **public** announcement. To allow for "soft" belief revision, an operation $\Uparrow \mathbf{P}$ was introduced by van Benthem (2004), essentially adapting to public announcements the 'lexicographic' policy for belief revision described by Rott (1989). This operation, called "lexicographic update" consists of changing the cutrent plausibility order on any given state model as follows: *all* $\mathbf{P}$*-worlds become more plausible than all* $\neg\mathbf{P}$*-worlds, and within the two zones, the old ordering remacns.* In our setting, this action corresponds to the following local plausibility action model:



Taking the anti-lexicographic update product with this action will give an exact "simulation" of the lexicographic upgrade operation.

Proposition 3.1. The update product of a state plausibility model and an action plausibility model is a state plausibility model.

4 Product update in CDM form

As for the "static" conditional doxastic logic, the axioms of "dynamic" belief revision logic can be easily derived if we first work out the CDM version of the above update product. We do this from scratch, by first introducing a "dynamic" version of the notion of CDM, equivalent to the above concept of action plausibility model:

A *conditional doxastic action model* (CDAM, *[or short}* **Σ** is just a conditional doxastic frame $(\Sigma, \{\bullet_a^\Pi\}_{a\in\mathcal{A},\Pi\subseteq\Sigma})$, together with a *precondition map* pre : $\Sigma \to$ *Prop* as above. A set of actions II $\subseteq \Sigma$ can be interpreted as *pariiel injormaiion* about some real (basic) action σ E II, or equivalently, as a *non-deterministic action* (in which one of the actions σ E II happens, but we are not told which). The conditional appearance σ_a^Π captures the way *action σ eppears to agent a, given additional (plausible, but noi tiecessarily truthful) injormaiion* II *about this action.* This means that, in normal circumstances, if aftel' σ happens the agent is told that (one of the actions in) II has happened, then the agent will believe that in fact (one of the basic actions in) σ_a^Π has happened. As before, any action plausibility model induces in a canonieal way a CDAM, and conversely any CD AM can be represented as the canonieal CDAM of some action plausibility model.

Example: Lying, revisited. In the successful lying example, if we convert the plausibility model into its canonieal CDM, we obtain, e.g., that $(\text{Lie}_a \mathbf{P})_b^Q$ = {True_a **P}** for $b \neq a$ and $Q \not\subseteq$ {Lie_a P}. So this lying is *indeed genemlly "successful":* no matter what other information is given to *b,* if it is consistent with *a* telling the truth, then *b* believes that atelIs the truth. The only case in which the appearance of this action to *b* is different is when $Q \subseteq$ {Lie, P}, in which case $(\text{Lie.}, \mathbf{P})_b^Q = Q$, and in particular, $(\text{Lie}, \mathbf{P})_b^{\{\text{Lie}_a\}}$ = {Lie, P}: so the hearer can discover the lying only if given information that excludes all other possible actions.

Independenee of action's **appearance from prior beliefs.** The above description assumes that *the aqerü's beliefs about the action are independent of his beliefs about the state:* indeed, σ_a^Π contains no information about, or reference to, the cutrent state's doxastic appearance s_a to the agent, so it is assumed that this does not inftuence in any way the appearance of the action. This assumption embodies a certain interpretation of our "appearance" maps: we take an action's appearance to simply denote the action itself, as it appears to the agent. **In** other words: for the agent, the appearance is Lhe act.ion, pure and sirnple. When Lhe action σ happens (say, in an unconditional context), it *realls]* appears to the agent as if (the apparent, un-conditionalized action) $\sigma_a := \sigma_a^\Sigma$ happens. **If** the agent makes the additional hypothesis that one of the actions in II happens, then it

appears to him that σ_a^Π is happening. The action's appearance is simply taken here as *a brute, new fact:* the agent really believes this apparent action is happening (otherwise this would *noi* really be the appearance of this action). *This belief cannot be retnsed at the same time that it* is *being held: any reoision of the action's appeamnce can only happen in the future.* But for the moment, this appearance correctly reftects what the agent thinks to be happening. **In** contrast, his prior beliefs about the state are just that: *prior* beliefs. They may be subject to revision at this very moment, due to cutrent action (or, more precisely, due to its appearance): indeed, the (apparent) action is the one that *induces* the revision (or update) of the static belief. **In** a certain sense, the action, as it appears, is the belief update: the apparent action simply eneodes the way the agent is compelled to update his prior beliefs. Hence, the action's appearance cannot, by definition, be dependent, or be inftuenced, by these prior beliefs: the action's appearance is a given, it is what it is, and the prior beliefs are the ones that may be changed by the apparent action, not vice-versa.

Taking the action's appearance as a correct description of the action as seen by the agent, the above independence (of this appearance from prior beliefs) can be understood as a *mtionality postulate:* agents should be prepared to revise their prior beliefs when faced with *(what eppears to them as) iruihful new information.* Rational agents are not fundamentalists: if given compelling evidence to the contrary (as encoded in the "apparent action"), they will not refuse it due to prior beliefs, but will change these prior beliefs to fit the new evidence. And it does not matter in the least that, at some later point, this "compelling evidence" might turn out to have been a belief (an "apparent acr.ion"), noL a realiLy: when this will happen, rational agents might change their minds again. But for the moment, they have to accept the cutrent action *as it eppears to them,* and adjust their previous beliefs appropriately.

An action's contextual appearance. **In** the context of belief revision, there is a subtie point to be made here: the above independence only refers to the agent's prior *beliefs,* but not to the *aqent's knowiedge. No action's eppearanee can be assumed to be independent of prior knowiedge:* it might happen that the cutrent state s is such that agent *a knows* that the believed action σ_a^Π *cannot happen* at this state. This is perfectly possible, even in states in which *a* does happen, and even if the information II is correct (i.e., $a \in \text{II}$). **In** such a state, the agent cannot accept the default appearance σ_a^Π. Prior knowledge may thus inftuence the action's appearance.

Example 2, revisited: *"Successjul" lying cannot always be successful!* Indeed, if the original input-state s is such that an outsider *b alreeds] knows that P* is *false,* then lying cannot succeed. **In** this context, the appearance

of the action Lie, P to *b* is *noi* its default appearance True, P: *b* cannot believe that *a* is telling the truth. Instead, the *contextual appeamnce of this action at* s is *itself:* (Lie., *P)b* = Lie, P. The hearer knows the speaker is lying.

So σ_a^Π should only be thought of as the action's *default appeamnce* (conditional on 11) to the agent: *in the absence of any other additional information* (except for 11), or *uiheneoer the aqent's prior knowledge allows this,* the agent *a* will believe that σ_a^Π has happened.

So how will this action appeal' in a context in which the default appearance is known to be impossibie? We can answer this question by defining a *contextual appeamnce* $\sigma_a^{s,\Pi}$ of action *a* to agent *a at state s, given* 11. We can do this by strengthening our conditionalization: at a given state s ∈ *S,* an agent has *already some information about the next action,* namely that it cannot be inconsistent with his knowledge *s(a)* of the state. **In** other words, agent *a* knows that the action must belong to the set $\Sigma_{s(a)} := \{p \in \Sigma : s(a) \cap \text{pre}(p)s \neq \emptyset\} = \{p \in \Sigma : s \not\models_S K_a \neg \text{pre}(p)\}$. Putting this information together with the new information 11, we obtain the contextual appearance by *conditionalizing the aqent's belief about the action with* $\Sigma_{s(a)} \cap$ 11:

$$\boldsymbol{\sigma_a}^{s,\Pi} := \sigma_a^{\Sigma_{s(a)} \cap \Pi} = \boldsymbol{\sigma_a}^{\{\rho \in \Pi : s(a) \cap \text{pre}(\rho) \neq \emptyset\}}$$

This contextual appearance is the one that fully captures the agent's *actual belief about the action a* in state s, whenever he is given information 11.

An action's **effect: Deterministic change of state.** As announced, we take the basic actions *a* ∈ Σ to represent deterministic changes of states. **In** the following, we will always represent *the output-state a* (s) *of applying basic action a to state* s ∈ *S* by an ordered pair *a* (s) := *(s,* o). So, for a given CDM S of possible input-states and a given CD AM of possible actions, the set of all possible output-states will be a subset of the Cartesian product $S \times \Sigma$. Thus, we could represent *post-conditions,* i.e., conditions restricting the possible output-states of some (unspecified) action acting on some (unspecified) input-state as *subsets* $P \subseteq S \times \Sigma$ *of the Cartessari product.* Given a basic action *a* ∈ Σ and a post-condition $P \subseteq S \times \Sigma$, we may denote the *set of possible input-states of action a ensuring post-condition P* by:

$$\sigma^{-1}(P) = \{s \in S : \sigma(s)\ EP\} = \{s \in S : (s, \sigma)\ EP\}$$

Post-conditional contextual appearance. Sometimes, the additional information the agent may be given (or the hypothesis that he may entertain) refers, not directly to the range 11 ⊆ Σ of possible actions currently happening, but to some *post-condition P;* i.e., the agent might be told that

the cutrent action will result in a state $\sigma(s) = (s,\sigma)$ satisfying some post-condition $P \subseteq S$ x Σ. He should be able to conditionalize his belief about the cutrent action with this information, in a given context. For this, we define the *contextual eppearanee* $\sigma_a^{s,P}$ *of action* σ *at state s, in the hypothesis (that the action wil! ertsure postcondition) P,* by putting:

$$\sigma_a^{s,P} := \sigma_a^{\{\rho\in\Sigma : s(a)\cap\rho^{-1}(P)\neq\varnothing\}}$$

Example: lying, again. Let again Lie, P be the action of successfullying by agent *a,* and suppose that P denotes *afactual* ("ontic") statement (which happens to be false, and thus will remain false aftel' lying). Even if in the original state, the hearer *b* did *not know* that P was false (so that lying *was* successful, and its appearance to *b* was the default one True, P), he may be given later that information, as a post-condition $\neg\mathbf{P}$. *Then,* the hearer discovers the lying: the post-conditional contextual appearance of lying (given $\neg\mathbf{P}$) is... lying!

Belief revision induced by an action and a postcondition. We want to calculate now the *reoisiori of an aqerü's beliefs (about an input-state s) induced by an action* σ *when given some post-condition* $P \subseteq S$ x Σ. We denote this by $s_a^{\sigma,P}$. This captures the *appeamnce of the input-state s to agent a, afier action* σ *and ajier being given the informaiion that P holds at the output-state.* As explained already, the agent revises his prior beliefs not in accordance with the actual action, but in accordance to how this action appears to him. As we have seen, the appearance of action σ at state *s* when given post-condition *P* is $\sigma_a^{s,P}$. So the new information obtained post-factum about the original input-state *s* is that this state was capable of supporting (one of the actions in) $\sigma_a^{s,P}$, and moreover that it yielded an output-state satisfying post-condition *P.* **In** other words, the agents learns that the original state was in $(\sigma_a^{s,P})^{-1}(P)$. So he has to revise (conditionalize) his prior beliefs about *s* with this information, obtaining:

$$s_a^{\sigma,P} := s(\sigma_a^{s,P})^{-1}(P)$$

Product update, in CDM **form.** We now give a CDM equivalent of the above notion of product update: the *product update* of a conditional doxastic model S with a conditional doxastic action model $\mathbf{\Sigma}$ is a new conditional doxastic model S $\otimes$ $\mathbf{\Sigma}$, whose states are elements $\sigma(s) := (s,\sigma)$ of a subset $S\otimes\Sigma$ of the Cartesian product $S\times\Sigma$. Note that we prefer here the functional notation $\sigma(s)$, instead of (s,σ). As before, preconditions select the surviving states:

$$S \otimes \Sigma := \{\sigma(s) : s \in \mathrm{pre}(\sigma)_S\}$$

For any hypothesis $P \subseteq S \otimes \Sigma$ about the output-state, the conditional appearance (conditioned by *P)* of an output-state σ(s) to an agent *a* is given by

$$\sigma(s)_a^P := \sigma_a^{s,P}(s_a^{\sigma,P}) \; \boldsymbol{nP}$$

In words: *Agent a 's updated belief* (about the output-state of a basic action σ applied to an input-state s, when given condition *P)* $\sigma(s)_a^P$ can be obtained by applying the *action that* is *believed to happen* (i.e., the appearance $\sigma_a^{s,P}$ of σ to *a* at s, given post-condition *P)* to the agent's *reoised belief about the input-state* $s_a^{\sigma,P}$ (belief revised with the information provided by the apparent action $\sigma_a^{s,P}$), then *reetrictinç to the given post-condition P*. Finally (as for plausibility modeis), the valuation on output-states comes from the original states:

$$\|p\|_{\mathbf{S}\otimes\mathbf{\Sigma}} := \{\sigma(s) \; \mathsf{ES} \otimes \Sigma : \mathrm{s} \; \mathsf{E||p||S}\}$$

Proposition 4.1. The two "product update" operations defined above agree: the canonical CDM associated to the (anti-lexicographic) product update of two plausibility models is the product update of their canonieal CDM's.

5 The dynamic logic

As in (Baltag and Moss, 2004), we consider a doxastic *sumature,* i.e., a *finite (fixed) plausibility frame* (or, equivalently, a finite conditional doxastic frame) $\mathbf{\Sigma}$, together with an *ordered list without repetitiotie* $(\sigma_1, \ldots, \sigma_n)$ of some of the elements of $\mathbf{\Sigma}$. Each signature gives rise to a dynamic-doxastic *logic* $\mathrm{CDL}(\mathbf{\Sigma})$, as in (Baltag and Moss, 2004): one defines by double recursion a set of *sentences* φ and a set of *proqram terms* π; the *basic proqrams* are of the form $tt = \sigma\vec{\varphi} = \sigma\varphi_1 \ldots \varphi_n$, where $\sigma \; \mathsf{E} \; \Sigma$ and φ_i are sentences in our logic; program terms are generated from basic programs using *nondetermirusiic sum (choice)* $\pi \cup \pi'$ and *sequential composition* $\pi; n'$, Sentences are built using the operators of CDL, and in addition a dynamic modality $\langle\pi\rangle\varphi$, taking program terms and sentences into other sentences. As in (Baltag and Moss, 2004), the conditional doxastic maps on the signature $\mathbf{\Sigma}$ induce in a natural way conditional doxastic maps on basic programs in $\mathrm{CDL}(\mathbf{\Sigma})$: we put $(\sigma\vec{\varphi})_a^{\Pi\vec{\varphi}} := \{\sigma'\vec{\varphi} : \sigma' \; \mathsf{E} \; \sigma_a^{\Pi}\}$. The given listing can be used to assign syntactic preconditions for basic programs, by putting: $\mathrm{pre}(\sigma_i\vec{\varphi}) := \varphi_i$, and $\mathrm{pre}(\sigma\vec{\varphi}) :=$ T (the trivially true sentence) if σ is not in the listing. Thus, the basic programs of the form $\sigma\vec{\varphi}$ form a (finite) *syntactic* CDAM[15] $\mathbf{\Sigma}\vec{\varphi}$. Every given interpretation $\mathrm{11\_11} : \mathrm{CDL}(\mathbf{\Sigma}) \to$ *PTOp* of

15 A *syntactic* COAM is just a conditional doxastic frame endowed with a *syntactic* precondition map, associating sentences to basic action. For justification and examples, in the context of *epistemic* action models, see (Baltag and Moss, 2004).

sentences as doxastic propositions will convert this syntactic model into a "reai" (semantic) CDAM, called $\mathbf{\Sigma}\|\vec{\varphi}\|$.

To give the semantics, we define by induction two *inierpretation maps*, one taking any sentence φ to a doxastic proposition $\|\varphi\| \in PTOP$, the second taking any program term α to a (possibly non-deterministic) doxastic "program", i.e., a *set* of basic actions in some CDAM. The definition is completely similar to the one in (Baltag and Moss, 2004), so we skip the details here. Suffice to say that the semantics of basic dynamic modalities is given by the inverse map:

$$\|\langle\sigma\vec{\varphi}\rangle\psi\|_{\mathbf{S}} = \left(\sigma\|\vec{\varphi}\|_{\mathbf{S}}\right)^{-1}\|\psi\|_{\mathbf{S}\otimes\mathbf{\Sigma}\|\vec{\varphi}\|}$$

Notation. To state our proof system, we encode the notion of *post-conditional contextual appeamnce of an action* in our syntax. For sentences ϑ, ψ and basic program $\alpha = \sigma\vec{\varphi}$, we put:

$$\langle\alpha_a^{\vartheta}\rangle\psi := \bigvee_{\mathsf{TI}\subseteq\mathbf{\Sigma}\vec{\varphi}}\left(\langle\alpha_a^{\Pi}\rangle\psi \wedge \bigwedge_{\beta\in\mathsf{TI}}\neg K_a\neg\langle\beta\rangle\vartheta \wedge \bigwedge_{\beta'\notin\mathsf{TI}} K_a\neg\langle\beta'\rangle\vartheta\right)$$

This notation can be justified by observing that it semantically matches the modality corresponding to post-conditional contextual appearance:

$$\|\langle\langle\sigma\vec{\varphi}\rangle_a^{\vartheta}\rangle\psi\|_{\mathbf{S}} = \left\{s \in S : s \in \left((\sigma\|\vec{\varphi}\|_{\mathbf{S}})_a^{s,\|\vartheta\|_{\mathbf{S}}}\right)^{-1}\|\psi\|_{\mathbf{S}\otimes\mathbf{\Sigma}\|\tilde{\varphi}\|}\right\}$$

Theorem 5.1. A *complete proof system* for the logic $\mathrm{CDL}(\mathbf{\Sigma})$ is obtained by adding to the above axioms and rules of CDL the following Reduction Axioms:

$$\begin{aligned}
\langle\pi \cup \pi'\rangle\varphi &\leftrightarrow \langle\pi\rangle\varphi \vee \langle\pi'\rangle\varphi\\
\langle\pi;\pi'\rangle\varphi &\leftrightarrow \langle\pi\rangle\langle\pi'\rangle\varphi\\
\langle\alpha\rangle p &\leftrightarrow \mathrm{pre}(\alpha)\wedge p\\
\langle\alpha\rangle\neg\varphi &\leftrightarrow \mathrm{pre}(\alpha)\wedge\neg\langle\alpha\rangle\varphi\\
\langle\alpha\rangle(\varphi\vee\psi) &\leftrightarrow \langle\alpha\rangle\varphi \vee \langle\alpha\rangle\psi\\
\langle\alpha\rangle B_a^{\vartheta}\varphi &\leftrightarrow \mathrm{pre}(\alpha)\wedge B_a^{\langle\alpha_a^{\vartheta}\rangle\vartheta}[\alpha_a^{\vartheta}](\vartheta\to\varphi)
\end{aligned}$$

where pis any atomie sentence, π, π' are programs and α is a *basic* program in $L(\mathbf{\Sigma})$.

The soundness of the last reduction axiom is obvious, once we see that its holding at a state s follows immediately from the definition of the product update in CDM form

$$\sigma(s)_{\mathbf{a}}^{P} := \sigma_{\mathbf{a}}^{s,P}(s_{\mathbf{a}}^{\sigma,P}) \cap P,$$

by taking $\sigma := \alpha$, $P := \|\vartheta\|$ and using the semantics of dynamic modalities.

Special cases. **If** we put the last reduction axiom in its dual (universal modality) form, we obtain

$$[\alpha]B_a^{\vartheta}\varphi \leftrightarrow \mathrm{pre}(\alpha) \to B_a^{\langle\alpha_a^{\vartheta}\rangle\vartheta}\,[\alpha_a^{\vartheta}]\,(\vartheta \to \varphi).$$

As special cases of the Action-Conditional-Belief Law, we can derive *the redmetion laws* from (van Benthem, 2004) for (conditional) belief aftel' the events $!\psi$ and $\Uparrow\psi$:

$$[!\psi]B_a^{\vartheta}\varphi \;\leftrightarrow\; \psi \to B_a^{\psi\wedge[!\psi]\vartheta}[!\psi]\varphi\,,$$
$$[\Uparrow\psi]B_a^{\vartheta}\varphi \;\leftrightarrow\; \left(\tilde{K}_a^{\psi}[\Uparrow\psi]\vartheta \wedge B_a^{\psi\wedge[\Uparrow\psi]\vartheta}[\Uparrow\psi]\varphi\right) \vee \left(\neg\tilde{K}_a^{\psi}[\Uparrow\psi]\vartheta \wedge B_a^{[\Uparrow\psi]\vartheta}[\Uparrow\psi]\varphi\right)$$

where $K_a^{\psi}\vartheta := K_a(\psi \to \vartheta)$, $\tilde{K}_a^{\psi}\vartheta := \neg K_a^{\psi}\neg\vartheta$, $\tilde{B}_a^{\psi}\vartheta := \neg B_a^{\psi}\neg\vartheta$.

Acknowledgments

Sonja Smets' contribution to this research was made possible by the post-doctoral fellowship awarded to her by the Flemish Fund for Scientific Research. We thank Johan van Benthem, Hans van Ditmarseh, Jan van Eijck, Larry Moss and Hans Rott for their valuable feedback.

References

Adams, KW. (1965). A Logic of Conditionals. *lnquirs],* 8:166-197.

Alchourrón, C.K, Gärdenfors, P. & Makinson, D. (1985). On the Logic of Theory Change: Partial Meet Contraction and Revision Functions. *The Journal of Symbolic Logic, 50(2):510-530.*

Aucher, G. (2003). A Combined System for Update Logic and Belief Revision. Master's thesis, ILLC, University of Amsterdam, Amsterdam, The Netherlands. *ILLC Fublications* MoL-2003-03.

Baltag, A. & Moss, L.S. (2004). Logies for Epistemic Programs. *Synthese,* 139(2): 165-224.

Baltag, A., Moss, L.S. & Solecki, S. (1998). The Logic of Common Knowledge, Public Announcements, and Private Suspicions. **In** Gilboa, 1., ed., *Fmceedings of the 'lth Confer-ence on Theoretical Aspects of Rationality and Knowiedge, (TARK'98),* pp. 43-56.

Baltag, A. & Sadrzadeh, M. (2006). The Algebra of Multi-Agent Dynamic Belief Revision. *Electrome Notes in Theoretical Computer- Science, 157(4).*

Baltag, A. & Smets, S. (2006a). Conditional Doxastic Modeis: A Qualitative Approach to Dynamic Belief Revision. *Electronse Notes in Theoreiical Computer Science,* 165:5-2l.

Baltag, A. & Smets, S. (2006b). Dynamic Belief Revision over Multi-Agent Plausibility Modeis. **In** Bonanno, *G.,* van der Hoek, W. & Wooldridge, M., eds., *Proceedinqs of the 'lih Conference on Logic and the Foundations of Game and Decision Theoru (LOFT*$^{JO6)}$, pp. 11-24.

Baltag, A. & Smets, S. (2006c). The Logic of Conditional Doxastic Actions: A Theory of Dynamic Multi-Agent Belief Revision. **In** Artemov, S. & Parikh, R., eds., *Proceedinqs of the ESSLLI'06 Workshop on Rationality and Knowiedge,* pp. 13-30.

Baltag, A. & Smets, S. (2007a). From Conditional Probability to the Logic of Doxastic Actions. **In** Samet, D., ed., *Proceedinqs of the l1th Conference on Theoretical Aspects of Rationality and Knowledge (TARK-2007), Brussels, Belgium, June* 25-27, *2007,* pp. 52-6l.

Baltag, A. & Smets, S. (2007b). Probabilistic Dynamic Belief Revision. **In** van Benthem, .1., Ju, S. & Veltman, F., eds., *Proceedinqs of the Workshop Logic, Rationality and Interaction (LORI'07),* London. College Publications. An extended version to appeal' in *Synthese.*

Baltag, A. & Smets, S. (2008). A qualitative theory of dynamic interactive belief revision. **In** Bonanno, *G.,* van der Hoek, W. & Wooldridge, M., eds., *Logic and the Foundations of Game and Decision Theoru (LOFT* 7), Vol. 3 of *Texts in Logic and Games,* pp. 11-58. Amsterdam University Press.

Bennett, .1. (2003). *A Fhilosophical Guide to Conditionals.* Oxford University Press.

van Benthem, J. (2004). Dynamic Logic for Belief Revision. *Joumal of Applied Non-Classical Logies, 14(2):129-155.*

van Benthem, .1. & Liu, F. (2004). Dynamic Logic of Preferenee Upgrade. *Joumal of Applied Non-Classical Logies, 14:157-182.*

van Benthem, .1., van Otterloo, S. & Roy, O. (2005). Preferenee Logic, Conditionals and Solution Concepts in Games. Technical Report *ILLC Fublications* PP-2005-28, University of Amsterdam.

Board, O. (2002). Dynamic Interactive Epistemology. *Games and Economie Behaoiour,* 49(1):49-80.

van Ditmarseh, H. (2005). Prolegomena to Dynamic Logic for Belief Revision. *Synthese, 147(2):229-275.*

Friedmann, N. & Halpern, J.Y. (1994). Conditionallogics of belief revision. **In** Hayes-Roth, B. & Korf, R., eds., *Proceedinqs of the Tweljth National Conference on Artificial Intelligence. AAAI-94. Seattle, Washington.,* pp. 915-921, Menlo Park, CA. AAAI Press.

Fuhrmann, A. & Levi, I. (1994). Undercutting and the Ramsey Test for Conditionals. *Synthese, 101(2):157-169.*

Gärdenfors, P. (1986). Belief Revisions and the Ramsey Test for Conditionais. *Fhilosophical Review, 95(1):81-93.*

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

Gerbrandy, .1. (1999). Dynamic Epistemie Logic. **In** Moss, L.S., Ginzburg, .1. & de Rijke, M., eds., *Logic, Language and Computation, Vol.* 2. **CSLI** Publications, Stanford University.

Grove, A. (1988). Two Modellings for Theory Change. *Joumal of Fhilosophical Logic, 17(2):157-170.*

Halpern, J.Y. (2003). *Reasoning about Uncertainty.* **MIT** Press.

Katsuno, H. & Mendelzon, A.O. (1992). On the Difference between Updating a Knowledge Base and Revising It. *Cambridge Tracis in Theoretical Computer Science, 29:183-203.*

Lehrer, K. (1990). *Theory of Knowiedge.* Dimensions of Philosophy. Routledge, London.

Lehrer, K. & Paxson, Jr., T. (1969). Knowiedge: Undefeated Justified True Belief. *Joumal of Fhilosophy, 66(8):225-237.*

Lewis, D. (1973). *Counterfactuals.* Blackwell Publishing, Oxford.

Meyer, .L-J.e. & van der Hoek, W. (1995). *Epistemic Logic [or AI and Computer Science.* Cambridge University Press.

Ramsey, F. (1931). *The Foundations of Mathematics and Oiher Essays.* Kegan Paul, London.

Rott, H. (1989). Conditionals and Theory Change: Revisions, Expansions, and Additions. *Synthese,* 81(1):91-113.

Ryan, M. & Schobbens, P.Y. (1997). Counterfactuals and Updates as Inverse Modalities. *Joumal of Logic, Language and Injormation, 6(2):123-*146.

Segerberg, K. (1998). Irrevoeable Belief Revision in Dynamie Doxastie Logie. *Notre Dame Journalof Formol Logic, 39(3):287-306.*

Spohn, W. (1988). Ordinal Conditional Funetions: ADynamie Theory of Epistemie States. 42:105-134.

Stalnaker, RC. (1968). A Theory of Conditionals. **In** Reseher, N., ed., *Studies in Logical Theorsj,* Vol. 2 of *APQ Monogmphs.* Blaekwell, Oxford.

Stalnaker, RC. (1996). Knowiedge, Belief and Counterfaetual Reasoning in Games. *Economics and Philosophy, 12:133-163.*

Stalnaker, RC. (2006). On Logies of Knowledge and Belief. *Philosophical Studies, 128(1):169-199.*

Comments on 'The Logic of Conditional Doxastic Actions'

Hans van Ditmarsch

Department of Computer Scîence
University of Otago
PO Box 56
Dunedin 9054, New Zealand

lnstitut de Recherche en lnformatique de Toulouse
Université Paul Sabatier
118 Route de Narbonne
31062 Toulouse Cedex 4, France

hans@cs.otago.ac.nz, hans@irit.fr

Abstract

I first summarize Baltag and Smets' contribution to this volume, and praise their work. Then I compare the anti-lexicographic plausibility update that they propose to a proposal by Aucher, as an illustration of the difference between a qualitative and a quantitive formulation of updates. I quote Spohn's original work that is at the root of research in plausibility updates and of the notion of anti-lexicographic update. Some technical notes on different partial orders used in belief revision serve as a prelude to an observed relation between the qualitative and quantitative representation of structures. Finally I address Baltag and Smets' analysis of the action of lying.

In this commentary on 'The Logic of Conditional Doxastic Actions' I am in the delightful position of having the last word in an argument with Alexandru Baltag. This position is very hard to obtain. But because in this volume my commentary follows the chapter by Alexandru Baltag and Sonja Smets, any further elaborations and involutions will be out of reach to the readers of the volume. lam going to use this rare advantage to the limit.

Having said that, I sent a preliminary version of these comments to Alexandru and Sonja for comments, and immediately received in response an email of about the same length as this submission. I am very grateful for Alexandru's last words. Weil, nearly last words. I made some further changes. Now, the work is done.

1 The logic of conditional doxastic actions

In AGM belief revision a distinction is made between belief expansion and (proper) belief revision. Given a set of consistent beliefs, in belief expan-

Krzysztof R. Apt, Robert van Rooij (eds.). *New Perspectives on Games and Interaction.* Texts in Logic and Games 4, Amsterdam University Press 2008, pp. 33-44.

sion new information can typically be added as such, without conflict with existing beliefs. But in belief revision, the incoming information is typically inconsistent with the prior beliefs. A long line of work in dynamic epistemic logic, prominently including the well-known framework of action modeis, also by Baltag but with different collaborators, can be seen as a generalization of belief expansion. Unlike AGM belief revision in its original formulation, this dynamic epistemic logic also models belief expansion for more than one agent, and what is known as higher-order belief change: given explicit operators for 'the agent believes that', self-reference to one's own belief or to the beliefs of ot hers can also be formalized. A problem in that line of research remained that the typical belief *reoision,* i.e., how to process inconsistent new beliefs, cannot be modelled. Belief in factual information, for example, cannot be given up when confronted with new belief that is considered as acceptable evidence to the contrary. And this is not just impossible within the setting of knowiedge, where one does not expeet proper revision to be possible, because knowledge is truthfu1. It is also impossible for weaker epistemic notions.

In this contribution to the volume, Alexandru Baltag and Sonja Smets introduce adynamie epistemic framework in which belief revision in the proper sense is, aftel' all, possible. Given a structure (an *epistemic plausibility frame*) wherein one does not merely have epistemic indistinguishability between states but also plausibility relations between states, one can define both knowledge and conditional belief operators. Unconditional belief is defined as belief that is conditional to the trivial state of information. (The trivial state of information is the epistemic equivalence class occupied by Lhe agent, which is described by Lhe Iorrnula $\top$.) **In** this seu.ing belief revision is possible where the agent (unconditionally) believed some factual information p but aftel' having been presented with convincing evidence to the contrary, changes his mind, and then believes the exact opposite. Just as for the relation between classic AGM expansion and dynamic epistemic logic, we now have again that this approach also models muiti-agent and higher-order belief revision.

The authors go much further, beyond that. **In** a multi-agent setting there are more complex forms of belief revision than revision with a publicly announced formula φ. That is merely an example of a doxastic action. More complex doxastic actions, where the action appears differently to each agent, are also conceivable. They present a very good example, namely the action where *agent a* is *lying that* φ. To a credulous agent *b,* this action will appeal' as a truthful announcement that φ. But not to *a* of course, who knows that she is Iying. The general form of doxastic actions is like an epistemic plausibility model and is called an *action plausibility model;* the difference is that instead of a valuation of atoms in each state of an

epistemic plausibility model we now have a *precondiiion* for the execution of each action (i.e., element of the domain) of an action plausibility model. The execution of such a doxastic action in a epistemic plausibility state is a restricted modal product, where I am tempted to say 'as usual', to avoid the obligation of having to explain this in detail. The only unusual aspect of this procedure is the very well-chosen mechanism to compute new plausibilities from given plausibities, called *anti-lexicogmphic preorder relaiion.* This says that plausibility among actions takes precedence over plausibility among states. **It** is the natural generalization of the implicit AGM principle that the revision formula takes precedence over the already believed formulas. Anti-lexicographic preorder prescribes that: a new state of affairs is more plausible than another new state of affairs, if it results from an action that is strictly more plausible than the action from which the other state results, or if the states result from equally plausible actions but the former state already was more plausible than the latter state before action execution.

So far, this overview also describes the authors' other publication (Baltag and Smets, 2006). A main focus of their underlying contribution is the interpretation of these results in terms of conditional reasoning and *conditional doxastic action models.* The *conditional eppearanee maps* of these conditional doxastic action models take the place of the plausibility relations among the actions in an action plausibility model. They motivate and justify in great detail various notions for conditional belief, and their interdependencies. A fabulous finale is a complete axiomatization with a reduction axiom that relates conditional belief aftel' an action to conditional belief befere that action. The technicalit.ies of this logic wiLh dynarnic operators for conditional action execution may be hard to follow unless the reader is intimately familiar with the BMS action model framework, as these technical details are only somewhat summarily presented. **In** that case, *just* focus on this reduction axiom, the *action-conditional-beliej law,* and the well-chosen examples given of its application. I can assure you, it's all true.

2 Quality is better than quantity

The authors claim that their approach is 'in its spirit closer to qualitative logics than to approaches of a more quantitative ftavour.' This is a very well-considered phrasing. Let us see why they are right.

One such approach of a more quantitative ftavour is the proposal by Guillaume Aucher (2003). **In** terms of action plausibility models his proposal is a version of van Benthem's soft update referred to in Example 5, but with a different recipe to compute new plausibilities. Aucher also employs structures with epistemic equivalence classes and plausibility relations, but the plausibility relations are derived from a finite total order of degrees of

plausibility. If s is more plausible than s_i, its degree of plausibility is lower. The lowest degree of plausibility is 0. To each degree of plausibility corresponds a degree of belief, expressing that the believed proposition is *at least as plausible* as that. Unconditional belief is belief of degree 0.

Given some model with domain S, Aucher's 'belief revision with a formula φ' now amounts to the following. Whenever φ is true, subtract the minimum degree of the cp-states from the current degree. Otherwise, subtract the minimum degree of the $\neg\varphi$-states from the cutrent degree *and add one*. This ensures that at least one cp-state will get degree 0, and thus factual information φ will be unconditionally believed aftel' revision, as required.

For an example, consider one agent a only and a domain consisting of four states 0, 1,2,3 comprising a single equivalence class (all four states are considered possible by the agent) and such that $0 \leq_a 1 \leq_a 2 \leq_a 3$. In this initial epistemic plausibility structure, the degree of each state is its name.

First, suppose that factual information p is true in state 3 only (the valuation of p is {3}). According to the recipe above, the result is the order $3 \leq_a 0 \leq_a 1 \leq_a 2$. How come? Write dg($s$) for the old degree of state s and dg'{s) for its new degree, aftel' revision. Then $dg'(3)$ = dg(3) - Min{dg(s) Is $\models p$} = 3-3 = 0. Whereas dg'{l) = dg(l) -Min{dg(s) Is $\not\models p$} + 1 = 1 - 0 + 1 = 2. Etcetera. So far so good.

Now for some other examples, demonstrating issues with such quantitatively formulated proposals for belief revision. Suppose that, instead, p is true in states 1 and 2. We now obtain that $1 \leq_a 2 \simeq_a 0 \leq 3$. As a result of this revision, states 2 and 0 have become equally plausible. As a side effect of the revision, such 'Ioss of plausibility information' may be considered less desirable.

Finally, suppose that p was already believed: suppose that p is true in states 0 and 1. We then get $0 \leq_a 1 \simeq_a 2 < 3$. This is also strange: instead of reinforcing belief in p, the $\neg p$-states have become more plausible instead!

This example demonstrates some issues with a quantified formulation of belief revision. Of course Aucher is aware of all these issues. See Aucher's PhD thesis (2008) for a quite novel way to perform higher-order and muIti-agent belief revision, based on plausibility relations among sets of formulas describing the structure in which the revision is executed.

3 Belief revision known as maximal-Spohn

When I first heard from Baltag and Smets' work on plausibility reasoning my first response was: "But this has all been done already! It's maximal-Spohn belief revision!" Aftel' some heavy internal combustion, I told Alexandru, who has his own response cycle, and this is all a long time ago. At the time I thought to remember specific phrasing in Spohn's well-known 'Ordinal Conditional Functions' (1988). But I never got down to be precise about

this souree of their work. Now lam. This section of my comments can be seen as yet another footnote to the extensive references and motivation of the authors' contribution. **In** their Example 5 they explain that for revision with single formulas all the following amount to more or less the same: anti-lexicographic update, lexicographic update, soft public update. To this list we can add yet another term: what Hans Rott (in his presentations and in Rott, 2006) and I call 'maximal-Spohn belief revision'.

Spohn introduces the 'simple conditional functions' (SCF) and the equivalent notion of 'well-ordered partitions' (WOP) to represent the extent of disbelief in propositions. **In** terms of Baltag and Smets, a WOP defines a totally ordered plausibility relation on the domain. Spohn then observes that such WOPs (plausibility relations) need to be updated when confronted with incoming new information in the form of a proposition A. **In** our terms A is the denotation of some revision formula φ. He then proceeds to discuss some specific plausibility updates. His presentation is based on ordinals $\alpha, \beta, \hat{\imath}, \cdots$ that label sets $E_\alpha, E_\beta, E_\gamma, \ldots$ of equally plausible states (all the E_α-states are more plausible than all the E{3-states, etc.). For a simplifying example, consider a partition of a domain W into a well-ordered partition $Eo, El, \ldots, E_6$. The set Eo are the most believed states, etc. Assume that a proposition A has non-empty intersection with E_4 and Es. Thus, the most plausible A-states are found in E_4. **If** we now also read 'state x is less plausible than state y' for 'world x is more disbelieved than world y' we are ready for an original quote from (Spohn, 1988), explaining two different ways to adjust $El, \ldots, E_6$ relative to A. A clear sign of a great writer is that one can take his work out of context but that it remains immediately inLelligible.

> A first proposal might be this: It seems plausibie to assume that, after information A is accepted, all the possible worlds in A are less disbelieved than the worlds in $\overline{A}$ (where $\overline{A}$ is the relative complement $W \setminus A$ of A). Further, it seems reasonable to assume that, by getting information only about A, the ordering of disbelief of the worlds within A remains unchanged, and likewise for the worlds in A. (Spohn, 1988, pp. 112-113)
>
> (...) the assumption that, after getting informed about A, all worlds in $\overline{A}$ are more disbelieved than all worlds in Aseems too strong. Certainly, the first member, i.e. the net content of the new WüP, must be a subset of A; thus at least some worlds in A must get less disbelieved than the worlds in $\overline{A}$. But it is utterly questionable whether even the most disbelieved world in A should get less disbelieved than even the least disbelieved world in $\overline{A}$; this could be effected at best by the most certain information.
>
> This last consideration suggests a second proposal. Perhaps one

> should put only the least disbelieved and not all worlds in A at the top of the new WOP (...). (Spohn, 1988, pp. 113-114)

The first proposal has become known as *maximal-Spohn*. The second proposal has become known as *minimal-Spohn*. Applied to the example partition the results are as follows; on purpose I use a fotmatting that is *veTY* similar to the displayed formulas on pages 113 and 114 in (Spohn, 1988). Further down in the sequence means less plausible.

$E_4 nA, E_S nA, Eo, El, E2, E3, E4\ nA, E_S$ nA maximal-Spohn
$E_4 \cap A$, *Eo, El,* $E_2, E_3, E_4 \cap \overline{A}$, *Es* minimal-Spohn

In maximal-Spohn, as in antilexicographic update, the A-states now come first, respecting the already existing plausibility distinctions among A-states, so that we start with E_4 n *A, Es* nA. The order among the non-A-states also remains the same (whether intersecting with A or not), thus we end with *Eo, El,* $E_2, E_3, E_4 \cap \overline{A}$, *Es* $\cap \overline{A}$. **In** minimal-Spohn, the states in *Es* are not affected by proposition A; only the equivalence class containing most plausible A-states is split in two, and only those most plausible A-states, namely E_4 n *A,* are shifted to the front of the line. These are now the most plausible states in the domain, such that A is now (in terms of Baltag and Smets again) unconditionally believed.

Aucher's plausibility update (Aucher, 2003), that we discussed in the previous section, implements a particular kind of 'minimal-Spohn' that also employs Spohn's ordinal conditional functions. We do not wish to discuss those here-things are quantitative enough as it is, already. Aucher's is not as minimal as it can be, e.g., I demonstrated the side-effect of merging plausibilities. **It** would be interesting to see a truly qualitative form of plausibility update that amounts to minimality in the Spohn-sense, or at least to something less maximal than anti-lexicographic update but equally intuitively convincing; but I do not know of one.

4 Well-preorders

In epistemic plausibility frames $(5, r\text{-}:\!\ll, \leq_a)_{a \in \mathcal{A}}$ the epistemic indistinguishability relations are equivalence relations and the plausibility relations are required to be well-preorders, i.e., refiexive and transitive relations where every non-empty subset has minimal elements. The non-empty-subset requirement ensures that something non-trivial is always conditionally believed. I will clarify some technical points concerning these primitives, to illustrate their richness for modelling purposes.

On the meaning of minimal. A weil-order is a total order where every non-empty subset has a least element, so by analogy a well-preorder should indeed be, as Baltag and Smets propose, a pre-order where every non-empty

subset has minimal elements. So far so good. Right? Wrong! Because we haven't seen the definition of minimal yet. I thought I did not need one. For me, *an element* is *minimal in an ordered set ij nothing* is *smaller:*

$$\mathrm{Min}(\leq_a, S) := \{s \mid t \in S \text{ and } t \leq_a \text{ simplies } t \simeq_a s\}.$$

Also, their first occurrence of 'minimal' before that definition was in a familiar setting, 'something like a weil-order', so that I did not bother about the precise meaning. And accidentally the definition of minimal was not just aftel' that first textual occurrence but even on the next page. So I had skipped that.

This was unwise. For Baltag and Smets, *an element* is *minimal in an ordered set ij eveTything* is *biggeT:*

$$\mathrm{Min}^{\mathrm{bs}}(\leq_a, S) = \{s \mid t \in S \text{ implies } s \leq_a t\}.$$

This is a powerful device, particularly as for partial orders some elements may not be related. The constraint that all non-empty sets have minima in their sense applies to two-element sets and thus enforces that such relations are connected orders (as explained in Footnote 4 of Baltag and Smets' text). So every well-preorder is a connected order. On connected orders the two definitions of minimal (Min'" and Min) coincide. We can further observe that the quotient relation $\leq_a/\simeq_a$ is a total order, and that it is also a well-order. Given a non-empty subset S' of $\leq_a/\simeq_a$, there is a non-empty subset S'' of $\leq_a$ such that $S'' \setminus \simeq_a = S'$, The $\simeq_a$-equivalence class of the minimal elements of S'' is the least element of S', The well-preorders of the authors are sometimes known as templated orders (Meyer et al., 2000). All this corresponds to Grove systems of spheres, as the authors rightly state.

Partial orders in belief revision. Partial orders that are not connected or not well-ordered according to the authors' definition *do* occur in belief revision settings. From now on I will only use 'minimal' in the standard sense.

Given one agent *a,* consider the frame consisting of five states {0, 1, 2, 3, 4}, all epistemically indistinguishable, and such that the relation $\leq_a$ is the transitive and refiexive closure of $0 \leq_a 1 < 4$ and $0 \leq_a 2 \leq_a 3 \leq_a 4$. It is a partial order, and every non-empty subset has minima. The reader can easily check this, for example, $\mathrm{Min}(\leq_a, \{0, 1,2, 3,4\}) = \{0\}$, $\mathrm{Min}(\leq_a, \{1, 2, 3\}) = \{1,2\} = \mathrm{Min}(\leq_a, \{1, 2\})$, and so on. If neither $s \leq_a t$ nor $t \leq_a s$, states s and *t* are called *incompamble.* States 1 and 2 are incomparable, as are 1 and 3.

Consider the symmetrie closure $\mathrm{Sy}(\leq_a)$ of a plausibility relation $\leq_a$ that is a partial order and where every non-empty subset has minima. Given a state s, we call a state t $\leq_a$ s *plausible* iff $(s, t) \in \mathrm{Sy}(\leq_a)$. Conditionalization

to implausible but epistemically possible states is clearly problematic. So as long as all states in an equivalence class are plausible regardless of the actual state in that class, we are out of trouble. This requirement states that the epistemic indistinguishability relation $\sim_a$ must be a refinement of $\mathrm{Sy}(\leq_a)$, or, differently said, that $\sim_a \cap \mathrm{Sy}(\leq_a) = \ <<$.

Incomparable states in a partial order can be 'compared in a way' aftel' all. Define $s \equiv_a t$ iff for all $u \in S$, $u. \leq_a s$ iff $u. \leq_a t$. Let's say that the agent is *indijJeTent* between s and t in that case. Clearly, equally plausible states are indifferent: $\simeq_a \subseteq \equiv_a$. But the agent is also indifferent between the incomparable states 1 and 2 in the above example. The quotient relation $\leq_a/\equiv_a$ is a total order. **In** belief contraction this identification of incomparabie objects in a preorder typically occurs between sets of formulas, not between semantic objects. See work on epistemic entrenchment involving templated orders, e.g., (Meyer et al., 2000).

Qualitative to quantitative. As already quoted by me above, the authors consider their approach 'in its spirit closer to qualitative logics than to approaches of a more quantitative ftavour.' Well-chosen wording, because-as the authors know-in its *ruiture* their approach is fairly quantitative aftel' all. Let us see why.

From a preorder where all non-empty subsets have minimal elements we can create degrees of plausibility as follows. Given that all sets have minimal elements, we give the $\leq_a$-minimal states of the entire domain S degree of plausibility 0. This set is non-empty. Now the entire domain minus the set of states with degree 0 also has a non-empty set of minimal elements. Again, this set exists. These are the states of degree 1. And so on. Write $\mathrm{Degree}^i(\leq_a)$ for the set of states of degree i. We now have:

$$\begin{array}{ll} \mathrm{Degree}^0(\leq_a) & \mathrm{Min}(\leq_a, S) \\ \mathrm{Degree}^{k+1}(\leq_a) & \mathrm{Min}(\leq_a, S \setminus \mathrm{Uj=O.k}\, \mathrm{Degree}^k(\leq_a)) \end{array}$$

Note the relation with the total order $\leq_a/\equiv_a$ introduced above.

Of course, I entirely agree that a qualitative *preseniatioti* of an epistemic plausibility framework is to be preferred over a quantitative representation. And-this is once again Alexandru Baltag providing an essential comment to the preliminary version of this commentary-although this comparison can be made on the structural level, the *language* of conditional doxastic logic is apparently not expressive enough to define degrees of belief, that use the above order. This matter is explained in their related publication (Baltag and Smets, 2006). But with that result one can wonder if a weaker structural framework, more qualitative in nature, would already have sufficed to obtain the same logical result.

It seems to me that the quest for the nature and the spirit of qualitative belief revision has not yet been ended. Other frameworks for belief revision,

such as the referenced work by Fenrong Liu, her **PhD** thesis (2008), and my own work (van Ditmarseh, 2005) (where the non-empty minima requirement is only for the entire domain, thus al!owing the real number interval [0,1 D, sometimes employ other partial orders and basic assumptions and mayalso contribute to this quest.

5 This is a lie

The authors' analysis of "lying about φ" involves an action plausibility model consisting of two actions Lie.,(φ) and True,(φ) with preconditions $\neg K_a\varphi$ and $K_a\varphi$ respectively. These actions can be distinguished by the lying agent, the speaker *a,* but are indistinguishable for the target, the listener *b.* Further, *b* considers it more plausible that *a* speaks the truth, than not: Truea(cp) $\leq_b$ Liea(cp). So 'agent *a* lies about φ' means that *a* armounces that φ is true, thus suggesting that she knows that φ is true, although *a* does in fact not know that. For convenience I am presenting this action as a dynamic operator that is part of the language (which can be justified as the authors do in Section 5).

In the authors' subsequent analysis it is explained how the contextual appearance of an action mayalso determines its meaning, both the context of states wherein the action may be executed and the context of states resulting from the action's execution. Again, 'lying about φ' makes for a fine example. **If** the listener balready knows that φ is false, the act of lying does not appeal' to *b* as the truth that φ, but as a lie that φ.

I have two observations to this analysis.

Lying and bluffing. **1** think that the precondition of *a* 'lying that φ' is not *ignomnce* of the truth, but *knowledge* to the contrary: the precondition of the action Liea(cp) should not be $\neg K_a\varphi$ but $K_a\neg\varphi$. **If** the precondition is $\neg K_a\varphi$ instead, I cal! this *bluffing,* not lying. As I am not a native speaker of English, and neither are the authors, this seems to be as good a moment as any to consult a dictionary (Merriam-Webster). To *blufJ* is "to cause to believe what is untrue." Whereas to *lie* is "to make a statement one knows to be untrue." **It** is further informative to read that the etymology for 'bluff' gives "probably from the Dutch *bluffen,* for 'to boast', 'to play a kind of card game." **It** is of course typical Anglo-Saxon prejudice that al! bad things concerning short-changing, scrooging, boasting, diseases, and unfair play ('Dutch book') are called Duteh. But let's not pursue that matter further. Given the action Truea(cp), that expresses the for *b* more plausible alternat.ive, I t.hink that its precondit.ion $K_a\varphi$ properly expresses Lhe part 'to cause to believe what is untrue'. On the other hand, given that the action Lie.,(φ) that is considered less plausible by *b,* the precondition $K_a\neg\varphi$ seems to express accurately 'to make a statement one knows to be untrue,'

and this condition is stronger than the precondition $\neg K_a\varphi$ suggested by Baltag and Smets.

When I proposed this commentary to the authors, Alexandru Baltag came with an interesting response: the precondition $\neg K_a\varphi$ of action $\mathrm{Lie}_a(\varphi)$ *also* involves 'to make a statement one knows to be untrue', namely the statement 'I know *ip',* In fact, *a* knows that she does not know φ. This is true. But a bit further-fetched, if I may. For me, the prototypical example of a lie remains the situation where, way back in time, my mother asks me if I washed my hands before dinner and I say: "Yes." Whereas when my grandfather held up his arm, with a closed fist obscuring a rubber (for Americans: eraser) and asked me: "What have I got in my hand?" and I then respond "A marbie!" he never accused me of being a liar. Or did he? I'd like to investigate these matters further. I am unaware of much work on lying in dynamic epistemics. For a setting involving only belief and not knowiedge, and public but not truthful announcements, see (van Ditmarsch et al., 2008).

Is contextual appearance relevant? I question the need for contextual appearances of actions. I make my point by resorting to lying, again. The authors say that the precondition of a lie is $\neg K_a\varphi$ but that, if the listener *b* already knows that φ is false, the act of lying no Ionger appears to *b* as the truth that φ, but as a lie that φ. I would be more inclined to strengthen the precondition for lying about φ from $\neg K_a\varphi$ to $\neg K_a\varphi \wedge \neg K_b\neg\varphi$. In which case there is no need for this contextual precondition.

Combining this with the previous I therefore think that the precondition of $\mathrm{Lie}_a(\varphi)$ should be $K_a\neg\varphi \wedge \neg K_b\neg\varphi$ rather than $\neg K_a\varphi$. And this is only the beginning of a more and more fine-grained analysis of lying, not the end. For example, it is reasonable to expect that the speaker is aware of the listener's ignorance about φ. That makes yet another precondition, namely $K_a\neg K_b\neg\varphi$. A distinction between knowledge and belief mayalso be important to model lying. The typical convention is to assume common belief that the speaker is knowledgeable about φ but the listener not, although in fact the speaker knows (or at least believes) the opposite of φ; so we get

$$K_a\neg\varphi \wedge CB_{ab}(B_b(K_a\varphi \vee K_a\neg\varphi) \wedge \neg K_b\varphi \wedge \neg K_b\neg\varphi)$$

where *CB* is the common belief operator. We cannot replace common belief by common knowledge in this expression. Then it would be inconsistent. (We can also replace all other K-operators in this expression by *B*-operators.) There are also truly multi-agent scenarios involving lying, where only the addressee is unaware of the truth about φ but other listeners in the audience may have a different communicative stance.

If this is only the beginning and not the end, why should there be an end at all? It is in fact unclear (as Alexandru Baltag also mentioned in

response to reading a version of these comments) if by incorporating more and more 'context' we finally have taken all possible contexts into account. Maybe there will always turn up yet another scenario that we might also want to incorporate in the precondition of lying. On the other hand-me again trying to have to last word-it seems that by employing infinitary operators in preconditions such as common knowledge and common belief, as above, we can already pretty weil take any kind of envisaged variation into account. So my cutrent bet is that the preconditions of contextual appearances (not the postconditional aspect) can be eliminated altogether.

I am detracting myself, and the reader. So let me stop here. Does this show that the authors' analysis of lying is ftawed? Not at all! In fact it is very weil chosen, as it is a very rich speech act with many hidden aspects that are crying aloud for analysis, and the authors' framework of doxastic actions is the obvious and very suitable formalization for such an analysis. Also, different arguments than the above can be put forward, in support of $\neg K_a\varphi$ as precondition of $\mathrm{Lie}_a(\varphi)$ instead of my preferred $K_a\neg\varphi$. Let me therefore conclude by complimenting the authors again on their rich contribution, and hope for more from this productive duo.

References

Aucher, G. (2003). A Combined System for Update Logic and Belief Revision. Master's thesis, ILLC, University of Amsterdam, Amsterdam, The Netherlands. *ILLC Fublications* MoL-2003-03.

Aucher, G. (2008). *Perspeciioes on Belief and Change.* PhD thesis, University of Otago, New Zealand & Institut de Recherche en Informatique de Toulouse, France.

Baltag, A. & Smets, S. (2006). Dynamic Belief Revision over Multi-Agent Plausibility Modeis. In Bonanno, *G.,* van der Hoek, W. & Wooldridge, M., eds., *Proceedinqs of the 'lih Conference on Logic and the Foundations of Game and Decision Theoru (LOFT'06),* pp. 11-24.

van Ditmarsch, H. (2005). Prolegomena to Dynamic Logic for Belief Revision. *Synthese, 147(2):229-275.*

van Ditmarsch, H., van Eijck, .1., Sietsma, F. & Wang, Y. (2008). On the Logic of Lying. Under submission.

Liu, F. (2008). *Changing fOT the Better. Preferenee Dynamics and Agent Dioersiiç.* PhD thesis, University of Amsterdam. *ILLC Fublications* DS-2008-02.

Meyer, T.A., Labuschagne, W.A. & Heidema, .1. (2000). Refined Epistemic Entrenchment. *Joumal of Logic, Language, and Ïnformaiion; 9(2):237-259.*

Rott, H. (2006). Shifting Priorities: Simple Representations for Twenty-seven Iterated Theory Change Operators. **In** Lagerlund, H., Lindström, S. & Sliwinski, R., eds., *Modality Matters: Twenty-Five Essays in Honour of Krister Seqerberq,* Uppsala Philosophical Studies Volume 53, pp. 359-384. Uppsala Universitet.

Spohn, W. (1988). Ordinal Conditional Functions: ADynamie Theory of Epistemic States. **In** Harper, W.L. & Skyrms, B., eds., *Causation in Decision, Belief Change and Statistics,* Vol. 42 of *The Westem Ontario Series in Philosophy of Science,* pp. 105-134. Springer.

Belief Revision In a Temporal Framework

Giacomo Bonanno

Department of Economics
University of California
Davis CA 95616-8578, United States of America
gfbonanno@ucdavis.edu

Abstract

We study a branching-time temporallogic of belief revision where the interaction of belief and information is modeled explicitly. The logic is based on three modal operators: a belief operator, an information operator and a next-time operator. We consider three logics of increasing strength. The first captures the most basic notion of minimal belief revision. The second characterizes the qualitative content of Bayes' rule. The third is the logic proposed in (Bonanno, 2üü7a), where some aspects of its relationship with the AGNI theory of belief revision were investigated. We further explore the relationship to AG *NI* with the help of semantic structures that have been used in the rational choice literature. Further strengthenings of the logic are also investigated.

1 Introduction

Since the foundational work of Alchourrón, Gärdenfors and Makinson (1985), the theory of belief revision has been a very active area of research. Recently several authors have been attempting to re-cast belief revision within a modal framework. Pioneering work in this new area was done by Segerberg (1995, 1999) in the context of dynarnic doxastic logic, Board (2002) in the context of multi-agent doxastic logic and van Benthem (2004) in the context of dynarnic epistemic logic. Much progress has been made both in dynarnic episternic logic (see, for example, Baltag and Smets, 2006; van Ditmarseh, 2005; van Ditmarsch and Labuschagne, 2007 and the recent survey in van Ditmarsch et al., 2007) as weil as in dynamic doxastic logic (see Leitgeb and Segerberg, 2007). Another very active area of research has been iterated belief revision (see, for example, Boutilier, 1996; Darwiche and Pearl, 1997; Nayak et al., 2003; Rott, 2006).

This paper joins the recent attempts to establish a qualitative view of belief revision in a modal framework, by continuing the study of belief revision within a temporal framework that was first proposed in (Bonanno, 2007a). Since belief revision deals with the interaction of belief and information over

Krzysztof R. Apt, Robert van Rooij (eds.). *New Perspectives on Games and Interaction.* Texts in Logic and Games 4, Amsterdam University Press 2008, pp. 45-79.

time, branching-time temporal logic seems a natural setting for a theory of belief change. On the semantic side we consider branching-time frames with the addition of a belief relation and an information relation for every instant *t*. We thus extend to a temporal setting the standard (Kripke, 1963) semantics used in the theory of statie belief pioneered by Hintikka (1962). On the syntactic side we consider a propositionallanguage with a next-time operator, a belief operator and an information operator. Three logies of increasing strength are studied. The first is a logic that expresses the most basic notion of minimal belief revision. The second captures the qualitative content of Bayes' rule, thus generalizing the two-date result of (Bonanno, 2005) to a branching-time framework. The third logic is the logic proposed in (Bonanno, 2007a), where some aspects of the relationship between that logic and the AGM theory of belief revision were investigated. **In** this paper we provide frame characterization results for al! three logies and we further investigate the relationship between the strongest of the three logies and the notion of AGM belief revision functions. We do so with the help of semantic structures that have been used in the rational choice literature. We cal! these structures one-stage revision frames and show that there is a correspondence between the set of one-stage revision frames and the set of AGM belief revision functions. Further strengthening of the logic are also investigated.

While the structures that we consider accommodate iterated belief revision in a natural way, we do not attempt to axiomatize iterated revision in this paper. First steps in this direction have been taken in (Zvesper, 2007).

We provide frame characterization results and do not address the issue of cornpleteness of our logics. Cornpleteness of Lhe basic logic wiLh respect to a more general class of temporal belief revision frames (where the set of states is al!owed to change over time) is proved in (Bonanno, 2008); that result has been extended in (Zvesper, 2007) to the set of frames considered in this paper.

2 Temporal belief revision frames

We consider the semantic frames introduced in (Bonanno, 2007a), which are branching-time structures with the addition of a belief relation and an information relation for every instant *t*.

A *next-time bmnching frame* is a pair $(T, \rightarrowtail)$ where T is a non-empty, countable set of instants and $\rightarrowtail$ is a binary relation on T satisfying the fol!owing properties: $\forall t_1, t_2, t_3 \in T$,

(1) backward uniqueness — if $t_1 \rightarrowtail t_3$ and $t_2 \rightarrowtail t_3$ then $t_1 = t_2$

(2) acyclicity — if $(t_1, \ldots, t_n)$ is a sequence with $t_i \rightarrowtail t_{i+1}$ for every $i = 1, \ldots, n-1$, then $t_n \neq t_1$.

The interpretation of $t_1 \rightarrowtail t_2$ is that t_2 is an *immediate successar* 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.

Definition 2.1. A *temporal belief reoisiori frame* is a quadruple $(T, \rightarrowtail, \Omega, \{\mathcal{B}_t, \mathcal{I}_t\}_{t \in T})$ where $(T, \rightarrowtail)$ is a next-time branching frame, Ω is a non-empty set of states (or possible worlds) and, for every $t \in T$, $\mathcal{B}_t$ and T, are binary relations on Ω.

The interpretation of *wBtw'* is that at state wand time *t* the individual considers state w' possible (an alternative expression is "w' is a doxastic alternative to *w* at time *t"),* while the interpretation of $wItw'$ is that at state wand time *t,* according to the information received, it is possible that the true state is w'. We shal! use the fol!owing notation:

$$Bt(w) = \{w' \in \Omega : wBtw'\} \text{ and, similarly, } It(w) = \{w' \in \Omega : wItw'\}.$$

Figure 1 il!ustrates a temporal belief revision frame. For simplicity, in al! the figures we assume that the information relations I_t are equivalence relations (whose equivalence classes are denoted by rectangles) and the belief relations $\mathcal{B}_t$ are serial, transitive and euclidean' (we represent this fact by enclosing states in ovals and, within an equivalence class for I_t, we have that-for every two states wand w_1---$w_1 \in Bt(w)$ if and only if w_1 belongs to an oval)." For example, in Figure 1 we have that $\mathcal{I}_{t_1}(\gamma) = \{\alpha, \beta, \gamma\}$ and $\mathcal{B}_{t_1}(\gamma) = \{\alpha, \beta\}$.

Temporal belief revision frames 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 cal!ed belief revision, while the latter is cal!ed belief update (see Katsuno and Mendelzon, 1991). We shal! focus on belief revision.

On the syntactic side we consider a propositional language with five modal operators: the next-time operator $\bigcirc$ and its inverse $\bigcirc^{-1}$, the belief operator *B,* the information operator land the "al! state" operator *A.* The intended interpretation is as fol!ows:

$\bigcirc\phi$:	"at every next instant it wil! be the case that ϕ"
$\bigcirc^{-1}\phi$:	"at every previous instant it was 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".

1 *Bt* is serlal if, $\forall \omega \in \Omega$, $Bt(w) \neq 0$; it is transitive if $w' \in Bt(w)$ implies that $Bt(w') \subseteq Bt(w)$; it is euclidean if $w' \in Bt(w)$ implies that $Bt(w) \subseteq Bt(w')$.

2 Note, however, that our results do *not* require *T,* to be an equivalence relation, nor do they require *Bt* to be serial, transitive and euclidean.

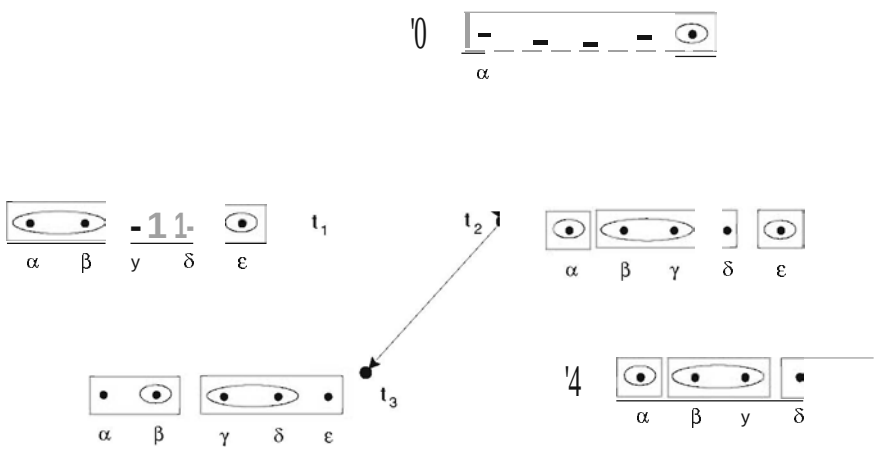


FIGURE 1.

The "all state" operator A is needed in order to eapture the non-normality ofthe information operator I (see below). For a thorough discussion ofthe "all state" operator see (Goranko and Passy, 1992).

Note that, while the other operators apply to arbitrary formulas, *we restriet the informaiion operator to apply to Boolean [ormulas only,* that is, to formulas that do not eontain modal operators. Boolean formulas are defined reeursively as follows: (1) every atomie proposition is a Boolean formula, and (2) if ϕ and ψ are Boolean formulas then so are $\neg\phi$ and $(\phi \mathrm{V} \psi)$. The set of Boolean formulas is denoted by Φ^B. Boolean formulas represent faets and, therefore, we restriet information to be about faets.[3]

Given a temporal belief revision frame $(T, \rightarrowtail, \Omega, \{B_t, \mathcal{I}_t\}_{t\in T})$ one obtains a *model based on it* by adding a function $V : S \to 2^{\Omega}$ (where S is the set of atomie propositions and 2^{Ω} denotes the set of subsets of Ω) that assoeiates with every atomie proposition p the set of states at whieh p is true. Note that defining a valuation this way is what frames the problem as one of belief revision, sinee the truth value of an atomie proposition p depends only on the state and not on the time." Given a model, a state w, an instant $t$and a formula ϕ, we write $(w, t) \models \phi$ to denote that ϕ is true at state $w$and time t. Let $\|\phi\|$ denote the truth set of ϕ, that is, $\|\phi\| = \{(w, t) \in \Omega \times T : (w, t) \models \phi\}$ and let $\lceil\phi\rceil_t \subseteq \Omega$ denote the set of states at whieh ϕ is true at time t, that is, $\lceil\phi\rceil_t = \{w \in \Omega : (w,t) \models \phi\}$. Truth of an arbitrary formula at a pair (w, t) is defined reeursively as follows:

3 Zvesper (2007) has recently proposed a version of our logic where the restrietion to Boolean forrnulas is dropped.

4 Belief update would require a valuation to be defined as a function $V : S \to 2^{\Omega\times T}$.

if p ES,	$(w,t) \models p$ if and only if wE $V(p)$;
$(w, t) \models \neg\phi$	if and only if $(w,t) \not\models \phi$;
$(w,t) \models \phi \vee \psi$	if and only if either $(w,t) \models \phi$ or $(w, t) \models \psi$ (or both);
$(w, t) \models \bigcirc\phi$	if and only if $(w,ti) \models \phi$ for every ti such that $t \rightarrowtail ti$;
$(w, t) \models \bigcirc^{-1}\phi$	if and only if $(w,til) \models \phi$ for every til such that $til \rightarrowtail t$;
$(w, t) \models B\phi$	if and only if $Bt(w) \subseteq \lceil\phi\rceil_t$, that is, if $(Wl, t) \models \phi$ for all io' E $Bt(w)$;
$(w,t) \models I\phi$	if and only if $It(w) = \lceil\phi\rceil_t$, that is, if (1) $(Wl, t) \models \phi$ for all wl E $It(w)$, and (2) if $(wl, t) \models \phi$ then io' E $It(w)$;
$(w, t) \models A\phi$	if and only if $\lceil\phi\rceil_t = \Omega$, that is, if $(Wl, t) \models \phi$ for all io' E Ω.

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 $It(w) \subseteq \lceil\phi\rceil_t$ we require equality: $It(w) = \lceil\phi\rceil_t$. Thus our information operator is formally similar to the "all and only" operator introduced in (Humberstone, 1987) and the "only knowing" operator studied in (Levesque, 1990), although the interpretation is different. It is also similar to the "assumption" operator used by Brandenburger and Keisier (2006).

Remark 2.2. The truth value of a Boolean formula does not change over time: it is only a function of the state. That is, fix an arbitrary model and suppose that $(w,t) \models \phi$ where ϕ E Φ^B; then, for every ti E T, $(w,ti) \models \phi$ (for a proof see Bonanno, 2007a, p. 148).

A formula ϕ is *valid in a model* if $\|\phi\| = \Omega$ x T, that is, if ϕ is true at every state-instant pair (w,t). A formula ϕ is *valid in a frame* if it is valid in every model based on it.

3 The basic logie

The formal language is built in the usual way (see Blackburn et al., 2001) from a countable set of atomie propositions, the connectives -, and V (from which the connectives A, $\rightarrow$ and $\leftrightarrow$ are defined as usual) and the modal operators 0, O-^{1}, B, 1 and A, with the restriction that $I\phi$ is a well-formed formula if and only if ϕ is a Boolean formula. Let $\Diamond\phi \stackrel{\text{def}}{=}$ -,0 $\neg\phi$, and $\Diamond^{-1}\phi \stackrel{\text{def}}{=}$ -,O-^{1} $\neg\phi$. Thus the interpretation of $\Diamond\phi$ is "at *some* next instant it will be the case that ϕ" while the interpretation of $\Diamond$-1ϕ is "at some immediately preceding instant it was the case that ϕ".

Wc dcnotc by $\mathbb{L}_a$ thc basic logic dcfincd by thc following axioms and rules of inference.

Axioms:

1. All propositional tautologies.
2. Axiom K for $\bigcirc, \bigcirc^{-1}$, B and A:[5] for $D \in \{\bigcirc, \bigcirc^{-1}, B, A\}$:
$$(\Box\phi \wedge \Box(\phi \to \psi)) \to \Box\psi \qquad (K)$$
3. Temporal axioms relating $\bigcirc$ and $\bigcirc^{-1}$:
$$\phi \to \bigcirc\Diamond^{-1}\phi$$
$$\phi \to \bigcirc^{-1}\Diamond\phi$$
4. Backward Uniqueness axiom:
$$\Diamond^{-1}\phi \to \bigcirc^{-1}\phi \qquad (BU)$$
5. 85 axioms for A:
$$A\phi \to \phi$$
$$\neg A\phi \to A\neg A\phi$$
6. Inclusion axiom for B (note the absence of an analogous axiom for 1):
$$A\phi \to B\phi \qquad (IndB)$$
7. Axioms to capture the non-standard semantics for I: for $\phi, \psi \in \Phi^B$ (recall that Φ^B denotes the set of Boolean formulas),
$$(I\phi \wedge I\psi) \to A(\phi \leftrightarrow \psi)$$
$$A(\phi \leftrightarrow \psi) \to (I\phi \leftrightarrow I\psi)$$

Rules of Inference:

1. Modus Ponens:
$$\frac{\phi, \quad \phi \to \psi}{\psi} \qquad (MP)$$
2. Necessitation for A, $\bigcirc$ and $\bigcirc^{-1}$:
$$\text{For every } D \in \{A, \bigcirc, \bigcirc^{-1}\}, \ \frac{\phi}{\Box\phi}. \qquad (Nec)$$

[5] Axiom K for I is superftuous, since it can be derived from axioms 1_1 and 1_2 below (see Bonanno, 2005, p. 204).

Note that from MP, Inclj, and Necessitation for A one can derive necessitation for B. On the other hand, necessitation for J is *noi* a rule of inference of this logic (indeed it is not validity preserving).

Remark 3.1. By MP, axiom K and Necessitation, the following is a derived rule of inference for the operators $\bigcirc$ „ $\bigcirc^{-1}$ Band A: $\frac{\phi \to \psi}{\bigcirc\phi \to \Box\psi}$ for D E $\{O,O\text{-}l,B,A\}$. We call this rule RK. On the other hand, rule **RK** is not avalid rule of inference for the operator J.

4 The weakest logic of belief revision

Our purpose is to model how the beliefs of an individual change over time in response to *factual* information. Thus *the axioms we introduce are resiricted to Boolean [ormulas,* which are formulas that do not contain any modal operators.

We shall consider axioms of increasing strength that capture the notion of minimal change of beliefs.

The first axiom says that if ϕ and ψ are facts (Boolean formulas) and-currently-the agent believes that ϕ and also believes that ψ and his belief that ϕ is non-trivial (in the sense that he considers ϕ possible) then-at every next instant-if he is informed that ϕ it will still be the case that he believes that.é. That is, if at a next instant he is informed of some fact that he currently believes non trivially, then he *cannot drop* any of his cutrent factual beliefs ('W' stands for 'Weak' and 'ND' for 'No Drop'):6 if ϕ and ψ are Boolean,

$$(B\phi \wedge \neg B\neg\phi \wedge B\psi) \to \bigcirc(I\phi \to B\psi). \qquad \text{(WND)}$$

The second axiom says that if ϕ and ψ are facts (Boolean formulas) and-currently-the agent believes that ϕ and does not believe that ψ, then-at every next instant-if he is informed that ϕ it will still be the case that he does not believe that ψ. That is, at any next instant at which he is informed of some fact that he currently believes he *cannot add* a factual belief that he does not currently have ('W' stands for 'Weak' and 'NA' stands for 'No Add'):" ϕ and ψ are Boolean,

$$(B\phi \wedge \neg B\psi) \to \bigcirc(I\phi \to \neg B\psi). \qquad \text{(WNA)}$$

6 lt is shown in the appendix that the following axiom (which says that if the individual is informed of some fact that he believed non-trivially at a previous instant then he must continue to believe every fact that he believed at that time) is equivalent to WND: if ϕ and ψ are Boolean,

$$O\text{-}^{1}(B\phi \wedge B\psi \wedge \neg B\neg\phi) \wedge I\phi \to B\psi.$$

This, in turn, is propositionally equivalent to $\bigcirc^{-1}(B\phi \wedge B\psi \wedge \neg B\neg\phi) \to (I\phi \to B\psi)$.

7 lt is shown in the appendix that the following is an equivalent formulation of WNA: if ϕ and ψ are Boolean,

$$\Diamond^{-1}(B\phi \wedge \neg B\psi) \wedge I\phi \to \neg B\psi.$$

Thus, by WND, no belief can be dropped and, by WNA, no belief can be added, at any next instant at which the individual is informed of a fact that he currently believes.

An axiom is *characterized by* (or *characterizes*) a property of frames if it is valid in a frame if and only if the frame satisfies that property.

Ali the propositions are proved in the appendix.

Proposition 4.1.

(1) The axiom WND is characterized by the foliowing property: $\forall \omega \in \Omega$, $\forall t_1, t2 \in T$,

$$\text{if } t1 \rightarrowtail t2, B_{t1}(w) \neq 0 \text{ and } B_{t1}(w) < 1_{\tau 2}(w) \text{ then } B_{t2}(w) < \eta_{,}(w). \qquad (P_{WNO})$$

(2) The axiom WN A is characterized by the foliowing property: $\forall \omega \in \Omega$, $\forall t_1, t\text{-:} \in T$,

$$\text{if } t_1 \rightarrowtail t_2 \text{ and } \mathcal{B}_{t_1}(\omega) \subseteq \mathcal{I}_{t_2}(\omega) \text{ then } \mathcal{B}_{t_1}(\omega) \subseteq \mathcal{B}_{t_2}(\omega). \qquad (P_{WNA})$$

Let lLw (where 'W' stands for 'Weak') be the logic obtained by adding WND and WNA to lL_a. We denote this by writing lLw = lL_a + WNA + WND. The foliowing is a coroliary of Proposition 4.1.

Corollary 4.2. The logic lLw is characterized by the class of temporal belief revision frames that satisfy the foliowing property: $\forall \omega \in \Omega$, $\forall t_1, t\text{-:} \in T$,

$$\text{if } t_l \rightarrowtail t2, B_{t1}(w) \neq 0 \text{ and } B_{t1}(w) \subseteq 1_{\tau 2}(w) \text{ then } B_{t1}(w) = B_{t2}(w).$$

The frame of Figure 1 violates the property of Coroliary 4.2, since $t\text{-:} \rightarrow t3$, $\mathcal{B}_{t_2}(\alpha) = \{\alpha\} \subseteq 1_{\tau 3}(a) = \{\alpha, \beta\}$ and $\mathcal{B}_{t_3}(\alpha) = \{\beta\} \neq \mathcal{B}_{t_2}(\alpha)$.

The logic lLw captures a weak notion of minimal change of beliefs in that it requires the agent not to change his beliefs if he is informed of some fact that he already believes. This requirement is stated explicitly in the foliowing axiom ('WNC' stands for 'Weak No Change'): if ϕ and ψ are Boolean formulas,

$$(I\phi \wedge \Diamond^{-1}(B\phi \wedge \neg B\neg\phi)) \rightarrow (B\psi \leftrightarrow \Diamond^{-1}B\psi). \qquad \text{(WNC)}$$

WNC says that if the agent is informed of something that he believed non-trivialiy in the immediately preceding past, then he now believes a fact if and only if he believed it then.

Proposition 4.3. WNC is a theorem of lL_w.

We now turn to a strengthening of lL_w.

5 The logic of the Qualitative Bayes Rule

The logic $\mathsf{IL_W}$ imposes no restrictions on belief revision whenever the individual is informed of some fact that he did not previously believe. We now consider a stronger logic than $\mathsf{IL_W}$. The foliowing axiom strengthens WND by requiring the individual not to drop any of his cutrent factual beliefs at any next instant at which he is informed of some fact that he currently *conciders possible* (without necessarily believing it: the condition $B\phi$ in the antecedent of WND is dropped): if ϕ and ψ are Boolean,

$$(\neg B\neg\phi \wedge B\psi) \rightarrow \bigcirc(I\phi \rightarrow B\psi). \tag{ND}$$

The cortesponding strengthening of WN A requires that if the individual considers it possible that $(\phi \wedge \neg\psi)$ then at any next instant at which he is informed that ϕ he does not believe that ψ:[8] if ϕ and ψ are Boolean,

$$\neg B\neg(\phi \wedge \neg\psi) \rightarrow \bigcirc(I\phi \rightarrow \neg B\psi). \tag{NA}$$

One of the axioms of the AGM theory of belief revision (see Gärdenfors, 1988) is that information is believed. Such axiom is often referred to as "Success" or "Acceptance". The foliowing axiom is a weaker form of it: information is believed when it is not surprising. If the agent considers a fact ϕ possible, then he wili believe ϕ at any next instant at which he is informed that ϕ. We cali this axiom *Qualified Acceptance* (QA): if ϕ is Boolean,

$$\neg B\neg\phi \rightarrow \bigcirc(I\phi \rightarrow B\phi). \tag{QA}$$

Proposition 5.1.

(1) The axiom ND is characterized by the foliowing property: $\forall\omega \in \Omega$, $\forall t_1, t_2 \in \mathsf{ET}$,

$$\text{if } t_1 \rightarrowtail t_2 \text{ and } \mathcal{B}_{t_1}(\omega) \cap \mathcal{I}_{t_2}(\omega) \neq \varnothing \text{ then } \mathcal{B}_{t_2}(\omega) \subseteq \mathcal{B}_{t_1}(\omega). \quad (\mathrm{P_{ND}})$$

(2) The axiom NA is characterized by the foliowing property: $\forall\omega \in \Omega$, $\forall t_1, t_2 \in \mathsf{ET}$,

$$\text{if } t_1 \rightarrowtail t_2 \text{ then } \mathcal{B}_{t_1}(\omega) \cap \mathcal{I}_{t_2}(\omega) \subseteq \mathcal{B}_{t_2}(\omega).$$

8 Axiom NA can alternatively be written as $\Diamond(I\phi \wedge B\psi) \rightarrow B(\phi \rightarrow \psi)$, which says that if there is a next instant at which the individual is informed that ϕ and believes that ψ, then he must now believe that whenever ϕ is the case then ψ is the case. Another, propositionally equivalent, formulation of NA is the following: $\neg B(\phi \rightarrow \psi) \rightarrow \bigcirc(I\phi \rightarrow \neg B\psi)$, which says that if the individual does not believe that whenever ϕ is the case then ψ is the case, then-at any next instant-if he is informed that ϕ then he cannot believe that ψ.

(3) The axiom QA is characterized by the foliowing property: $\forall \omega \in \Omega$, $\forall t_1, t_2 \in T$,

$$\text{if } t_1 \rightarrowtail t_2 \text{ and } \mathcal{B}_{t_1}(\omega) \cap \mathcal{I}_{t_2}(\omega) \neq \varnothing \text{ then } \mathcal{B}_{t_2}(\omega) \subseteq \mathcal{I}_{t_2}(\omega). \quad (\mathrm{P_{QA}})$$

We cali the foliowing property of temporal belief revision frames "Qualitative Bayes Rule" (QBR): $\forall t_1, t_2 \in T, \forall \omega \in \Omega$,

$$\text{if } t_1 \rightarrowtail t_2 \text{ and } \mathcal{B}_{t_1}(\omega) \cap \mathcal{I}_{t_2}(\omega) \neq \varnothing \text{ then } \mathcal{B}_{t_2}(\omega) = \mathcal{B}_{t_1}(\omega) \cap \mathcal{I}_{t_2}(\omega). \quad (\mathrm{QBR})$$

The expression "Qualitative Bayes Rule" is motivated by the foliowing observation (see Bonanno, 2005). **In** a probabilistic setting, let *Pw,t,* be the probability measure over a set of states Ω representing the individual's beliefs at state wand time t_1; let $F \subseteq \Omega$ be an event representing the information received by the individual at a later date t_2 and let P_{ω,t_2} be the posterior probability measure representing the revised beliefs at state wand date *t2.* Bayes' rule requires that, if *Pw,t,(F)* > 0, then, for every event $E < \Omega$, $P_{w,t2}(E) = \frac{P_{\omega,t}(E \cap F)}{P_{w,t}(F)}$. Bayes' rule thus implies the foliowing (where supp(P) denotes the support of the probability measure *P):*

$$\text{if } \mathrm{supp(Pw,t,)} \cap F \neq \mathrm{o}, \text{ then } \mathrm{supp}(P_{\omega,t_2}) = \mathrm{supp(Pw,t,)} \cap \mathrm{F}.$$

If we set *Bt,(w)* = supp(Pw,t,), $F = \mathcal{I}_{t_2}(\omega)$ (with *t-:* $\rightarrowtail$ *t2)* and $B_{t2(w)}$ = sUPP($P_{w,t2}$) then we get the Qualitative Bayes Rule as stated above. Thus in a probabilistic setting the proposition "at date *t* the individual believes ϕ" would be interpreted as "the individual assigns probability 1 to the event $[\phi]_t < \Omega$".

The foliowing is a coroliary of Proposition 5.1.

Corollary 5.2. The conjunction ofaxioms ND, NA and QA characterizes the Qualitative Bayes Rule.

The frame of Figure 1 violates QBR, since $t_2 \rightarrow t_3$, $B_{t2}(\delta) = \{\beta, \gamma\}$ and $\mathcal{I}_{t_3}(\delta) = \{\gamma, \delta, \varepsilon\}$, so that $\mathcal{B}_{t_2}(\delta) \cap \mathcal{I}_{t_3}(\delta) = \{\gamma\} \neq \mathrm{0}$; however, $\mathcal{B}_{t_3}(\delta) = \{\gamma, \delta\} \neq B_{t2}(\delta) \cap \mathcal{I}_{t_3}(\delta)$. On the other hand, the frame of Figure 2 does satisfy QBR.

Definition 5.3. Let $\mathbb{L}_{\mathrm{QBR}} = \mathbb{L}_{\mathrm{a}} + \mathrm{ND} + \mathrm{NA} + \mathrm{QA}$.

Remark 5.4. The logic $\mathbb{L}_{\mathrm{QBR}}$ contains (is a strengthening of) $\mathbb{L}_{\mathrm{W}}$. **In** fact, WND is a theorem of the logic $\mathbb{L}_{\mathrm{a}} + ND$, since $(B\phi \wedge \neg B\neg\phi \wedge B\psi) \rightarrow (\neg B\neg\phi \wedge B\psi)$ is a tautology, and WN A is a theorem of the logic $\mathbb{L}_{\mathrm{a}} + NA$ (see the appendix).

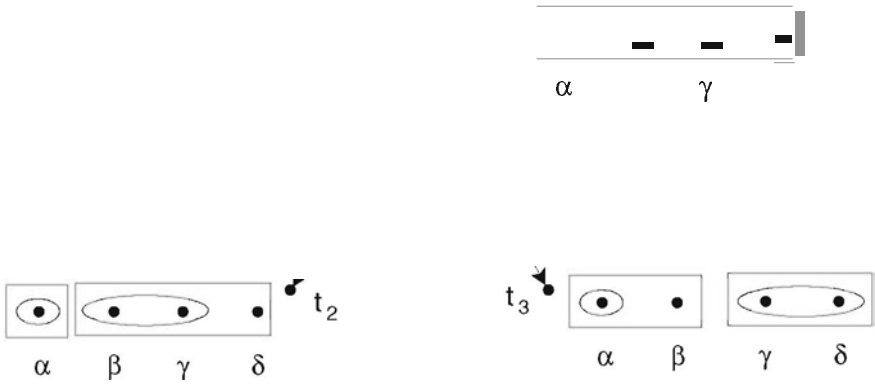


FIGURE 2.

6 The logic of AGM

We now strengthen logic $\mathbb{L}_{\mathrm{QBR}}$ by adding four more axioms.

The first axiom is the Acceptance axiom, which is a strengthening of Qualified Acceptance: if ϕ is Boolean,

$$I\phi \rightarrow B\phi. \tag{A}$$

The second axiom says that if there is a next instant where the individual is informed that $\phi \mathbin{!\!\backslash\!.} \psi$ and believes that X, then at every next instant it must be the case that if the individual is informed that ϕ then he must believe that $(\phi \mathbin{!\!\backslash\!.} \psi) \rightarrow X$ (we call this axiom K7 because it corresponds to AGM postulate $(\circledast 7)$: see the next section): if ϕ, ψ and X are Boolean formulas,

$$\Diamond(I(\phi \mathbin{!\!\backslash\!.} \psi) \mathbin{!\!\backslash\!.} BX) \rightarrow \bigcirc(I\phi \rightarrow B((\phi \mathbin{!\!\backslash\!.} \psi) \rightarrow X)). \tag{K7}$$

The third axiom says that if there is a next instant where the individual is informed that ϕ, considers $\phi \mathbin{!\!\backslash\!.} \psi$ possible and believes that $\psi \rightarrow X$, then at every next instant it must be the case that if the individual is informed that $\phi \mathbin{!\!\backslash\!.} \psi$ then he believes that X (we call this axiom K8 because it corresponds to AGM postulate $(\circledast 8)$: see the next section): if ϕ, ψ and X are Boolean formulas,

$$\Diamond(I\phi \mathbin{!\!\backslash\!.} \neg B\neg(\phi \mathbin{!\!\backslash\!.} \psi) \mathbin{!\!\backslash\!.} B(\psi \rightarrow X)) \rightarrow \bigcirc(I(\phi \mathbin{!\!\backslash\!.} \psi) \rightarrow BX). \tag{K8}$$

The Iourt.h axiorn says Lhat if Lhe individual recei ves consistent information then his beliefs are consistent, in the sense that he does not simultaneously believe a formula and its negation ('WC' stands for 'Weak Consistency'): if ϕ is a Boolean formula,

$$(I\phi \wedge \neg A\neg\phi) \rightarrow (B\psi \rightarrow \neg B\neg\psi). \qquad \text{(WC)}$$

Proposition 6.1.

(1) The axiom (A) is characterized by the foliowing property: $\forall\omega$ E Ω, $\forall t$ E T,

$$\mathcal{B}_t(\omega) \subseteq \mathcal{I}_t(\omega).$$

(2) The axiom *(K7)* is characterized by the foliowing property: $\forall\omega$ E Ω, $\forall t_1, t2, t3$ E T,

$$\text{if } t_1 \rightarrowtail t2,\ t\,l \rightarrowtail t3 \text{ and } \mathcal{I}_{t_3}(\omega) \subseteq \mathcal{I}_{t_2}(\omega) \text{ then } \mathcal{I}_{t_3}(w) \cap B_{t2(w)} < B_{t3(w)}. \qquad (\mathrm{P_{K7}})$$

(3) The axiom *(K8)* is characterized by the foliowing property: $\forall\omega$ E Ω, $\forall t_1, t2, t3$ E T,

$$\text{if } t_1 \rightarrowtail t2, t\,l \rightarrowtail t3, \mathcal{I}_{t_3}(\omega) \subseteq \mathcal{I}_{t_2}(\omega) \text{ and } \mathcal{I}_{t_3}(\omega)\, nB_{t2(w)} \neq 0 \text{ then } B_{t3(w)} < \mathcal{I}_{t_3}(\omega)\, nB_{t2(w)}. \qquad (\mathrm{P_{KS}})$$

(4) The axiom (WC) is characterized by the foliowing property: $\forall\omega$ E Ω, $\forall t$ E T,

$$\text{if } \mathcal{I}_t(\omega) \neq \varnothing \text{ then } \mathcal{B}_t(\omega) \neq \varnothing. \qquad (\mathrm{P_{WC}})$$

Let $\mathbb{L}_{\mathrm{AGM}} = \mathbb{L}_a$ + A + ND + NA + K7 + K8 + Wc. Since QA can be derived from *A*, logic $\mathbb{L}_{\mathrm{AGM}}$ contains (is a strengthening of) logic $\mathbb{L}_{\mathrm{QBR}}$.

Definition 6.2. An lLAGlvI-fmme is a temporal belief revision frame that satisfies the foliowing properties:

(1) the Qualitative Bayes Rule,

(2) $\forall\omega$ E Ω, $\forall t$ E T, $Bt(w) \subseteq \mathcal{I}_t(\omega)$,

(3) $\forall\omega$ E Ω, $\forall t$ E T, if $\mathcal{I}_t(\omega) \neq 0$ then $Bt(w) \neq$ o,

(4) $\forall\omega \in \Omega,\ \forall t_1, t_2, t_3 \in T,$

$$\text{if } t_1 \rightarrowtail t2,\ t\,l \rightarrowtail t3,\ \mathcal{I}_{t_3}(\omega) \subseteq \mathcal{I}_{t_2}(\omega) \text{ and } \mathcal{I}_{t_3}(\omega)\, nB_{t2(w)} \neq 0 \text{ then } B_{t3(w)} = \mathcal{I}_{t_3}(w) \cap B_{t2(w)}.$$

An *lLAGlvI-model* is a model based on an lLAGlvI-frame.

The frame of Figure 2 is not an AGM frame, although it satisfies QBR. **In** fact, we have that $t_1 \rightarrowtail t_2$, $t_1 \rightarrowtail t_3$, $\mathcal{I}_{t_3}(\gamma) = \{\gamma, \delta\}$, $\mathcal{I}_{t_2}(\gamma) = \{\beta, \gamma, \delta\}$ and $\mathcal{B}_{t_2}(\gamma) = \{\beta, \hat{I}\}$, so that $1_{t3}(\gamma) \subseteq 1_{t2}(\gamma)$ and $1_{t3}(\gamma) \cap \mathcal{B}_{t_2}(\gamma) = \{\gamma\} \neq 0$ but $\mathcal{B}_{t_3}(\gamma) = \{\gamma, \delta\} \neq 1_{t3}(w) \cap B_{t2}(w) = \{\gamma\}$.

Corollary 6.3. It follows from Proposition 6.1 that logic $\mathbb{L}_{\mathrm{AGlvI}}$ is characterized by the class of lLAGlvI-frames.

Some aspects of the relationship between logic $\mathbb{L}_{\mathrm{AGlvI}}$ and the AGM theory of belief revision were investigated in (Bonanno, 2üü7a). **In** the next section we explore this relationship in more detail, with the help of structures borrowed from the rational choice literature.

7 Relationship to the AGM theory

We begin by recalling the theory of belief revision due to Alchourrón, Gärdenfors and Makinson (1985), known as the AGM theory (see also Gärdenfors, 1988). **In** their approach beliefs are modeled as sets of formulas in a given syntactic language and belief revision is construed as an operation that associates with every deductively closed set of formulas K (thought of as the initial beliefs) and formula ϕ (thought of as new information) a new set of formulas $K^{\circledast}_{\phi}$ representing the revised beliefs.

7.1 AGM belief revision functions

Let S be a countable set of atomie propositions and $\mathcal{L}_0$ the propositional language built on S. Thus the set Φ_0 of formulas of $\mathcal{L}_0$ is defined recursively as follows: if p **ESthen** p E Φ_0 and if ϕ, ψ E Φ_0 then $\neg\phi$ E Φ_0 and $\phi \vee \psi$ E Φ_0.

Given a subset $K \subseteq \Phi_0$, its PL-deductive closure $[K]^{\mathrm{PL}}$ (where 'PL' stands for Propositional Logic) is defined as follows: ψ E $[K]^{\mathrm{PL}}$ if and only if there exist $\phi_1, \ldots \phi_n$ E K such that $(\phi_1 \wedge \ldots \wedge \phi_n) \rightarrow \psi$ is a tautology (that is, a theorem of Propositional Logic). A set $K \subseteq \Phi_0$ is *consistent* if $[K]^{\mathrm{PL}} \neq \Phi_0$ (equivalently, if there is no formula ϕ such that both ϕ and $\neg\phi$ belong to $[K]^{\mathrm{PL}}$). A set $K \subseteq \Phi_0$ is *deductively closed* if $K = [K]^{\mathrm{PL}}$. A *belief set* is a set $K \subseteq \Phi_0$ which is deductively closed. The set of belief sets will be denoted by $\mathbb{K}$ and the set of consistent belief sets by $\mathbb{K}^{\mathrm{con}}$.

Let $K \in \mathbb{K}^{\mathrm{con}}$ be a consistent belief set representing the agent's initial beliefs. A *belief reoisiori function fOT K* is a function

$$K^{\circledast} : \Phi_0 \rightarrow 2^{\Phi_0}$$

that associates with every formula ϕ E Φ_0 (thought of as new information) a set $K^{\circledast}(\phi) \subseteq \Phi_0$ (thought of as the new belief). **It** is common in the literature to use the notation $K^{\circledast}_{\phi}$ instead of $K^{\circledast}(\phi)$, but we prefer the latter.

A belief revision function is called an *AGM reoision function* if it satisfies the following properties, known as the AGM postulates: $\forall \phi, \psi \in \Phi_0$,

$$K^{\circledast}(\phi) \in \mathbb{K} \tag{$\circledast 1$}$$

$$\phi \in K^{\circledast}(\phi) \tag{$\circledast 2$}$$

$$K^{\circledast}(\phi) \subset [K \cup \{\phi\}]^{\mathrm{PL}} \tag{$\circledast 3$}$$

$$\text{if } \neg\phi \notin K, \text{ then } [K \cup \{\phi\}]^{\mathrm{PL}} \subset K^{\circledast}(\phi) \tag{$\circledast 4$}$$

$$\text{if } \phi \text{ is a contradiction then } K^{\circledast}(\phi) = \Phi_0 \tag{$\circledast 5a$}$$

$$\text{if } \phi \text{ is not a contradiction then } K^{\circledast}(\phi) \neq \Phi_0 \tag{$\circledast 5b$}$$

$$\text{if } \phi \leftrightarrow \psi \text{ is a tautology then } K^{\circledast}(\phi) = K^{\circledast}(\psi) \tag{$\circledast 6$}$$

$$K^{\circledast}(\phi \,\mathrm{r.}\, \psi) \subset \left[K^{\circledast}(\phi) \cup \{\psi\}\right]^{\mathrm{PL}} \tag{$\circledast 7$}$$

$$\text{if } \neg\psi \notin K^{\circledast}(\phi), \text{ then } \left[K^{\circledast}(\phi) \cup \{\psi\}\right]^{\mathrm{PL}} \subset K^{\circledast}(\phi \wedge \psi). \tag{$\circledast 8$}$$

$(\circledast 1)$ requires the revised belief set to be deductively closed.

$(\circledast 2)$ requires that the information be believed.

$(\circledast 3)$ 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.

$(\circledast 4)$ 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.

$(\circledast 5ab)$ require the revised beliefs to be consistent, unless the information ϕ is contradictory (that is, $\neg\phi$ is a tautology).

$(\circledast 6)$ 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.

$(\circledast 7)$ and $(\circledast 8)$ are a generalization of $(\circledast 3)$ and $(\circledast 4)$ that

> "applies to *itemted* changes of belief. The idea is that if $K^{\circledast}(\phi)$ is a revision of K [prompted by ϕ] and $K^{\circledast}(\phi)$ is to be changed by adding further sentences, such a change should be made by using expansions of $K^{\circledast}(\phi)$ whenever possible. More generally, the minimal change of K to include both ϕ and ψ (that is, $K^{\circledast}(\phi \wedge \psi)$) ought to be the same as the expansion of $K^{\circledast}(\phi)$ by ψ, so long as ψ does not contradiet the beliefs in $K^{\circledast}(\phi)$" (Gärdenfors, 1988, p. 55).9

We now turn to a semantic counterpart to the AGM belief revision functions, which is in the spirit of Grove's (1988) system of spheres. The structures we will consider are known in rational choice theory as *choice functions* (see, for example, Rott, 2001; Suzumura, 1983).

9 The expansion of $K^{\circledast}(\phi)$ by ψ is $\left[K^{\circledast}(\phi) \cup \{\psi\}\right]^{\mathrm{PL}}$.

7.2 Choice structures and one-stage revision frames

Definition 7.1. A *choice struciure* is a quadruple $\langle \Omega, \mathcal{E}, \mathbb{O}, \mathrm{R} \rangle$ where

- Ω is a non-empty set of *states;* subsets of Ω are called *eoenis.*
- $\mathcal{E} \subseteq 2^{\Omega}$ is a collection of events (2^{Ω} denotes the set of subsets of Ω).
- $\mathsf{R} : \mathcal{E} \to 2^{\Omega}$ is a function that associates with every event $E \in \mathcal{E}$ an event $RE \subseteq \Omega$ (we use the notation RE rather than $R(E)$).
- $\mathbb{O} \in \mathcal{E}$ is a distinguished element of $\mathcal{E}$ with $\mathbb{O} \neq \mathsf{O}$.

In rational choice theory a set $E \in \mathcal{E}$ is interpreted as a set of available alternatives and RE is interpreted as the subset of E which consists of those alternatives that could be rationally chosen. In our case, we interpret the elements of $\mathcal{E}$ as possible items of information that the agent might receive and the interpretation of RE is that, if informed that event E has occurred, the agent considers as possible all and only the states in RE. For the distinguished element $\mathbb{O}$, we interpret $\mathbf{R}_{\mathbb{O}}$ as the *oriqinol* or *initial* beliefs of the agent.I''

Note that we do not impose the requirement that $\Omega \in \mathcal{E}$.

Definition 7.2. A *one-stage remsioti frame* is a choice structure $\langle \Omega, \mathcal{E}, \mathbb{O}, \mathsf{R} \rangle$ that satisfies the following properties: $\forall E, F \in \mathcal{E}$,

$$RE \subseteq E, \tag{BR1}$$

$$\text{if } E \neq 0 \text{ then } RE \neq \mathsf{O}, \tag{BR2}$$

$$\text{if } E \subseteq F \text{ and } RF \cap E \neq 0 \text{ then } RE = RF \cap E, \text{[11]} \tag{BR3}$$

$$\text{if } \mathbf{R}_{\mathbb{O}} \cap E \neq 0 \text{ then } RE = \mathbf{R}_{\mathbb{O}} \cap E. \tag{BR4}$$

In the rational choice literature, (BR1) and (BR2) are taken to be part of the definition of a choice function, while (BR3) is known as Arrow's axiom (see Suzumura, 1983, p. 25). Property (BR4), which corresponds to our Qualitative Bayes Rule, has not been investigated in that literature.

The following is an example of a belief revision frame: $\Omega = \{\alpha, 13, \gamma, \delta\}$, $\mathcal{E} = \{\{\alpha, j3\}, \{\gamma, \delta\}, \{\alpha, 13, \hat{I}\}\}$, $\mathbb{O} = \{\alpha, 13, \hat{I}\}$, $\mathbf{R}_{\{\alpha,\beta\}} = \{j3\}$, $\mathbf{R}_{\{\gamma,\delta\}} = \{\gamma\}$, $\mathbf{R}_{\{\alpha,\beta,\gamma\}} = \{j3'\hat{I}\}'$

---

10 In the rational choice literature there is no counterpart to the distinguished set $\mathbb{O}$.

11 It is proved in the appendix that, in the presence of (BRI), (BR.3) is equivalent to: $\forall E, FEE$,

$$\text{if } RF \cap E \neq 0 \text{ and } E \in FEE \text{ then } REnF = RF \cap E. \tag{BR.3'}$$

A *one-stage reoisiori model* is a quintuple $\langle \Omega, \mathcal{E}, \mathbb{O}, \mathrm{R}, V \rangle$ where $\langle \Omega, \mathcal{E}, \mathbb{O}, \mathrm{R} \rangle$ is a one-stage revision frame and $V : S \to 2^{\Omega}$ is a function (called a *valuation)* that assoeiates with every atomie proposition P the set of states at whieh P is true. Truth of an arbitrary formula in a model is defined reeursively as follows ($w \models \phi$ means that formula ϕ is true at state w): (1) for $P \in S$, $w \models p$ if and only if $w \in V(p)$, (2) $w \models \neg\phi$ if and only if $w \not\models \phi$ and (3) $w \models \phi \vee \psi$ if and only if either $w \models \phi$ or $w \models \psi$ (or both). The truth set of a formula ϕ is denoted by $\|\phi\|$. Thus $\|\phi\| = \{w \in \Omega : w \models \phi\}$.

Given a one-stage revision model, we say that

(1) the agent *initially believes that* ϕ if and only if $\mathbf{R}_{\mathbb{O}} \subseteq \|\phi\|$,

(2) the agent *believes that* ϕ *upon learning that* ψ if and only if $\|\psi\| \in \mathcal{E}$ and $\mathbf{R}_{\|\psi\|} < \|\phi\|$.

Definition 7.3. A one-stage revision model is *comprehensioe* if for every formula ϕ, $\|\phi\| \in \mathcal{E}$. It is *ricli* if, for every finite set of atomie propositions $P = \{\mathrm{Pl}, \ldots, Pn, ql, \cdots, qm\}$, there is a state $w_P \in \Omega$ sueh that $ur_P \models Pi$ for every $\mathrm{i} = 1, \ldots, n$ and $w_P \models qj$ for every $j = 1, \ldots, \mathrm{m}$.

Thus in a eomprehensive one-stage revision model every formula is a possible item of information. For example, a model based on a one-stage revision frame where $\mathcal{E} = 2^{\Omega}$ is eomprehensive. In a rieh model every formula eonsisting of a conjunction of atomie proposition or the negation of atomie propositions is true at some state.

7.3 Correspondenee

We now show that the set of AGM belief revision functions eorresponds to the set of eomprehensive and rieh one-stage revision modeis, in the sense that

(1) given a eomprehensive and rieh one-stage revision model, we ean assoeiate with it a consistent belief set K and a corresponding AGM belief revision function $K^{\circledast}$, and

(2) given a consistent belief set K and an AGM belief revision function $K^{\circledast}$ there exists a eomprehensive and rieh one-stage revision model whose assoeiated belief set and AGM belief revision function eoineide with K and $K^{\circledast}$, respeetively.

Proposition 7.4. Let $\langle \Omega, \mathcal{E}, \mathbb{O}, \mathrm{R}, V \rangle$ be a eomprehensive one-stage revision model. Define $K = \{\psi \in \Phi_0 : \mathbf{R}_{\mathbb{O}} \subseteq \|\psi\|\}$. Then K is a consistent belief set. For every $\phi \in \Phi_0$ define $K^{\circledast}(\phi) = \{\psi \in \Phi_0 : \mathbf{R}_{\|\phi\|} \subseteq \|\psi\|\}$. Then the function $K^{\circledast} : \Phi_0 \to 2^{\Phi_0}$ so defined satisfies AGM postulates $(\circledast 1)$–$(\circledast 5\mathrm{a})$ and $(\circledast 6)$–$(\circledast 8)$. If the model is rieh then also $(\circledast 5\mathrm{b})$ is satisfied.

Proposition 7.5. Let $K \in \mathbb{K}$ be a consistent belief set and $K^{\circledast} : \Phi_0 \to 2^{\Phi_0}$ be an AGM belief revision function (that is, $K^{\circledast}$ satisfies the AGM postulates $(\circledast 1)$–$(\circledast 8)$). Then there exists a comprehensive and rich one-stage revision model $\langle \Omega, \mathcal{E}, \mathbb{O}, \mathbf{R}, V \rangle$ such that $K = \{\psi \in \Phi_0 : \mathbf{R}_{\mathbb{O}} \subseteq \|\psi\|\}$ and, for every $\phi \in \Phi_0$, $K^{\circledast}(\phi) = \{\psi \in \Phi_0 : \mathbf{R}_{\|\phi\|} \subseteq \|\psi\|\}$.

7.4 Back to $\mathbb{L}_{\mathrm{AGM}}$ frames

Given an $\mathbb{L}_{\mathrm{AGM}}$ frame $\langle T, \rightarrowtail, \Omega, \{B_t, \mathcal{I}_t\}_{t \in T} \rangle$ (see Definition 6.2) we can associate with every state-instant pair (w, t) a one-stage revision frame (see Definition 7.2) $\langle \Omega^0, \mathcal{E}^0, \mathbb{O}^0, \mathrm{Ra} \rangle$ as follows. Let $t_0 = \{t \in T : t_0 \rightarrowtail t\}$, then

- $\Omega^0 = \Omega$,
- $\mathcal{E}^0 = \{E \subseteq \Omega : E = I_t(w)$ for some $t \in \overrightarrow{t_0}\}$,
- $\mathbb{O}^0 = \mathcal{I}_{t_0}(w)$,
- $\mathbf{R}_{\mathbb{O}^0} = B_{t_0}(w)$,
- for every $E \in \mathcal{E}$, if $E = I_t(w)$ (for some $t \in \overrightarrow{t_0}$) then $\mathbf{R}^0_E = B_t(w)$.

By Property (2) of $\mathbb{L}_{\mathrm{AGM}}$-frames the frame $\langle \Omega^0, \mathcal{E}^0, \mathbb{O}^0, \mathrm{Ra} \rangle$ so defined satisfies property BR1 of the definition of one-stage revision frame, while Property (3) ensures that BR2 is satisfied, Property (4) ensures that BR3 is satisfied and Property (1) ensures that BR4 is satisfied.

Consider now the subset of the set of $\mathbb{L}_{\mathrm{AGM}}$ frames consisting of those frames satisfying the following properties:

$$\forall t \in T, \forall \omega \in \Omega, \forall E \in 2^{\Omega} \setminus \{\varnothing\}, \exists t' \in T : t \rightarrowtail t' \text{ and } \mathcal{I}_{t'}(\omega) = E. \qquad (\mathrm{P_{CMP}})$$

$$\forall t \in T, \forall \omega \in \Omega, \mathcal{I}_t(\omega) \neq 0. \qquad \text{(seriality of } \mathcal{I}_t)$$

Let $\mathbb{L}_{\mathrm{comp}}$ ("comp" stands for "comprehensive") be $\mathbb{L}_{\mathrm{AGM}} + \mathrm{CMP} + \mathrm{I}_{\mathrm{con}}$ where CMP and Icon are the following axioms: for every Boolean ϕ

$$\neg A \neg \phi \to \Diamond I \phi, \qquad \text{(CMP)}$$

$$\neg I(\phi \wedge \neg \phi). \qquad \text{(Icon)}$$

Axiom CMP says that, for every Boolean formula ϕ, if there is a state where ϕ is true, then there is a next instant where the agent is informed that ϕ, while axiom Icon rules out contradictory or inconsistent information.

Proposition 7.6. The logic $\mathbb{L}_{comp}$ is characterized by the class of $\mathbb{L}_{AGM}$-frames that satisfy PCMP and seriality of $\mathcal{I}_t$.[12]

We can view logic $\mathbb{L}_{comp}$ as an axiomatization of the AGM belief revision functions. **In** fact, if we take any model based on a $\mathbb{L}_{comp}$ frame and any state-instant pair, the one-stage revision frame associated with it is such that $\mathcal{E} = 2^{\Omega}\backslash\{\varnothing\}$. Thus the corresponding one-stage revision model is comprehensive (see Definition 7.3) and therefore, by Proposition 7.4, the associated AGM belief revision function $K^{\circledast} : \Phi_0 \to 2^{\Phi_0}$ satisfies AGM postulates $(\circledast 1)$–$(\circledast 5a)$ and $(\circledast 6)$–$(\circledast 8)$. Conversely, by Proposition 7.5, for every consistent belief set K and AGM belief revision function $K^{\circledast} : \Phi_0 \to 2^{\Phi_0}$ there is a model based on an $\mathbb{L}_{comp}$ frame whose associated AGM belief revision function coincides with $K^{\circledast}$.[13]

Models of $\mathbb{L}_{comp}$, however, are "very large" in that, for every state-instant pair and for every Boolean formula ϕ whose truth set is non-empty, there is a next instant where the agent is informed that ϕ. This requirement corresponds to assuming a complete belief revision policy for the agent, whereby the agent contemplates his potential reaction to every conceivable (and consistent) item of information. **In** a typical $\mathbb{L}_{AGM}$ frame, on the other hand, the items of information that the individual might receive at the next instant might be few, so that the agent's belief revision policy is limited to a few (perhaps the most likely) pieces of information. How does this limited belief revision policy associated with $\mathbb{L}_{AGM}$ frames relate to the AGM postulates for belief revision? The answer is given in the foliowing proposition, which was proved in (Bonanno, 2üü7a) (we have reworded it to fit the set-up of this section). We can no longer reeover an entire AGM belief revision function from a model based on an arbitrary $\mathbb{L}_{AGM}$ frame. However, we can recover, for every pair of Boolean formulas ϕ and ψ, the values $K^{\circledast}(\phi)$ and $K^{\circledast}(\phi \wedge \psi)$ of an AGM belief revision function whenever there is a next instant at which the agent is informed that ϕ and there is another next instant where he is informed that $(\phi \wedge \psi)$.

Proposition 7.7.
(A) Let $K \subseteq \Phi^B$ be a consistent and deductively closed set and let $K^{\circledast}$ $\Phi_0 \to 2^{\Phi_0}$ be an AGM belief revision function. Fix arbitrary $\phi, \psi \in \Phi^B$.

12 Note that, given the non-standard validation rule for $I\phi$ the equivalence ofaxiom D $(I\phi \to \neg I\neg\phi)$ and seriality of Z, breaks down. lt is still true that if Z, is serial then the axiom $I\phi \to \neg I\neg\phi$ is valid, but the converse is not true (see Bonanno, 2005, p. 226, Footnote 25).

13 All we need to do in this respect is to eliminate the empty set from $\mathcal{E}$ in the proof of Proposition 7.5, that is, discard the possibility that ϕ is a contradiction.

Then there is an lLAGlvI-model, t_i, $t2, t3$ ET and α E Ω such that:

$$t\text{-:} \rightarrowtail t\text{-:} \tag{A.1}$$

$$K = \{X \in \Phi^B : (\alpha, t_1) \models BX\} \tag{A.2}$$

$$(\alpha, t_2) \models I\phi \tag{A.3}$$

$$K^{\circledast}(\phi) = \{X \in \Phi^B : (\alpha, t2) \models BX\} \tag{AA}$$

$$\text{if } \phi \text{ is consistent then } (\beta, t) \models \phi \text{ for some } \beta \in \Omega \text{ and } t \text{ ET} \tag{A.5}$$

$$t_1 \rightarrowtail t_3 \tag{A.6}$$

$$(\alpha, t_3) \models I(\phi \wedge \psi) \tag{A.7}$$

$$K^{\circledast}(\phi \wedge \psi) = \{X \in \Phi^B : (\alpha, t_3) \models BX\} \tag{A.8}$$

$$\text{if } (\phi \,!\backslash\, \psi) \text{ is consistent then } (\hat{I}' ti) \models (\phi \wedge \psi) \text{ for some } \gamma \in \Omega \text{ and } ti \in T. \tag{A.9}$$

(B) Fix an lLAGlvI-model such that (1) for some $t_1, t\text{-}i, t_3 \in T$, $\alpha \in \Omega$ and $\phi, \psi \in \Phi^B$, $t_1 \rightarrowtail t\text{-}z$, $t_1 \rightarrowtail t3$, $(\alpha, t_2) \models I\phi$ and $(\alpha, t_3) \models I(\phi \wedge \psi)$, (2) if ϕ is not a contradiction then $(\beta, t) \models \phi$, for some $\beta \in \Omega$ and $t \in T$ and (3) if $(\phi \wedge \psi)$ is not a contradiction then $(\gamma, ti) \models (\phi \wedge \psi)$, for some γ E Ω and ti E T. Define $K = \{X \in \Phi^B : (\alpha, t_1) \models BX\}$. Then there exists an AGM belief revision function $K^{\circledast} : \Phi_0 \rightarrow 2^{\Phi_0}$ such that $K^{\circledast}(\phi) = \{X \in \Phi^B : (\alpha, t2) \models BX\}$ and $K^{\circledast}(\phi \wedge \psi) = \{X \in \Phi^B : (\alpha, t3) \models BX\}$. Furthermore, for every $\phi, \psi \in \Phi^B$, there exists an lLAGlvI-model such that, for some α E Ω and $t_2, t3$ ET, (1) $(\alpha, t_2) \models I\phi$ and $(\alpha, t_3) \models I(\phi \wedge \psi)$, (2) if ϕ is not a contradiction then $(\beta, t) \models \phi$, for some β E Ω and t ET and (3) if $(\phi \wedge \psi)$ is not a contradiction then $(\gamma, ti) \models (\phi \wedge \psi)$, for some $\gamma \in \Omega$ and ti ET.

8 Conclusion

We proposed a temporal logic where information and beliefs are modeled by means of two modal operators land *B,* respectively. A third modal operator, the next-time operator 0, enables one to express the dynamic interaction of information and beliefs over time. The proposed logic can be viewed as a temporal generalization of the theory of static belief pioneered by Hintikka (1962).

The combined syntactic-semantic approach of modal logic allows one to state properties of beliefs in a clear and transparent way by means ofaxioms and to show the correspondence between axioms and semantic properties. Natural extensions of our lLAGM logic would impose, besides consistency of information (axiom Icon)[14], the standard KD45 axioms for belief (axiom 4:

14 As pointed out by Friedman and Halperu (1999), it is not clear how one could be informed of a contradiction.

$B\phi \rightarrow BB\phi$ and axiom 5: $\neg B\phi \rightarrow B\neg B\phi$, while the D axiom: $B\phi \rightarrow \neg B\neg\phi$ would follow from axioms Icon and WC). Furthermore, one might want to investigate axioms that capture the notion of memory or recall, for instanee $B\phi \rightarrow \bigcirc B\bigcirc^{-1}B\phi$ and $\neg B\phi \rightarrow \bigcirc B\bigcirc^{-1}\neg B\phi$ (the agent always remembers what he believed and what he did not believe in the immediately preceding past). Further strengthenings might add the requirement that information be correct $(I\phi \rightarrow \phi)$ or the weaker requirement that the agent trusts the information souree $(B(I\phi \rightarrow \phi))$. Another natural direction to explore is the axiomatization of *iteraied reoisiori,* a topic that has received considerable attention in recent years (see, for example, Boutilier, 1996; Darwiche and Pearl, 1997; Nayak et al., 2003; Rott, 2006). Extensions of the logic $\mathsf{IL}_{\mathsf{AGM}}$ that incorporate axioms for iterated revision have been recently investigated in (Zvesper, 2007). Finally, another line of research, which is pursued in (Bonanno, 2007b), deals with the conditions under which belief revision can be rationalized by a plausibility ordering on the set of states, in the sense that the set of states that are considered possible aftel' being informed that ϕ coincides with the most plausible states that are compatible with ϕ.

Acknowledgments

Parts of this paper were presented at the Seventh Conference on *Logic and the Foundations of Game and Decision Theoru* (LOFT7; Liverpool, July 2006) and at the KNAW Academy Colloquium on *New perspectcoes on Games and Interaction* (Amsterdam, February 2007). lam grateful to Jan van Eijck and Johan van Benthem for helpful comments.

Appendix

Proof of the claim in Footnote 6, namely that axiom WND is equivalent to the following axiom: if ϕ and ψ are Boolean,

$$\Diamond^{-1}(B\phi \wedge B\psi \wedge \neg B\neg\phi) \wedge I\phi \rightarrow B\psi.$$

Derivation of WND from the above axiom ('PL' stands for 'Propositional Logic'):

1.	$\bigcirc^{-1}(B\phi \wedge B\psi \wedge \neg B\neg\phi) \rightarrow (I\phi \rightarrow B\psi)$	above axiom, **PL**
2.	$\bigcirc\Diamond^{-1}(B\phi \wedge B\psi \wedge \neg B\neg\phi) \rightarrow \bigcirc(I\phi \rightarrow B\psi)$	1, rule **RK** for $\bigcirc$
3.	$(B\phi \wedge B\psi \wedge \neg B\neg\phi) \rightarrow \bigcirc\Diamond^{-1}(B\phi \wedge B\psi \wedge \neg B\neg\phi)$	Temporal axiom 01
4.	$(B\phi \wedge B\psi \wedge \neg B\neg\phi) \rightarrow \bigcirc(I\phi \rightarrow B\psi)$	2,3, PL.

Derivation of the above axiom from WND:

1. $(B\phi \mathsf{r}.\ B\psi \wedge \neg B\neg\phi) \rightarrow$ $\mathbf{O}(\mathrm{I}\phi \rightarrow B\psi)$ Axiom WND
2. $\neg 0\ (I\phi \rightarrow B\psi)$ $\rightarrow \neg(B\phi \wedge B\psi \wedge \neg B\neg\phi)$ 1, **PL**
3. $O\text{-}^{1}\text{-}'0\ (I\phi \rightarrow B\psi)$ $\rightarrow \bigcirc^{-1}\neg(B\phi \mathsf{r}.\ B\psi \mathsf{r}.\ \neg B\neg\phi)$ 2, rule **RK** for 0-^{1}
4. $\Diamond^{-1}(B\phi \mathsf{r}.\ B\psi \mathsf{r}.\ \neg B\neg\phi)$ $\rightarrow \Diamond^{-1} \bigcirc (I\phi \rightarrow B\psi)$ 3, PL, definition of 0-^{1}
5. $\neg(I\phi \rightarrow B\psi) \rightarrow$ $\bigcirc^{-1}\Diamond\neg(I\phi \rightarrow B\psi)$ Temporal axiom O_2
6. $O\text{-}^{1}O\ (I\phi \rightarrow B\psi) \rightarrow (I\phi \rightarrow B\psi)$ 5, PL, definition of 0-^{1} and $\Diamond$
7. $\Diamond^{-1}(B\phi \mathsf{r}.\ B\psi \mathsf{r}.\ \neg B\neg\phi) \rightarrow$ $(I\phi \rightarrow B\psi)$ 4,6, PL.

Q.E.D.

PTOof of the claim in Fooinote 7, namely that axiom WN A is equivalent to the foliowing axiom: if ϕ and ψ are Boolean,

$$\Diamond^{-1}(B\phi \wedge \neg B\psi) \wedge I\phi \rightarrow \neg B\psi.$$

Derivation of WNA from the above axiom:

1. $\Diamond^{-1}(B\phi \mathsf{r}.\ \neg B\psi)\ \mathsf{r}.\ I\phi \rightarrow \neg B\psi$ above axiom
2. $\Diamond^{-1}(B\phi \mathsf{r}.\ \neg B\psi) \rightarrow (I\phi \rightarrow \neg B\psi)$ 1, **PL**
3. $\bigcirc\Diamond^{-1}(B\phi \mathsf{r}.\ \neg B\psi) \rightarrow \bigcirc(I\phi \rightarrow \neg B\psi)$ 2, rule **RK** for 0
4. $(B\phi \mathsf{r}.\ \neg B\psi) \rightarrow \bigcirc\Diamond^{-1}(B\phi \mathsf{r}.\ \neg B\psi)$ Temporal axiom 01
5. $(B\phi \wedge \neg B\psi) \rightarrow \bigcirc(I\phi \rightarrow \neg B\psi)$ 3,4, PL.

Derivation of the above axiom from WNA:

1. $(B\phi \wedge \neg B\psi) \rightarrow \bigcirc(I\phi \rightarrow \neg B\psi)$ Axiom WNA
2. $\neg 0\ (I\phi \rightarrow \neg B\psi) \rightarrow$ $\neg(B\phi \wedge \neg B\psi)$ 1, **PL**
3. $\mathbf{O}\text{-}^{1}\text{-},\ 0\ (I\phi \rightarrow \neg B\psi)$ $\rightarrow \bigcirc^{-1}\neg(B\phi \wedge \neg B\psi)$ 2, rule **RK** for $O\text{-}^{1}$
4. $\Diamond^{-1}(B\phi \wedge \neg B\psi) \rightarrow$ $O\text{-}^{1}O\ (I\phi \rightarrow \neg B\psi)$ 3, **PL** and definition of 0-^{1}
5. $\text{-},(1\phi \rightarrow \neg B\psi) \rightarrow$ $\bigcirc^{-1}\Diamond\neg(I\phi \rightarrow \neg B\psi)$ Temporal axiom O_2
6. $O\text{-}^{1}O\ (I\phi \rightarrow \neg B\psi) \rightarrow$ $(I\phi \rightarrow \neg B\psi)$ 5, PL, definition of 0-^{1} and $\Diamond$
7. $\Diamond^{-1}(B\phi \wedge \neg B\psi) \rightarrow (I\phi \rightarrow \neg B\psi)$ 4,6, **PL**
8. $\Diamond^{-1}(B\phi \mathsf{r}.\ \neg B\psi)\ \mathsf{r}.\ I\phi \rightarrow \neg B\psi$ 7, PL.

Q.E.D.

Proof of Proposiiion 4.1. (1) Fix a frame that satisfies P_{WNO}, an arbitrary model based on it and arbitrary $\alpha \in \Omega$, $t_1 \in T$ and Boolean formulas ϕ and ψ and suppose that $(\alpha, t_1) \models (B\phi \wedge B\psi \wedge \neg B\neg\phi)$. Sinee $(\alpha, t_1) \models \neg B\neg\phi$, there exists an $w \in B_{t_1}(\alpha)$ sueh that $(w, t_1) \models \phi$. Thus $B_{t_1}(\alpha) \neq \emptyset$. Fix an arbitrary $t_2 \in T$ sueh that $t_1 \rightarrow t_2$ and suppose that $(\alpha, t_2) \models I\phi$. Then $\mathcal{I}_{t_2}(\alpha) = [\phi]_{t_2}$. Fix an arbitrary $\beta \in \mathcal{B}_{t_1}(\alpha)$. Sinee $(\alpha, t_1) \models B\phi$, $(\beta, t_1) \models \phi$. Sinee ϕ is Boolean, by Remark 2.2 $(\beta, t_2) \models \phi$. Henee $\beta \in \mathcal{I}_{t_2}(\alpha)$. Thus $\mathcal{B}_{t_1}(\alpha) \subset \mathcal{I}_{t_2}(\alpha)$. Hence, by P_{WNO}, $\mathcal{B}_{t_2}(\alpha) \subset \mathcal{B}_{t_1}(\alpha)$. Fix an arbitrary $w \in \mathcal{B}_{t_2}(\alpha)$. Then $w \in B_{t_1}(\alpha)$ and, sinee $(\alpha, t_1) \models B\psi$, $(w, t_1) \models \psi$. Since $'$ is Boolean, by Remark 2.2 $(w, t_2) \models \psi$. Thus $(\alpha, t_2) \models B\psi$.

Conversely, suppose that P_{WNO} is violated. Then there exist $\alpha \in \Omega$ and $t_1, t_2 \in T$ sueh that $t_1 \rightarrowtail t_2$, $B_{t_1}(\alpha) \neq \emptyset$, $B_{t_1}(\alpha) \subseteq \mathcal{I}_{t_2}(\alpha)$ and $B_{t_2}(\alpha) \not\subseteq B_{t_1}(\alpha)$. Let p and q be atomie propositions and construct a model where $\|p\| = \mathcal{I}_{t_2}(\alpha)$ x T and $\|q\| = \mathcal{E}_{t_1}(\alpha)$ x T. Then $(\alpha, t_1) \models (Bp \wedge Bq \wedge \neg B\neg p)$. By hypothesis, there exists a $\beta \in \mathcal{B}_{t_2}(\alpha)$ sueh that $\beta \notin B_{t_1}(\alpha)$, so that $(\beta, t_2) \not\models q$. Henee $(\alpha, t_2) \not\models Bq$ while $(\alpha, t_2) \models \neg p$, so that $(\alpha, t_2) \not\models \neg p \rightarrow Bq$. Thus, sinee $t_1 \rightarrowtail t_2$, WND is falsified at (α, t_1).

(2) Fix a frame that satisfies P_{WNA}, an arbitrary model based on it and arbitrary $\alpha \in \Omega$, $t_1 \in T$ and Boolean formulas ϕ and ψ and suppose that $(\alpha, t_1) \models B\phi \wedge \neg B\psi$. Then there exists a $\beta \in B_{t_1}(\alpha)$ sueh that $(\beta, t_1) \models \neg\psi$. Fix an arbitrary $t_2 \in T$ sueh that $t_1 \rightarrowtail t_2$ and suppose that $(\alpha, t_2) \models I\phi$. Then $\mathcal{I}_{t_2}(\alpha) = [\phi]_{t_2}$. Fix an arbitrary $w \in \mathcal{B}_{t_1}(\alpha)$. Sinee $(\alpha, t_1) \models B\phi$, $(w, t_1) \models \phi$. Sinee ϕ is Boolean, by Remark 2.2 $(w, t_2) \models \phi$ and therefore $w \in \mathcal{I}_{t_2}(\alpha)$. Thus $\mathcal{B}_{t_1}(\alpha) \subset \mathcal{I}_{t_2}(\alpha)$ and, by P_{WNA}, $\mathcal{B}_{t_1}(\alpha) \subset \mathcal{B}_{t_2}(\alpha)$. Sinee $(\beta, t_1) \models \neg\psi$ and $\neg\psi$ is Boolean (because $'$ is), by Remark 2.2 $(\beta, t_2) \models \neg\psi$. Sinee $\beta \in B_{t_1}(\alpha)$ and $B_{t_1}(\alpha) \subseteq B_{t_2}(\alpha)$, $\beta \in B_{t_2}(\alpha)$ and therefore $(\alpha, t_2) \models \neg B\psi$.

Conversely, suppose that P_{WNA} is violated. Then there exist $\alpha \in \Omega$ and $t_1, t_2 \in T$ sueh that $t_1 \rightarrowtail t_2$ and $\mathcal{B}_{t_1}(\alpha) \subseteq \mathcal{I}_{t_2}(\alpha)$ and $\mathcal{B}_{t_1}(\alpha) \not\subseteq \mathcal{B}_{t_2}(\alpha)$. Let p and q be atomie propositions and construct a model where $\|p\| = \mathcal{I}_{t_2}(\alpha)$ x T and $\|q\| = \mathcal{B}_{t_2}(\alpha)$ x T. Then $(\alpha, t_1) \models Bp \wedge \neg Bq$ and $(\alpha, t_2) \models Lpr.Bq$, so that, sinee $t_1 \rightarrowtail t_2$, $(\alpha, t_1) \models \neg\bigcirc(lp \rightarrow \neg Bq)$. Q.E.D.

Proof of Proposiiion 4.3. First of all, note that, sinee $\bigcirc^{-1}$ is a normal operator, the following is a theorem of $\mathbb{L}_a$ (henee of $\mathbb{L}_w$):

$$\Diamond^{-1}\chi \wedge \bigcirc^{-1}\xi \rightarrow \Diamond^{-1}(\chi \wedge \xi). \tag{1}$$

It follows from (1) and axiom $B\cup$ that the following is a theorem of $\mathbb{L}_a$:

$$\Diamond^{-1}\chi \wedge \Diamond^{-1}\xi \rightarrow \Diamond^{-1}(\chi \wedge \xi). \tag{2}$$

Figure 3 below is a syntaetie derivation of WNC. Q.E.D.

1.	$\Diamond^{-1}(B\phi \wedge \neg B\neg\phi) \wedge \mathbf{0}\text{-}^{1}B\psi \rightarrow \Diamond^{-1}(B\phi \wedge \neg B\neg\phi \wedge B\psi)$	Theorem of *La* (see (2) above)
2.	$\mathbf{0}\text{-}^{1}(B\phi \wedge \neg B\neg\phi \wedge B\psi) \wedge I\phi \rightarrow B\psi$	Equivalent to WND (see Footnote 6)
3.	$\Diamond^{-1}(B\phi \wedge \neg B\neg\phi) \wedge \mathbf{0}\text{-}^{1}B\psi \wedge I\phi \rightarrow B\psi$	1,2, PL
4.	$I\phi \wedge \Diamond^{-1}(B\phi \wedge \neg B\neg\phi) \rightarrow (\mathbf{0}\text{-}^{1}B\psi \rightarrow B\psi)$	3,PL
5.	$\Diamond^{-1}(B\phi \wedge \neg B\neg\phi) \wedge \bigcirc^{-1}\neg B\psi \rightarrow \Diamond^{-1}(B\phi \wedge \neg B\neg\phi \wedge \neg B\psi)$	Theorem of *La* (see (1) above)
6.	$\neg(B\phi \wedge \neg B\psi) \rightarrow \neg(B\phi \wedge \neg B\neg\phi \wedge \neg B\psi)$	Tautology
7.	$\bigcirc^{-1}\neg(B\phi \wedge \neg B\psi) \rightarrow \bigcirc^{-1}\neg(B\phi \wedge \neg B\neg\phi \wedge \neg B\psi)$	6, mie RK for $\mathbf{0}\text{-}^{1}$
8.	$\Diamond^{-1}(B\phi \wedge \neg B\neg\phi \wedge \neg B\psi) \rightarrow \Diamond^{-1}(B\phi \wedge \neg B\psi)$	7, PL, def. of $\mathbf{0}\text{-}^{1}$
9.	$\Diamond^{-1}(B\phi \wedge \neg B\neg\phi) \wedge \bigcirc^{-1}\neg B\psi \rightarrow \Diamond^{-1}(B\phi \wedge \neg B\psi)$	5,8, PL
10.	$\mathbf{0}\text{-}^{1}(B\phi \wedge \neg B\psi) \wedge I\phi \rightarrow \neg B\psi$	equivalent to WNA (see Footnote 7)
11.	$\mathbf{0}\text{-}^{1}(B\phi \wedge \neg B\neg\phi) \wedge \bigcirc^{-1}\neg B\psi \wedge I\phi \rightarrow \neg B\psi$	9,10, PL
12.	$I\phi \wedge \Diamond^{-1}(B\phi \wedge \neg B\neg\phi) \rightarrow (\bigcirc^{-1}\neg B\psi \rightarrow \neg B\psi)$	11, PL
13.	$(\bigcirc^{-1}\neg B\psi \rightarrow \neg B\psi) \rightarrow (B\psi \rightarrow \mathbf{0}\text{-}^{1}B\psi)$	tautology and definition of $\mathbf{0}\text{-}^{1}$
14.	$I\phi \wedge \Diamond^{-1}(B\phi \wedge \neg B\neg\phi) \rightarrow (B\psi \rightarrow \mathbf{0}\text{-}^{1}B\psi)$	12, 13, PL
15.	$I\phi \wedge \Diamond^{-1}(B\phi \wedge \neg B\neg\phi) \rightarrow (B\psi \leftrightarrow \mathbf{0}\text{-}^{1}B\psi)$	4,14, PL.

FIGURE 3.

PTOof of Pmposition 5.1. (1) Fix a frame that satisfies PND, an arbitrary model based on it and arbitrary $\alpha \in \Omega$, t-: $\in$ Tand Boolean formulas ϕ and ψ and suppose that $(\alpha, t_1) \models \neg B\neg\phi \wedge B\psi$. Fix an arbitrary t-: ET sueh that t-: $\rightarrowtail$ t-: and $(\alpha, t2) \models I\phi$. Then $\mathcal{I}_{t_2}(\alpha) = \lceil\phi\rceil_{t\text{-}z:}$ Sinee $(\alpha, t_1) \models \neg B\neg\phi$, there exists a $13 \in B_{tl}(\alpha)$ sueh that $(13, t_1) \models \phi$. Sinee ϕ is Boolean, by Remark 2.2 (j3,t2) $\models \phi$ and, therefore, $13 \in \mathcal{I}_{t_2}(\alpha)$. Thus $\mathcal{B}_{t_1}(\alpha) \cap \mathcal{I}_{t_2}(\alpha) \neq 0$ and, by PND, $\mathcal{B}_{t_2}(\alpha) \subseteq B_{tl}(\alpha)$. Fix an arbitrary $w \in \mathcal{B}_{t_2}(\alpha)$. Then $w \in B_{tl}(\alpha)$ and, sinee $(\alpha, t_1) \models B\psi$, (w, t 1) $\models \psi$. Since ψ is Boolean, by Remark 2.2, (w, t_2) $\models \psi$. Henee $(\alpha, t_2) \models B\psi$.

Conversely, fix a frame that does not satisfy PND. Then there exist $\alpha \in \Omega$ and $t_1, t2 \in T$ such that t-: $\rightarrowtail$ $t2$, $B_{tl}(\alpha) \cap \mathcal{I}_{t_2}(\alpha) \neq 0$ and $\mathcal{B}_{t_2}(\alpha) \nsubseteq B_{tl}(\alpha)$. Let pand q be atomie propositions and construct a model where Ilpll = $B_{tl}(\alpha)$ x Tand Ilqll = $\mathcal{I}_{t_2}(\alpha)$ x T. Then $(\alpha, t_1) \models \mathcal{B}, q \wedge Bp$ and $(\alpha, t2) \models Iq$. By hypothesis there exists a $13 \in B_{t2}(\alpha)$ sueh that $13 \notin B_{tl}(\alpha)$.

Thus $(\beta, t2) \not\models$ pand therefore $(\alpha, t2) \models$ -,Bp. Henee $(\alpha, t_1) \models \neg$ 0 $(Iq \rightarrow Bp)$.

(2) Fix a frame that satisfies P_{NA}, an arbitrary model based on it and arbitrary α E Ω, t-: E Tand Boolean formulas ϕ and ψ and suppose that $(\alpha, t_1) \models \neg B \neg(\phi \wedge \neg\psi)$. Fix an arbitrary t-: E T sueh that t-: $\rightarrowtail$ t-: and suppose that $(\alpha, t_2) \models I\phi$. Then $1_{t2}(a) = [\phi]_{t_2}$. Sinee $(\alpha, t_1) \models \neg B \neg(\phi \wedge \neg\psi)$, there exists a β E $B_{tl}(\alpha)$ sueh that $(\beta, t_1) \models \phi \wedge \neg\psi$. Sinee ϕ and ψ are Boolean, by Remark 2.2 (;3,t2) $\models \phi \wedge \neg\psi$. Thus;3 E $1_{t2}(a)$ and, by P_{NA},;3 E $\mathcal{B}_{t_2}(\alpha)$. Thus, sinee (;3,t2) $\models \neg\psi$, $(\alpha, t_2) \models \neg B\psi$.

Conversely, fix a frame that does not satisfy P_{NA}. Then there exist α E Ω and $t_{1,t2}$ ET sueh that t-: $\rightarrowtail$ t-: and $\mathcal{B}_{t_1}(\alpha) \cap 1_{t2}(a) \not\subseteq \mathcal{B}_{t_2}(\alpha)$. Let pand q be atomie propositions and construct a model where $\|p\| = 1_{t2}(a)$ x Tand $\|q\| = \mathcal{B}_{t_2}(\alpha)$ x T. Then $(\alpha, t2) \models lp \wedge Bq$ and, therefore, $(\alpha, t_1) \models \neg$ 0 $(lp \rightarrow$ -,Bq). Sinee $B_{tl}(\alpha) \cap 1_{t2}(a) \not\subseteq \mathcal{B}_{t_2}(\alpha)$ there exists a β E $e_{,}(\alpha) \cap 1_{t2}(a)$ sueh that $\beta \notin \mathcal{B}_{t_2}(\alpha)$. Thus $(\beta, t_1) \models p \wedge$ -'q. Henee $(\alpha, t_1) \models$ -,B-,(p/\ -,q), so that axiom NA is falsified at (α, t_1).

(3) Fix a frame that satisfies P_{QA}, an arbitrary model based on it and arbitrary α E Ω, t-: E Tand Boolean formula ϕ and suppose that $(\alpha, t_1) \models \neg B \neg\phi$. Then there exists a β E $B_{tl}(\alpha)$ sueh that $(\beta, t_1) \models \phi$. Fix an arbitrary $t2$ sueh that $t_1 \rightarrowtail t2$ and suppose that $(\alpha, t2) \models I\phi$. Then $1_{t2}(\alpha) = [\phi]_{t\text{-}z'}$ Sinee ϕ is Boolean and $(\beta, t_1) \models \phi$, by Remark 2.2 $(\beta, t2) \models \phi$. Thus β E $1_{t2}(a)$ and, therefore, $\mathcal{B}_{t_1}(\alpha) \cap 1_{t2}(a) \neq$ o. By P_{QA}, $\mathcal{B}_{t_2}(\alpha) < 1_{t2}(a)$. Thus $(\alpha, t_2) \models B\phi$. Henee $(\alpha, t_1) \models \bigcirc(I\phi \rightarrow B\phi)$.

Conversely, suppose that P_{QA} is violated. Then there exist α E Ω and t_1, t-: ET sueh that t-: $\rightarrowtail$ t2, $B_{tl}(\alpha) \cap 1_{t2}(a) \neq 0$ and $\mathcal{B}_{t_2}(\alpha) \not\subseteq 1_{t2}(\alpha)$. Let p be an atomie proposition and construct a model where $\|p\| = 1_{t2}(a)$ x T. Then $(\alpha, t_1) \models$ -,B-,p and $(\alpha, t2) \models lp$. By hypothesis, there exists a β E $B_{t2}(\alpha)$ sueh that $\beta \notin 1_{t2}(\alpha)$. Thus $(\beta, t2) \not\models$ pand therefore $(\alpha, t2) \models$ -,Bp. Henee $(\alpha, t_1) \not\models O(Ip \rightarrow Bp)$. Q.E.D.

Proof of the claim in Remark 5.4, namely that WNA is a theorem of the logie $\mathbb{L}_a$ + NA:

1.	$\neg B(\phi \rightarrow \psi) \rightarrow \bigcirc(I\phi \rightarrow \neg B\psi)$	Axiom NA (see Footnote 8)
2.	$B(\phi \rightarrow \psi) \rightarrow (B\phi \rightarrow B\psi)$	Axiom K for B
3.	$(B\phi \wedge \neg B\psi) \rightarrow \neg B(\phi \rightarrow \psi)$	2, **PL**
4.	$(B\phi \wedge \neg B\psi) \rightarrow O(1\phi \rightarrow \neg B\psi)$	1,3, PL.

Q.E.D.

Proof of Proposiiion 6.1. (1) The proof of this part is straightforward and is omitted.

(2) Fix a frame that satisfies property P_{K7}. Let α and t-: be sueh that $(a, t_1) \models \Diamond(I(\phi \wedge \psi) \wedge BX)$, where ϕ, ψ and X are Boolean formulas. Then

there exists a *t3* sueh that *t-:* $\rightarrowtail$ *t3* and $(\alpha, t3) \models I(\phi \wedge \psi) \wedge BX$. Thus $\mathcal{I}_{t_3}(\alpha) = \lceil \phi \wedge \psi \rceil_{t3'}$ Fix an arbitrary *t2* sueh that $t1 \rightarrowtail t2$ and suppose that $(\alpha, t_2) \models I\phi$. Then $\mathcal{I}_{t_2}(\alpha) = \lceil \phi \rceil_{t\text{-}z'}$ Sinee ϕ and ψ are Boolean, by Remark 2.2, $\lceil \phi \wedge \psi \rceil_{t_3} = \lceil \phi \wedge \psi \rceil_{t_2}$. Thus, sinee $\lceil \phi \wedge \psi \rceil_{t_2} < \lceil \phi \rceil_{t_2}$, $\mathcal{I}_{t_3}(\alpha) < \mathcal{I}_{t_2}(\alpha)$. Henee by P_{K7}, $\mathcal{I}_{t_3}(\alpha) \cap \mathcal{B}_{t_2}(\alpha) < \mathcal{B}_{t_3}(\alpha)$. Fix an arbitrary *13* E $\mathcal{B}_{t_2}(\alpha)$. If *(j3,t2)* $\models \neg(\phi \wedge \psi)$ then *(j3,t2)* $\models (\phi \wedge \psi) \rightarrow X$· If *(j3,t2)* $\models \phi \wedge \psi$, then, by Remark 2.2, *(13, t3)* $\models \phi \wedge \psi$ and, therefore, *13* E $\mathcal{I}_{t_3}(\alpha)$. Henee *13* E $B_{t3}(\alpha)$. Sinee $(\alpha, t3) \models BX$, *(13, t3)* $\models X$ and, therefore, *(13, t3)* $\models (\phi \wedge \psi) \rightarrow X$· Sinee $(\phi \wedge \psi) \rightarrow X$ is Boolean (beeause ϕ, ψ and X are), by Remark 2.2, *(j3,t2)* $\models (\phi \wedge \psi) \rightarrow X$· Thus, sinee *13* E $\mathcal{B}_{t_2}(\alpha)$ was ehosen arbitrarily, $(\alpha, t_2) \models B((\phi \wedge \psi) \rightarrow X)$·

Conversely, suppose that P_{K7} is violated. Then there exist *t-i, t2, t3* and α sueh that *t-:* $\rightarrowtail$ *t2, t-:* $\rightarrowtail$ *t3*, $\mathcal{I}_{t_3}(\alpha) \subseteq \mathcal{I}_{t_2}(\alpha)$ and $\mathcal{I}_{t_3}(\alpha) \cap \mathcal{B}_{t_2}(\alpha) \nsubseteq \mathcal{B}_{t_3}(\alpha)$. Let *p, q* and *r* be atomie propositions and construct a model where $\|p\| = \mathcal{I}_{t_2}(\alpha) xT$, $\|q\| = \mathcal{I}_{t_3}(\alpha) xT$ and $\|T\| = \mathcal{B}_{t_3}(\alpha) xT$. Then, $(\alpha, t3) \models Br$ and, sinee $\mathcal{I}_{t_3}(\alpha) < \mathcal{I}_{t_2}(\alpha)$, $\mathcal{I}_{t_3}(\alpha) = ip/\backslash qlt3$ so that $(\alpha, t_3) \models I(p/\backslash q)$. Thus, sinee *t-:* $\rightarrow$ *t3*, $(\alpha, t_1) \models O(I(p \wedge q) \wedge e\text{-};$ Byeonstruetion, $(\alpha, t2) \models lp$. Sinee $\mathcal{I}_{t_3}(\alpha) \cap \mathcal{B}_{t_2}(\alpha) \nsubseteq \mathcal{B}_{t_3}(\alpha)$, there exists a *13* E $\mathcal{I}_{t_3}(\alpha) \cap \mathcal{B}_{t_2}(\alpha)$ sueh that *13* $\notin B_{t3}(\alpha)$. Thus *(13, t 2)* $\models$ -*'T;* furthermore, sinee *13* E $\mathcal{I}_{t_3}(\alpha)$, *(13, t 3)* $\models p/\backslash q$ and, by Remark 2.2, *(13, t 2)* $\models p \wedge q$. Thus, *(13, t 2)* $\not\models (p \wedge q) \rightarrow r$. Sinee *13* E $B_{t2}(\alpha)$ it follows that $(\alpha, t2) \not\models B((p \wedge q) \rightarrow r)$. Hence, sinee *t-:* $\rightarrowtail$ *t2*, $(\alpha, t_1) \not\models Q(lp \rightarrow B((p \wedge q) \rightarrow T))$ so that axiom K7 is falsified at (α, t_1).

(3) Fix a frame that satisfies property P_{KS}. Let ϕ, ψ and X be Boolean formulas and let α and t_1 be sueh that $(\alpha, t_1) \models \Diamond(I\phi \wedge \neg B\neg(\phi \wedge \psi) \wedge B(\psi \rightarrow X))$. Then there exists a *t-:* sueh that *t-:* $\rightarrowtail$ *t-:* and $(\alpha, t2) \models I\phi \wedge$ -,*B-,(*$\phi \wedge \psi) \wedge B(\psi \rightarrow X)$. Thus $\mathcal{I}_{t_2}(\alpha) = \lceil \phi \rceil_{t_2}$ and there exists a *13* E $\mathcal{B}_{t_2}(\alpha)$ sueh that *(13, t2)* $\models \phi \wedge \psi$. Fix an arbitrary *t3* sueh that *t-:* $\rightarrowtail$ *t3* and suppose that $(\alpha, t_3) \models I(\phi \wedge \psi)$. Then $\mathcal{I}_{t_3}(\alpha) = \lceil \phi \wedge \psi \rceil_{t3'}$ Sinee $\phi \wedge \psi$ is a Boolean formula and *(13, t 2)* $\models \phi \wedge \psi$, by Remark 2.2, *(13, t 3)* $\models \phi \wedge \psi$ and therefore *13* E $\mathcal{I}_{t_3}(\alpha)$. Henee $\mathcal{I}_{t_3}(\alpha) \cap B_{t2}(\alpha) \neq$ o. Furthermore, sinee ϕ is Boolean, by Remark 2.2, $\lceil \phi \rceil_{t3} = \lceil \phi \rceil_{t\text{-}z'}$ Thus, sinee $\lceil \phi \wedge \psi \rceil_{t3} < \lceil \phi \rceil_{t3}$ it follows that $\mathcal{I}_{t_3}(\alpha) < \mathcal{I}_{t_2}(\alpha)$. Hence, by property P_{KS}, $\mathcal{B}_{t_3}(\alpha) < \mathcal{I}_{t_3}(\alpha) \cap \mathcal{B}_{t_2}(\alpha)$. Fix an arbitrary γ E $\mathcal{B}_{t_3}(\alpha)$. Then γ E $\mathcal{I}_{t_3}(\alpha) \cap \mathcal{B}_{t_2}(\alpha)$ and, sinee $(\alpha, t2) \models B(\psi \rightarrow X)$, $(\gamma, t2) \models \psi \rightarrow X$· Since $\psi \rightarrow X$ is a Boolean formula, by Remark 2.2 $(\gamma, t_3) \models \psi \rightarrow X$· Sinee γ E $\mathcal{I}_{t_3}(\alpha)$ and $\mathcal{I}_{t_3}(\alpha) = \lceil \phi \wedge \psi \rceil_{t_3}$, $(\gamma, t3) \models \psi$. Thus $(\gamma, t3) \models X$· Henee $(\alpha, t3) \models BX$·

Conversely, fix a frame that does not satisfy property P_{KS}. Then there exist t_1,*t2,t3* and α sueh that $t_1 \rightarrowtail t_2$, $t_1 \rightarrowtail t_3$, $\mathcal{I}_{t_3}(\alpha) \cap \mathcal{B}_{t_2}(\alpha) \neq$ o, $\mathcal{I}_{t_3}(\alpha) < \mathcal{I}_{t_2}(\alpha)$ and $\mathcal{B}_{t_3}(\alpha) \nsubseteq \mathcal{I}_{t_3}(\alpha) \cap \mathcal{B}_{t_2}(\alpha)$. Let ». *q* and *r* be atomie propositions and construct a model where $\|p\| = \mathcal{I}_{t_2}(\alpha) xT$, $\|q\| = \mathcal{I}_{t_3}(\alpha)$ x Tand $\|T\| = (\mathcal{I}_{t_3}(\alpha) \cap B_{t2}(\alpha))$ x T. Then $(\alpha, t2) \models lp$ and, sinee $\mathcal{I}_{t_3}(\alpha) \subseteq \mathcal{I}_{t_2}(\alpha)$, if *w* E $\mathcal{I}_{t_3}(\alpha)$ then *(w,t)* $\models p \wedge q$ for every *t* E T. Thus, sinee

$\mathcal{I}_{t_3}(\alpha) \cap \mathcal{B}_{t_2}(\alpha) \neq \emptyset$, $(\alpha, t_2) \models \neg B(p \wedge q)$. Fix an arbitrary $w \in \mathcal{B}_{t_2}(\alpha)$; if $w \in \mathcal{I}_{t_3}(\alpha)$ then $(w,t_2) \models r$; if $w \notin \mathcal{I}_{t_3}(\alpha)$ then $(w,t_2) \models \neg q$; in either case $(w,t_2) \models q \rightarrow r$. Thus $(\alpha,t_2) \models B(q \rightarrow r)$. Hence $(\alpha,t_2) \models lp \wedge \neg B(p \wedge q) \wedge B(q \rightarrow r)$ and thus $(\alpha,t_1) \models \Diamond(lp \wedge \neg B(p \wedge q) \wedge B(q \rightarrow r))$. Since $\mathcal{I}_{t_3}(\alpha) = [q]_{t_3}$ and $\mathcal{I}_{t_2}(\alpha) = [p]_{t_2}$ and, by Remark 2.2, $[p]_{t_2} = [p]_{t_3}$ and $\mathcal{I}_{t_3}(\alpha) \subset \mathcal{I}_{t_2}(\alpha)$, it follows that $\mathcal{I}_{t_3}(\alpha) = [p \wedge q]_{t_3}$, so that $(\alpha,t_3) \models I(p \wedge q)$. Since $B_{t_3}(\alpha) \not\subseteq I_{t_3}(\alpha) \cap B_{t_2}(\alpha)$, there exists a $\beta \in B_{t_3}(\alpha)$ such that $\beta \notin \mathcal{I}_{t_3}(\alpha) \cap \mathcal{B}_{t_2}(\alpha)$. Then $(\beta, t_3) \models \neg r$ and therefore $(\alpha,t_3) \models \neg Br$. Thus $(\alpha,t_3) \not\models I(p \wedge q) \rightarrow Br$ and hence, $(\alpha,t_1) \not\models Q(I(p \wedge q) \rightarrow Br)$, so that axiom K8 is falsified at (a,t_1).

(4) Let ϕ be a Boolean formula, $\alpha \in \Omega$, $t \in T$ and suppose that $(\alpha,t) \models I\phi \wedge \neg A\neg\phi$. Then $\mathcal{I}_t(\alpha) = [\phi]_t$ and there exist $\beta \in \Omega$ that $(\beta,t) \models \phi$. Thus $\mathcal{I}_t(\alpha) \neq \emptyset$ and, by the above property, $\mathcal{B}_t(\alpha) \neq \emptyset$. Fix an arbitrary formula ψ and suppose that $(\alpha,t) \models B\psi$. Then, $\forall \omega \in \mathcal{B}_t(\alpha)$, $(\omega,t) \models \psi$. Since $\mathcal{B}_t(\alpha) \neq \emptyset$, there exists a $\gamma \in \mathcal{B}_t(\alpha)$. Thus $(\gamma,t) \models \psi$ and hence $(\alpha,t) \models \neg B\neg\psi$.

Conversely, fix a frame that does not satisfy property P_{WC}. Then there exist $\alpha \in \Omega$ and $t \in T$ such that $\mathcal{I}_t(\alpha) \neq \emptyset$ while $\mathcal{B}_t(\alpha) = \emptyset$. Let p be an atomic proposition and construct a model where $\|p\| = \mathcal{I}_t(\alpha) \times T$. Then $(\alpha,t) \models lp$. Furthermore, since $\mathcal{I}_t(\alpha) \neq \emptyset$, there exists a $\beta \in \mathcal{I}_t(\alpha)$. Thus $(\beta,t) \models p$ and hence $(\alpha,t) \models \neg A\neg p$. Since $\mathcal{B}_t(\alpha) = \emptyset$, $(\alpha,t) \models B\psi$ for every formula ψ, so that $(\alpha,t) \models Bp \wedge \neg Bp$. Thus WC is falsified at (α,t). Q.E.D.

Proof of the claim in Footnote 11. (BR3! $\Longrightarrow$ BR3). Fix arbitrary $E, F \in \mathcal{E}$ such that $E \subseteq F$ and $RF \cap E \neq \emptyset$. Then $E \cap F = E$, so that $(E \cap F) \in \mathcal{E}$ and $R_{E \cap F} = RE$. Thus, by (BR3'), $RE = RF \cap E$.

(BR3 + BR1 $\Longrightarrow$ BR3!). Let $E, F \in \mathcal{E}$ be such that $(E \cap F) \in \mathcal{E}$ and $RF \cap E \neq \emptyset$. By (BR1), $RF \subset F$ so that $RF \cap F = RF$. Hence

$$RF \cap (E \cap F) = RF \cap E. \tag{t}$$

Thus $RF \cap (E \cap F) \neq \emptyset$. Hence, since $E \cap F \subset F$, it follows from (BR3) that $R_{E \cap F} = RF \cap (E \cap F)$. Thus, by (t), $R_{E \cap F} = RF \cap E$. Q.E.D.

In order to prove Proposition 7.4 we need the following lemma. We shall throughout denote the complement of a set E by $\neg E$.

Lemma A.1. Let $\langle \Omega, \mathcal{E}, \mathbb{O}, R, V \rangle$ be a *rich* belief revision model. Then, for every formula $\phi \in \Phi_0$, $\|\phi\| = \emptyset$ if and only if ϕ is a contradiction (that is, $\neg\phi$ is a tautology).

Proof. If ϕ is a tautology then $\|\phi\| = \Omega$. If ϕ is a contradiction then $\neg\phi$ is a tautology and thus $\|\neg\phi\| = \neg\|\phi\| = \Omega$, so that $\|\phi\| = \emptyset$. If ϕ is neither a tautology nor a contradiction then it is equivalent to a formula of

the form $\left(\bigvee_{i=1}^{n}\left(\bigwedge_{j=1}^{m} Qij\right)\right)$ where eaeh *Qij* is either an atomie proposition or the negation of a atomie proposition (see Hamilton, 1987, p. 17, Corollary 1.20). By definition of rieh model, for every formula $\bigwedge_{j=1}^{m} Qij$, there is a state *wi* sueh that $w_i \in \bigwedge_{j=1}^{m} Q_{ij}$; Thus $\|\phi\| = \left\|\bigvee_{i=1}^{n}\left(\bigwedge_{j=1}^{m} \text{Qij}\right)\right\| = \bigcup_{i=1}^{n}\left\|\bigwedge_{j=1}^{m} Qij\right\| \supseteq \{w_1, \ldots, w_n\} \neq \emptyset$. Q.E.D.

PTOof of Proposiiion 7.4. Let $\langle \Omega, \mathcal{E}, \mathbb{O}, \mathsf{R}, V\rangle$ be a eomprehensive belief revision model and define $K = \{\psi \in \Phi_0 : \mathbf{R}_{\mathbb{O}} \subseteq \|\psi\|\}$. First we show that K is deduetively closed, that is, $K = [K]^{\mathrm{PL}}$. If $\psi \in K$ then $\psi \in [K]^{\mathrm{PL}}$, beeause $\psi \to \psi$ is a tautology; thus $K \subseteq [K]^{\mathrm{PL}}$. To show that $[K]^{\mathrm{PL}} \subseteq K$, let $\psi \in [K]^{\mathrm{PL}}$, that is, there exist $\phi_1, \ldots, \phi_n \in K$ sueh that $(\phi_1 \wedge \ldots \wedge \phi_n) \to \psi$ is a tautology. Sinee $\|\phi_1 \wedge \ldots \wedge \phi_n\| = \|\phi_1\| \cap \ldots \cap \|\phi_n\|$, and $\phi_i \in K$ (that is, $\mathbf{R}_{\mathbb{O}} \subseteq \|\phi_i\|$) for all $i = 1, \ldots, n$, it follows that $\mathbf{R}_{\mathbb{O}} \subseteq \|\phi_1 \wedge \ldots \wedge \phi_n\|$. Sinee $(\phi_1 \wedge \ldots \wedge \phi_n) \to \psi$ is a tautology, $\|(\phi_1 \wedge \ldots \wedge \phi_n) \to \psi\| = \Omega$, that is, $\|\phi_1 \wedge \ldots \wedge \phi_n\| \subseteq \|\psi\|$. Thus $\mathbf{R}_{\mathbb{O}}(\alpha) \subseteq \|\psi\|$, that is, $\psi \in K$. Next we show that $[K]^{\mathrm{PL}} \neq \Phi_0$ (eonsisteney). By definition of one-stage revision frame (see Definition 7.2), $\mathbb{O} \neq \emptyset$; thus, by property BR2, $\mathbf{R}_{\mathbb{O}} \neq \emptyset$. Choose an arbitrary atomie proposition $p \in S$. Then $\|(p \wedge \neg p)\| = \emptyset$ and therefore $\mathbf{R}_{\mathbb{O}} \not\subseteq \|(p \wedge \neg p)\|$, so that $(p \wedge \neg p) \notin K$. Sinee $K = [K]^{\mathrm{PL}}$, $(p \wedge \neg p) \notin [K]^{\mathrm{PL}}$. Next we show that AGM postulates $(\circledast 1)$–$(\circledast 5\mathrm{a})$ and $(\circledast 6)$–$(\circledast 8)$ are satisfied. For every formula $\phi \in \Phi_0$, define $K^{\circledast}(\phi) = \{\psi \in \Phi_0 : \mathbf{R}_{\|\phi\|} \subseteq \|\psi\|\}$ (note that, sinee the model is eomprehensive, for every $\phi \in \Phi_0$, $\|\phi\| \in \mathcal{E}$).

$(\circledast 1)$ Fix an arbitrary $\phi \in \Phi_0$. We need to show that $\{\psi \in \Phi_0 : \mathbf{R}_{\|\phi\|} \subseteq \|\psi\|\}$ is deduetively closed. We omit this proof sinee it is a repetition of the argument given above for K.

$(\circledast 2)$ Fix an arbitrary $\phi \in \Phi_0$. We need to show that $\phi \in K^{\circledast}(\phi)$, that is, that $\mathbf{R}_{\|\phi\|} \subseteq \|\phi\|$. This is an immediate eonsequenee of property BR1 of Definition 7.2.

$(\circledast 3)$ Fix an arbitrary $\phi \in \Phi_0$. We need to show that $K^{\circledast}(\phi) \subseteq [K \cup \{\phi\}]^{\mathrm{PL}}$. Let $\psi \in K^{\circledast}(\phi)$, that is, $\mathbf{R}_{\|\phi\|} \subseteq \|\psi\|$. First we show that $(\phi \to \psi) \in K$, that is, $\mathbf{R}_{\mathbb{O}} \subseteq \|\phi \to \psi\| = \neg\|\phi\| \cup \|\psi\|$. If $\mathbf{R}_{\mathbb{O}} \subseteq \neg\|\phi\|$ there is nothing to prove. Suppose therefore that $\mathbf{R}_{\mathbb{O}} \cap \|\phi\| \neq \emptyset$. Then, by property BR4 of Definition 7.2,

$$\mathbf{R}_{\|\phi\|} = \mathbf{R}_{\mathbb{O}} \cap \|\phi\|. \tag{3}$$

Fix an arbitrary $w \in \mathbf{R}_{\mathbb{O}}$. If $w \notin \|\phi\|$ then $w \in \|\neg\phi\|$ and thus $w \in \|\phi \to \psi\|$; if $w \in \|\phi\|$, then, by (3), $w \in \mathbf{R}_{\|\phi\|}$ and thus, sinee $\mathbf{R}_{\|\phi\|} \subseteq \|\psi\|$, $w \in \|\psi\|$, so that $w \in \|\phi \to \psi\|$. Henee $(\phi \to \psi) \in K$. It follows that $\psi \in [K \cup \{\phi\}]^{\mathrm{PL}}$.

$(\circledast 4)$ Fix an arbitrary $\phi \in \Phi_0$. We need to show that if $\neg\phi \notin K$ then $[K \cup \{\phi\}]^{\mathrm{PL}} \subseteq K^{\circledast}(\phi)$. Suppose that $\neg\phi \notin K$, that is, $\mathbf{R}_{\mathbb{O}} \not\subseteq \|\neg\phi\| = \neg\|\phi\|$,

that is, $\mathbf{R}_{\mathbb{O}} \cap \|\phi\| \neq \emptyset$. Then by property BR4 of Definition 7.2,

$$\mathbf{R}_{\|\phi\|} = \mathbf{R}_{\mathbb{O}} \cap \|\phi\|. \tag{4}$$

Let $X \in [K \cup \{\phi\}]^{\mathrm{PL}}$, that is, there exist $\phi_1, \ldots, \phi_n \in K \cup \{\phi\}$ sueh that $(\phi_1 \wedge \ldots \wedge \phi_n) \to X$ is a tautology. We want to show that $X \in K^{\circledast}(\phi)$, that is, $\mathbf{R}_{\|\phi\|} \subseteq \|X\|$. Sinee $(\phi_1 \wedge \ldots \wedge \phi_n) \to X$ is a tautology, $\|(\phi_1 \wedge \cdots \wedge \phi_n) \to X\| = \Omega$, that is, $\|(\phi_1 \wedge \ldots \wedge \phi_n)\| \subseteq \|X\|$. If $\phi_i \in K$ for every $i = 1, \ldots, n$, then $\mathbf{R}_{\mathbb{O}} \subseteq \|(\phi_1 \wedge \cdots \wedge \phi_n)\|$ and thus $\mathbf{R}_{\mathbb{O}} \subseteq \|X\|$. Hence, by (4), $\mathbf{R}_{\|\phi\|} \subseteq \|X\|$. If, for some $j = 1, \ldots, n$, $\phi_j \notin K$, then we ean assume (renumbering the formulas, if neeessary) that $\phi_i \in K$, for every $i = 1, \ldots, n - 1$, and $\phi_n \notin K$, whieh implies (sinee $\phi_i \in K \cup \{\phi\}$ for all $i = 1, \ldots, n$) that $\phi_n = \phi$. Since, by hypothesis, $(\phi_1 \wedge \cdots \wedge \phi_{n-1} \wedge \phi \to X$ is a tautology and, furthermore, it is propositionally equivalent to $(\phi_1 \wedge \cdots \wedge \phi_{n-1}) \to (\phi \to X)$, $\|(\phi_1 \wedge \cdots \wedge \phi_{n-1}) \to (\phi \to X)\| = \Omega$, that is, $\|\phi_1 \wedge \cdots \wedge \phi_{n-1}\| \subseteq \|\phi \to X\|$, so that, sinee $\mathbf{R}_{\mathbb{O}} \subseteq \|\phi_1 \wedge \cdots \wedge \phi_{n-1}\|$ (beeause $\phi_1, \ldots, \phi_{n-1} \in K$), $\mathbf{R}_{\mathbb{O}} \subseteq \|\phi \to X\|$. Thus $\mathbf{R}_{\mathbb{O}} \cap \|\phi\| \subseteq \|\phi\| \cap \|\phi \to X\| \subseteq \|X\|$. Hence, by (4), $\mathbf{R}_{\|\phi\|} \subseteq \|X\|$.

$(\circledast 5a)$ If ϕ is a eontradietion, $\|\phi\| = \emptyset$. By property BR1, $\mathbf{R}_{\|\phi\|} \subseteq \|\phi\|$. Henee $\mathbf{R}_{\|\phi\|} = \emptyset$ and, therefore, $K^{\circledast}(\phi) = \{\psi \in \Phi_0 : \mathbf{R}_{\|\phi\|} \subseteq \|\psi\|\} = \Phi_0$.

$(\circledast 6)$ If $\phi \leftrightarrow \psi$ is a tautology then $\|\phi \leftrightarrow \psi\| = \Omega$, that is, $\|\phi\| = \|\psi\|$. Henee $\mathbf{R}_{\|\phi\|} = \mathbf{R}_{\|\psi\|}$ and thus $K^{\circledast}(\phi) = \{X \in \Phi_0 : \mathbf{R}_{\|\phi\|} \subseteq \|X\|\} = \{X \in \Phi_0 : \mathbf{R}_{\|\psi\|} \subseteq \|X\|\} = K^{\circledast}(\psi)$.

$(\circledast 7)$ Fix arbitrary $\phi, \psi \in \Phi_0$. We need to show that $K^{\circledast}(\phi \wedge \psi) \subseteq [K^{\circledast}(\phi) \cup \{\psi\}]^{\mathrm{PL}}$. Let $X \in K^{\circledast}(\phi \wedge \psi)$, that is,

$$\mathbf{R}_{\|\phi \wedge \psi\|} \subseteq \|\chi\|. \tag{5}$$

First we show that $\mathbf{R}_{\|\phi\|} \subseteq \|(\phi \wedge \psi \to X\| = \neg\|\phi \wedge \psi\| \cup \|X\|$. If $\mathbf{R}_{\|\phi\|} \subseteq \neg\|\phi \wedge \psi\|$ there is nothing to prove. Suppose therefore that $\mathbf{R}_{\|\phi\|} \cap \|\phi \wedge \psi\| \neq \emptyset$. Then, by property (BR3) (with $E = \|\phi \wedge \psi\|$ and $F = \|\phi\|$),

$$\mathbf{R}_{\|\phi\|} \cap \|\phi \wedge \psi\| = \mathbf{R}_{\|\phi \wedge \psi\|}. \tag{6}$$

Fix an arbitrary $w \in \mathbf{R}_{\|\phi\|}$. If $w \notin \|\phi \wedge \psi\|$ then $w \in \|\neg(\phi \wedge \psi)\|$ and thus $w \in \|(\phi \wedge \psi \to \chi]$: if $w \in \|\phi \wedge \psi\|$, then by (5) and (6), $w \in \|X\|$ so that $w \in \|(\phi \wedge \psi \to X\|$. Henee $\mathbf{R}_{\|\phi\|} \subseteq \|(\phi \wedge \psi \to X\|$, that is, $(\phi \wedge \psi \to X) \in K^{\circledast}(\phi)$. Sinee $(\phi \wedge \psi \to X)$ is tautologieally equivalent to $(\psi \to (\phi \to X))$, and, by $(\circledast 1)$ (proved above), $K^{\circledast}(\phi$ is deduetively closed, $(\psi \to (\phi \to X)) \in K^{\circledast}(\phi)$. Furthermore, by $(\circledast 2)$ $\phi \in K^{\circledast}(\phi)$. Thus $\{\psi, (\psi \to (\phi \to \chi)), \phi\} \subseteq K^{\circledast}(\phi \cup \{\psi\}$ and therefore $X \in [K^{\circledast}(\phi \cup \{\psi\}]^{\mathrm{PL}}$.

$(\circledast 8)$ Fix arbitrary $\phi, \psi \in \Phi_0$. We need to show that if $\neg\psi \notin K^{\circledast}(\phi$ then $[K^{\circledast}(\phi \cup \{\psi\}]^{\mathrm{PL}} \subseteq K^{\circledast}(\phi \wedge \psi)$. Suppose that $\neg\psi \notin K^{\circledast}(\phi)$, that is,

$\mathbf{R}_{\|\phi\|} \not\subseteq \neg\|\psi\| = \|\neg\psi\|$, i.e., $\mathbf{R}_{\|\phi\|} \cap \|\psi\| \neq \emptyset$. Then by property (BR3/) (see Footnote 11),

$$\mathbf{R}_{\|\phi\wedge\psi\|} = \mathbf{R}_{\|\phi\|} \cap \|\psi\|. \tag{7}$$

Let $X \in [K^{\circledast}(\phi) \cup \{\psi\}]^{\mathrm{PL}}$, that is, there exist $\phi_1, \ldots \phi_n \in K^{\circledast}(\phi) \cup \{\psi\}$ such that $(\phi_1 \wedge \ldots \wedge \phi_n) \rightarrow X$ is a tautology. We want to show that $X \in K^{\circledast}(\phi \wedge \psi)$, that is, $\mathbf{R}_{\|\phi\wedge\psi\|} \subseteq [X]$- Since $(\phi_1 \wedge \cdot\cdot \wedge \phi_n) \rightarrow X$ is a tautology, $\|(\phi_1 \wedge \cdot\cdot \wedge \phi_n) \rightarrow X\| = \Omega$, that is, $\|(\phi_1 \wedge \cdot\cdot \wedge \phi_n)\| \subseteq \|X\|$. If $\phi_i \in K^{\circledast}(\phi)$ for every $i = 1, \ldots, n$, then $\mathbf{R}_{\|\phi\|} \subseteq \|(\phi_1 \wedge \cdots \wedge \phi_n)\|$ and thus $\mathbf{R}_{\|\phi\|} < \|X\|$. Hence, by (7), $\mathbf{R}_{\|\phi\wedge\psi\|} < \|X\|$. If, for some $j = 1, \ldots, n$, $\phi_j \notin K^{\circledast}(\phi)$, then we can assume (renumbering the formulas, if necessary) that $\phi_i \in K^{\circledast}(\phi)$, for every $i = 1, \ldots, n-1$, and $\phi_n \notin K^{\circledast}(\phi)$, which implies (since $\phi_i \in K^{\circledast}(\phi) \cup \{\psi\}$ for all $i = 1, \ldots, n$) that $\phi_n = \psi$. Since, by hypothesis, $(\phi_1 \wedge \ldots \wedge \phi_{n-1} \wedge \psi) \rightarrow X$ is a tautology and it is propositionally equivalent to $(\phi_1 \wedge \cdot\cdot \wedge \phi_{n-1}) \rightarrow (\psi \rightarrow X)$, $\|(\phi_1 \wedge \cdot\cdot \wedge \phi_{n-1}) \rightarrow (\psi \rightarrow X)\| = \Omega$, that is, $\|(\phi_1 \wedge \cdot\cdot \wedge \phi_{n-1})\| \subseteq \|\psi \rightarrow X\|$, so that, since $\mathbf{R}_{\|\phi\|} < \|(\phi_1 \wedge \cdots \wedge \phi_{n-1})\|$ (because $\phi_1, \ldots, \phi_{n-1} \in K^{\circledast}(\phi)$) $\mathbf{R}_{\|\phi\|} < \|\psi \rightarrow X\|$. Thus $\mathbf{R}_{\|\phi\|} \cap \|\psi\| < \|\psi\| \cap \|\psi \rightarrow X\| < \|X\|$. Hence, by (7), $\mathbf{R}_{\|\phi\wedge\psi\|} < \|X\|$.

Next we show that, if the model is rich, then $(\circledast 5b)$ is satisfied.

$(\circledast 5b)$ If the model is rich and ϕ is not a contradiction, then by Lemma A.1 $\|\phi\| \neq \emptyset$. Thus, by property BR2, $\mathbf{R}_{\|\phi\|} \neq \emptyset$. Fix an arbitrary $p \in S$. Since $\|p \wedge \neg p\| = \emptyset$, it follows that $\mathbf{R}_{\|\phi\|} \not\subseteq \|p \wedge \neg p\|$ and therefore $(p \wedge \neg p) \notin K^{\circledast}(\phi)$. Since, by $(\circledast 1)$ (proved above), $K^{\circledast}(\phi) = [K^{\circledast}(\phi)]^{\mathrm{PL}}$, it follows that $[K^{\circledast}(\phi)]^{\mathrm{PL}} \neq \Phi_0$. Q.E.D.

Before proving Proposition 7.5 we note the following.

Definition A.2. A set $E \subseteq 2^{\Omega}$ of events is called an *algebm* if it satisfies the following properties: (1) $\Omega \in E$, (2) if $E \in E$ then $\neg E \in E$ and (3) if $E, F \in$ [then $(E \cup F) \in \mathcal{E}$.[15]

Remark A.3. In a belief revision frame where E is an algebra, property *(BR3/)* (see Footnote 11) is equivalent to: $\forall E, F \in \mathcal{E}$,

$$\text{if } R_F \cap E \neq 0 \text{ then } R_{E \cap F} = R_F \cap E. \tag{BR3//}$$

PTOof of Pmposition 7.5. Let M be the set of maximally consistent sets (MeS) of formulas for a propositional logic whose set of formulas is Φ_0. For any $F \subseteq \Phi_0$ let $M_F = \{w \in M : F \subseteq w\}$. By Lindenbaum's lemma,

15 Note that from (1) and (2) it follows that $0 \in \mathcal{E}$ and from (2) and (3) it follows that if E, PEE then $(E \cap P) \in E$. In fact., from E, PEE we get, by (2), $\neg E, \neg F \in \mathcal{E}$ and thus, by (3), $(\neg E \cup \neg F) \in E$; using (2) again we get that $\neg(\neg E \cup \neg F) = (E \cap P) \in E$.

$M_F \neq 0$ if and only if F is a consistent set, that is, $[F]^{\mathrm{PL}} \neq \Phi_0$. To simplify the notation, for $\phi \in \Phi_0$ we write $\mathbb{M}_\phi$ rather than $\mathbb{M}_{\{\phi\}}$.

Define the following belief revision frame: $\Omega = \mathbb{M}$, $\mathcal{E} = \{\mathbb{M}_\phi : \phi \in \Phi_0\}$, $\mathbb{O} = \Omega$, $\mathbf{R}_\Omega = M_K$ and, for every $\phi \in \Phi_0$,

$$\mathbf{R}_{\mathbb{M}_\phi} = \begin{cases} \varnothing & \text{if } \phi \text{ is a contradiction} \\ \mathbb{M}_\phi \cap \mathbb{M}_K & \text{if } \phi \text{ is consistent and } \mathbb{M}_\phi \cap M_K \neq 0 \\ \mathbb{M}_{K^\circledast(\phi)} & \text{if } \phi \text{ is consistent and } \mathbb{M}_\phi \cap M_K = \mathbb{O}. \end{cases}$$

First of all, note that $\mathcal{E}$ is an algebra. (1) $\mathbb{M} \in \mathcal{E}$ since $\mathbb{M} = \mathbb{M}_{(p \vee \neg p)}$ where p is any atomic proposition. (2) Let $\phi \in \Phi_0$. Then $\mathbb{M}_\phi \in \mathcal{E}$ and $\neg\|\mathbb{M}_\phi\| = \{w \in \mathbb{M} : \phi \notin w\}$. By definition of MCS, for every $\omega \in \mathbb{M}$, $\phi \notin w$ if and only if $\neg\phi \in w$. Thus $\neg\|\mathbb{M}_\phi\| = \mathbb{M}_{\neg\phi} \in \mathcal{E}$. (3) Let $\phi, \psi \in \Phi_0$. Then $\mathbb{M}_\phi\ \mathbb{M}_\psi \in \mathcal{E}$ and, by definition of MCS, $\mathbb{M}_\phi \cup \mathbb{M}_\psi = \mathbb{M}_{\phi \vee \psi} \in \mathcal{E}$.

Next we show that the frame so defined is indeed a one-stage revision frame, that is, it satisfies properties (BR1)-(BR4) of Definition 7.2.

(BR1) We need to show that, for every $\phi \in \Phi_0$, $\mathbf{R}_{\mathbb{M}_\phi} \subseteq \mathbb{M}_\phi$. If ϕ is a contradiction, then $\mathbb{M}_\phi = 0$ and, by construction, $\mathbf{R}_{\mathbb{M}_\phi} = \mathbb{O}$. If ϕ is consistent and $\mathbb{M}_\phi \cap M_K \neq 0$ then $\mathbf{R}_{\mathbb{M}_\phi} = \mathbb{M}_\phi \cap M_K \subseteq \mathbb{M}_\phi$. If ϕ is consistent and $\mathbb{M}_\phi \cap M_K = 0$ then $\mathbf{R}_{\mathbb{M}_\phi} = \mathbb{M}_{K^\circledast(\phi)}$. Now, if $w' \in \mathbb{M}_{K^\circledast(\phi)}$ then $K^\circledast(\phi) \subseteq w'$ and, since by AGM postulate $(\circledast 2)$, $\phi \in K^\circledast(\phi)$, it follows that $\phi \in w'$, that is, $w' \in \mathbb{M}_\phi$. Hence $\mathbb{M}_{K^\circledast(\phi)} \subseteq \mathbb{M}_\phi$.

(BR2) We need to show that, for every $\phi \in \Phi_0$, if $\mathbb{M}_\phi \neq 0$ then $\mathbf{R}_{\mathbb{M}_\phi} \neq \mathbb{O}$. Now, $\mathbb{M}_\phi \neq 0$ if and only if ϕ is a consistent formula, in which case either $\mathbf{R}_{\mathbb{M}_\phi} = \mathbb{M}_\phi \cap M_K$ if $\mathbb{M}_\phi \cap M_K \neq 0$ or $\mathbf{R}_{\mathbb{M}_\phi} = \mathbb{M}_{K^\circledast(\phi)}$ if $\mathbb{M}_\phi \cap M_K = \mathbb{O}$. In the latter case, by AGM postulate $(\circledast 5b)$, $K^\circledast(\phi)$ is a consistent set and therefore, by Lindenbaum's lemma, $\mathbb{M}_{K^\circledast(\phi)} \neq \mathbb{O}$.

(BR3) Instead of proving (BR3) we prove the equivalent *(BR3!!)* (see Remark A.3 and footnote 11), that is, we show that, for every $\phi, \psi \in \Phi_0$, if $\mathbf{R}_{\mathbb{M}_\phi} \cap \mathbb{M}_\psi \neq 0$ then $\mathbf{R}_{\mathbb{M}_\phi \cap \mathbb{M}_\psi} = \mathbf{R}_{\mathbb{M}_\phi} \cap \mathbb{M}_\psi$. First note that, by definition of MCS, $\mathbb{M}_\phi \cap \mathbb{M}_\psi = \mathbb{M}_{\phi \wedge \psi}$. Since $\mathbf{R}_{\mathbb{M}_\phi} \neq \mathbb{O}$, ϕ is a consistent formula and either $\mathbf{R}_{\mathbb{M}_\phi} = \mathbb{M}_\phi \cap \mathbb{M}_K$, if $\mathbb{M}_\phi \cap \mathbb{M}_K \neq \mathbb{O}$, or $\mathbf{R}_{\mathbb{M}_\phi} = \mathbb{M}_{K^\circledast(\phi)}$, if $\mathbb{M}_\phi \cap \mathbb{M}_K = \mathbb{O}$. Suppose first that $\mathbb{M}_\phi \cap M_K \neq \mathbb{O}$. Then $\mathbf{R}_{\mathbb{M}_\phi} \cap \mathbb{M}_\psi = \mathbb{M}_\phi \cap M_K \cap \mathbb{M}_\psi = \mathbb{M}_{\phi \wedge \psi} \cap M_K \neq \mathbb{O}$, and, thus, by construction, $\mathbf{R}_{\mathbb{M}_{\phi \wedge \psi}} = \mathbb{M}_{\phi \wedge \psi} \cap M_K$. Thus $\mathbf{R}_{\mathbb{M}_\phi \cap \mathbb{M}_\psi} = \mathbf{R}_{\mathbb{M}_{\phi \wedge \psi}} = \mathbf{R}_{\mathbb{M}_\phi} \cap \mathbb{M}_\psi$. Suppose now that $\mathbb{M}_\phi \cap M_K = \mathbb{O}$. Then, by construction, $\mathbf{R}_{\mathbb{M}_\phi} = \mathbb{M}_{K^\circledast(\phi)}$ and, since $\mathbb{M}_{\phi \wedge \psi} \cap M_K = \mathbb{M}_\phi \cap \mathbb{M}_\psi \cap M_K \subseteq \mathbb{M}_\phi \cap M_K = 0$ we also have that $\mathbf{R}_{\mathbb{M}_{\phi \wedge \psi}} = \mathbb{M}_{K^\circledast(\phi \wedge \psi)}$. Thus we need to show that $\mathbb{M}_{K^\circledast(\phi \wedge \psi)} = \mathbb{M}_{K^\circledast(\phi)} \cap \mathbb{M}_\psi$. By hypothesis, $\mathbf{R}_{\mathbb{M}_\phi} \cap \mathbb{M}_\psi \neq \mathbb{O}$, that is, $\mathbb{M}_{K^\circledast(\phi)} \cap \mathbb{M}_\psi \neq \mathbb{O}$. This implies that $\neg\psi \notin K^\circledast(\phi)$.[16] Hence, by AGM

16 Suppose that $\neg\psi \in K^\circledast(\phi)$. Then, for every $w \in \mathbb{M}_{K^\circledast(\phi)}$, $\omega \supseteq K^\circledast(\phi)$ and, therefore, $\neg\psi \in w$. But this implies that $\mathbb{M}_{K^\circledast(\phi)} \cap \mathbb{M}_\psi = \mathbb{O}$.

postulates $(\circledast 7)$ and $(\circledast 8)$,

$$\left[K^{\circledast}(\phi) \cup \{\psi\}\right]^{\mathrm{PL}} = K^{\circledast}(\phi \wedge \psi). \tag{8}$$

Let $w \in \mathbb{M}_{K^{\circledast}(\phi \wedge \psi)}$. Then $K^{\circledast}(\phi \wedge \psi) < w$, since $K^{\circledast}(\phi) \cup \{\psi\} < [K^{\circledast}(\phi) \cup \{\psi\}]^{\mathrm{PL}}$, it foliows from (8) that $K^{\circledast}(\phi) \cup \{\psi\} < w$. Thus $w \in \mathbb{M}_{K^{\circledast}(\phi)} \cap \mathbb{M}_{\psi}$. Conversely, let $w \in \mathbb{M}_{K^{\circledast}(\phi)} \cap \mathbb{M}_{\psi}$. Then $K^{\circledast}(\phi) \cup \{\psi\} < w$. Hence, by definition of MCS, $[K^{\circledast}(\phi) \cup \{\psi\}]^{\mathrm{PL}} \subseteq w$. It foliows from (8) that $K^{\circledast}(\phi \wedge \psi) < w$, that is, $w \in \mathbb{M}_{K^{\circledast}(\phi \wedge \psi)}$. Thus $\mathbb{M}_{K^{\circledast}(\phi \wedge \psi)} = \mathbb{M}_{K^{\circledast}(\phi)} \cap \mathbb{M}_{\psi}$, that is, $\mathbf{R}_{\mathbb{M}_{\phi} \cap \mathbb{M}_{\psi}} = \mathbf{R}_{\mathbb{M}_{\phi}} \cap \mathbb{M}_{\psi}$.

(BR4) Since $\mathbb{O} = \Omega$ and, by construction, $\mathbf{R}_{\Omega} = M_K$, we need to show that, for every formula ϕ, if $\mathbb{M}_{\phi} \cap M_K \neq 0$ then $\mathbf{R}_{\mathbb{M}_{\phi}} = \mathbb{M}_{\phi} \cap M_K$. But this is true by construction.

Consider now the model based on this frame given by the foliowing valuation: for every atomie proposition and for every $w \in \Omega$, $w \models p$ if and only if $p \in w$. It is weli-known that in this model, for every formula ϕ,[17]

$$\|\phi\| = \mathbb{M}_{\phi}. \tag{9}$$

Note also the foliowing (see Chelias, 1984, Theorem 2.20, p. 57): $\forall F \subseteq \Phi_0, \forall \phi \in \Phi_0$,

$$\phi \in [F]^{\mathrm{PL}} \text{ if and only if } \phi \in w, \ \forall \omega \in M_F. \tag{10}$$

We want to show that (1) $K = \{\psi \in \Phi_0 : \mathbf{R}_{\mathbb{O}} \subseteq \|\psi\|\}$ and, (2) for every $\phi \in \Phi_0$, $K^{\circledast}(\phi) = \{\psi \in \Phi_0 : \mathbf{R}_{\|\phi\|} \subseteq \|\psi\|\}$.

(1) By construction, $\mathbb{O} = \Omega$ and $\mathbf{R}_{\Omega} = M_K$ and, by (9), for every fornula ψ, $\|\psi\| = \mathbb{M}_{\psi}$. Thus we need to show that, for every formula ψ, $\psi \in K$ if and only if $M_K \subseteq \mathbb{M}_{\psi}$. Let $\psi \in K$ and fix an arbitrary $w \in M_K$. Then $K \subseteq w$ and thus $\psi \in w$, so that $\omega \in \mathbb{M}_{\psi}$. Conversely, suppose that $M_K \subseteq \mathbb{M}_{\psi}$. Then $\psi \in w$, for every $w \in M_K$. Thus, by (10), $\psi \in [K]^{\mathrm{PL}}$. By AGM postulate $(\circledast 1)$, $K = [K]^{\mathrm{PL}}$. Hence $\psi \in K$.

(2) Fix an arbitrary formula ϕ. First we show that $K^{\circledast}(\phi) \subseteq \{\psi \in \Phi_0 : \mathbf{R}_{\mathbb{M}_{\phi}} \subseteq \|\psi\|\}$. Let $\psi \in K^{\circledast}(\phi)$. If ϕ is a contradiction, $\mathbf{R}_{\mathbb{M}_{\phi}} = 0$ and there is nothing to prove. If ϕ is consistent then two cases are possible:

17 The proof is by induction on the complexity of ϕ. If $\phi = p$, for some sentence letter p, then the statement is true by construction. Now suppose that the statement is true of $\phi_1, \phi_2 \in \Phi_0$; we want to show that it is true for $\neg\phi_1$ and for $(\phi_1 \vee \phi_2)$. By definition, $w \models \neg\phi_1$ if and only if $w \not\models \phi_1$ if and only if (by the induction hypothesis) $\phi_1 \notin w$ if and only if, by definition of MCS, $\neg\phi_1 \in w$. By definition, $w \models (\phi_1 \vee \phi_2)$ if and only if either $w \models \phi_1$, in which case, by the induction hypothesis, $\phi_1 \in w$, or $w \models \phi_2$, in which case, by the induction hypothesis, $\phi_2 \in w$. By definition of MCS, $(\phi_1 \vee \phi_2) \in w$ if and only if either $\phi_1 \in w$ or $\phi_2 \in w$.

(i) $\mathbb{M}_\phi \cap M_K = 0$ and (ii) $\mathbb{M}_\phi \cap M_K \neq 0$. **In** case (i) $\mathbf{R}_{\mathbb{M}_\phi} = \mathbb{M}_{K^\circledast(\phi)}$. Since, by hypothesis, $\psi \in K^\circledast(\phi)$, $\mathbb{M}_{K^\circledast(\phi)} \subseteq \mathbb{M}_\psi$ and, by (9), $\mathbb{M}_\psi = \|\psi\|$. Thus $\mathbf{R}_{\mathbb{M}_\phi} < \|\psi\|$. **In** case (ii), $\mathbf{R}_{\mathbb{M}_\phi} = \mathbb{M}_\phi \cap M_K$. First of all, note that $\mathbb{M}_\phi \cap \mathbb{M}_K = \mathbb{M}_{K\cup\{\phi\}}$. Secondly, it must be that $\neg\phi \notin K$.[18] Hence, by AGM postulates $(\circledast 3)$ and $(\circledast 4)$, $K^\circledast(\phi) = [K \cup \{\phi\}]^{\mathrm{PL}}$. Since, by hypothesis, $\psi \in K^\circledast(\phi)$, $\psi \in [K\cup\{\phi\}]^{\mathrm{PL}}$. Hence, by (10), $\psi \in w$, for every $w \in \mathbb{M}_{K\cup\{\phi\}}$. Thus $\mathbb{M}_{K\cup\{\phi\}} < \mathbb{M}_\psi$. Hence, since $\mathbf{R}_{\mathbb{M}_\phi} = \mathbb{M}_\phi \cap M_K = \mathbb{M}_{K\cup\{\phi\}}$, $\mathbf{R}_{\mathbb{M}_\phi} < \mathbb{M}_\psi$. Next we show that $\{\psi \in \Phi_0 : \mathbf{R}_{\mathbb{M}_\phi} \subseteq \|\psi\|\} \subseteq K^\circledast(\phi)$. Suppose that $\mathbf{R}_{\mathbb{M}_\phi} \subseteq \|\psi\| = \mathbb{M}_\psi$. If ϕ is a contradiction, then, by AGM postulate $(\circledast 5a)$, $K^\circledast(\phi) = \Phi_0$ and, therefore, $\psi \in K^\circledast(\phi)$. If ϕ is not a contradiction, then either (i) $\mathbb{M}_\phi \cap M_K = 0$ or (ii) $\mathbb{M}_\phi \cap M_K \neq 0$. **In** case (i) $\mathbf{R}_{\mathbb{M}_\phi} = \mathbb{M}_{K^\circledast(\phi)}$. Thus, since, by hypothesis, $\mathbf{R}_{\mathbb{M}_\phi} < \mathbb{M}_\psi$, we have that $\mathbb{M}_{K^\circledast(\phi)} < \mathbb{M}_\psi$, that is, for every $w \in \mathbb{M}_{K^\circledast(\phi)}$, $\psi \in w$. By (10) $\psi \in [K^\circledast(\phi)]^{\mathrm{PL}}$ and, by AGM postulate $(\circledast 1)$, $[K^\circledast(\phi)]^{\mathrm{PL}} = K^\circledast(\phi)$. Thus $\psi \in K^\circledast(\phi)$. **In** case (ii), $\mathbf{R}_{\mathbb{M}_\phi} = \mathbb{M}_\phi \cap M_K$. Thus, since, by hypothesis, $\mathbf{R}_{\mathbb{M}_\phi} \subseteq \mathbb{M}_\psi$, we have that $\mathbb{M}_\phi \cap M_K \subseteq \mathbb{M}_\psi$, from which it follows (since $\mathbb{M}_\phi \cap M_K = \mathbb{M}_{K\cup\{\phi\}}$) that $\mathbb{M}_{K\cup\{\phi\}} \subseteq \mathbb{M}_\psi$. This means that, for every $w \in \mathbb{M}_{K\cup\{\phi\}}$, $\psi \in w$. Hence, by (10), $\psi \in [K \cup \{\phi\}]^{\mathrm{PL}}$. Since $\mathbb{M}_\phi \cap \mathbb{M}_K \neq 0$, $\neg\phi \notin K$ and, therefore, by AGM postulates $(\circledast 3)$ and $(\circledast 4)$, $K^\circledast(\phi) = [K \cup \{\phi\}]^{\mathrm{PL}}$. Thus $\psi \in K^\circledast(\phi)$. Q.E.D.

Proof of Proposiiion 7.6. **In** view of Corollary 6.3 it is sufheient to show that (1) axiom CMP is characterized by property $\mathrm{P_{CMP}}$ and (2) Icon is characterized by seriality of *It*.

(1) Fix an arbitrary model based on a frame that satisfies property $\mathrm{P_{CMP}}$. Fix arbitrary $\alpha \in \Omega$, $t_a \in T$ and Boolean formula ϕ and suppose that $(\alpha, ta) \models \neg A\neg\phi$. Let $E = \lceil\phi\rceil to$. Then $E \neq 0$. We want to show that $(\alpha, t_0) \models \Diamond I\phi$. By property $\mathrm{P_{CMP}}$, there exists $at \in T$ such that $ta \rightarrowtail t$ and $It(w) = E$. Since ϕ is Boolean, $\lceil\phi\rceil_{t_0} = \lceil\phi\rceil_t$. Thus $(\alpha, t) \models I\phi$ and hence $(\alpha, t_0) \models \Diamond I\phi$.

Conversely, fix a frame that violates property $\mathrm{P_{CMP}}$. Then there exist $\alpha \in \Omega$, $ta \in T$ and $E \in 2^\Omega \setminus \{\varnothing\}$ such that, $\forall t \in T$, $if ta \rightarrowtail t$ then $It(w) \neq E$. Construct a model where, for some atomie proposition p, $\|p\| = E \times T$. Then, $\forall t \in T$ with $ta \rightarrowtail t$, $(\alpha, t) \not\models lp$. Thus $(\alpha, ta) \not\models Olp$.

(2) Fix an arbitrary model based on a frame where $\mathcal{I}_t$ is serial and suppose that $I(\phi \wedge \neg\phi)$ is not valid, that is, for some $\alpha \in \Omega$, $t \in T$ and formula ψ, $(\alpha, t) \models I(\phi\wedge\neg\phi)$. Then $\mathcal{I}_t(\alpha) = \lceil\phi \wedge \neg\phi\rceil_t$. But $\lceil\phi \wedge \neg\phi\rceil_t = 0$, while by seriality $\mathcal{I}_t(\alpha) \neq 0$, yielding a contradiction.

Conversely, fix a frame where $\mathcal{I}_t$ is not serial, that is, there exist $t \in T$ and $\alpha \in \Omega$ such that $I_t(\alpha) = 0$. Since, for every formula ϕ, $\lceil\phi \wedge \neg\phi\rceil t = 0$, it follows that $(\alpha, t) \models I(\phi \wedge \neg\phi)$ so that $\neg I(\phi \wedge \neg\phi)$ is not valid. Q.E.D.

18 If $\neg\phi \in K$ then $\neg\phi \in w$ for every $w \in \mathbb{M}_K$ and therefore $\mathbb{M}_\phi \cap \mathbb{M}_K = 0$.

References

Alchourrón, C.E., Gärdenfors, P. & Makinson, D. (1985). On the Logic of Theory Change: Partial Meet Contraction and Revision Functions. *The Joumal of Symbolic Logic, 50(2):510-530.*

Baltag, A. & Smets, S. (2006). Dynamic Belief Revision over Multi-Agent Plausibility Modeis. **In** Bonanno, G., van der Hoek, W. & Wooldridge, M., eds., *Proceedinqs of the 'lih Conference on Logic and the Foundations of Game and Decision Theoru (LOFT*$^{JO6)}$, pp. 11-24.

van Benthem, .1. (2004). Dynamic Logic for Belief Revision. *Joumal of Applied Non-Classical Logies, 14(2):129-155.*

Blackburn, P., de Rijke, M. & Venema, Y. (2001). *Modal Logic,* Vol. 53 of *Cambridqe Tracis in Theoretical Computer Science.* Cambridge University Press.

Board, O. (2002). Dynamic Interactive Epistemology. *Games and Economie Behaoiour,* 49(1):49-80.

Bonanno, G. (2005). A Simple Modal Logic for Belief Revision. *Synthese,* 147(2):193-228.

Bonanno, G. (2007a). Axiomatic Characterization of the AGM Theory of Belief Revision in a Temporal Logic. *Artificial !ntel!igence, 171(2-3):144-*160.

Bonanno, G. (2007b). A Temporal Logic for Belief Revision Based on Plausibility. Manuscript, University of California, Davis.

Bonanno, G. (2008). A Sound and Complete Temporal Logic for Belief Revision. **In** Dégremont, C., Keiff, L. & Rückert, H., eds., *Essays in hortour of Shahid Rahman,* Tributes. College Publications. To appeal'.

Boutilier, C. (1996). Iterated Revision and Minimal Change of Conditional Beliefs. *Joumal of Fhilosophical Logic, 25(3):263-305.*

Brandenburger, A. & Keisier, J. (2006). An Impossibility Theorem on Beliefs in Games. *Studia Logica, 84(2):211-240.*

Chellas, B. (1984). *Modal Logic: an Introduetion.* Cambridge University Press.

Darwiche, A. & Pearl, .1. (1997). On the Logic of Iterated Belief Revision. *Artificial !ntel!igence, 89(1-2):1-29.*

van Ditmarseh, H. (2005). Prolegomena to Dynamic Logic for Belief Revision. *Synthese, 147(2):229-275.*

van Ditmarseh, H. & Labuschagne, W. (2007). My Beliefs about Your Beliefs: a Case Study in theory of mind and epistemic logic. *Synthese,* 155(2):191-209.

van Ditmarseh, H., van der Hoek, W. & Kooi, B. (2007). *Dynamic Epistemic Logic,* Vol. 337 of *Synthese Libraru.* Springer.

Friedman, N. & Halpern, .1. (1999). Belief Revision: a Critique. *Joumal of Logic, Language, and Injormation, 8(4):401-420.*

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

Goranko, V. & Passy, S. (1992). Using the Universal Modality: Gains and Questions. *Joumal of Logic and Computation, 2(1):5-30.*

Grove, A. (1988). Two Modellings for Theory Change. *Joumal of Philosophical Logic, 17(2):157-170.*

Hamilton, A.G. (1987). *Logic fOT Mathematicians.* Cambridge University Press.

Hintikka, .1. (1962). *Knowledge and Belief.* Cornell University Press.

Humberstone, I.L. (1987). The Modal Logic of All and Only. *Noire Dame Joumal of Formol Logic, 28(2):177-188.*

Katsuno, H. & Mendelzon, A.O. (1991). Propositional Knowledge Base Revision and Minimal Change. *Artificial !ntel!igence, 52(3):263-294.*

Kripke, S. (1963). ASemantical Analysis of Modal Logic I: Normal Propositional Calculi. *Zeitschrift fÜT Mathematische Logik und Crundlaqeti der Mathematik, 9(5--6):67-96.*

Leitgeb, H. & Segerberg, K. (2007). Dynamic Doxastic Logic: Why, How and Where To? *Synthese, 155(2):167-190.*

Levesque, H.J. (1990). All I Know: a Study in Autoepistemic Logic. *ATtificial !ntel!igence,* 42(2-3):263-309.

Nayak, A., Pagnucco, M. & Peppas, P. (2003). Dynamic Belief Revision Operators. *Artificial !ntel!igence, 146(2):193-228.*

Rott, H. (2001). *Change, Choice and Inference.* Clarendon Press, Oxford.

Rott, H. (2006). Shifting Priorities: Simple Representations for Twenty-seven Iterated Theory Change Operators. **In** Lagerlund, H., Lindström, S. & Sliwinski, R., eds., *Modality Matters: Twenty-Five Essays in HonOUT of Krister Seqerberq,* Uppsala Philosophical Studies Volume 53, pp. 359-384. Uppsala Universitet.

Segerberg, K. (1995). Belief Revision from the Point of View of Doxastic Logic. *Bulletin of the IGPL, 3(4):535-553.*

Segerberg, K. (1999). Two Traditions in the Logic of Belief: Bringing Them Together. **In** 0 hlbach, H.J. & Reyle, U., eds., *Logic, Language and Reasoning: Essays in HonOUT of Dov Gabbay,* Vol. 5 of *Trends in Logic,* pp. 135-147. Klmver Academie Publishers.

Suzumura, K. (1983). *Rational Choice, Collective Decisions and Social Wel-[are.* Cambridge University Press.

Zvesper, .1. (2007). How to keep on changing your mind, dynamically. Manuscript. University of Amsterdam.

Vet More Modal Logies of Preferenee Change and Belief Revision

Jan van Eijck

Centre for Mathematics and Computer Scîence (CWI)
Kruislaan 413
1098 SJ Amsterdam, The Netherlands
jve@cwi.nl

Abstract

We contrast Bonanno's 'Belief Revision in a Temporal Framework' (Bonanno, 2008) with preference change and belief revision from the perspective of dynamic epistemic logic (DEL). For that, we extend the logic of communication and change of van Benthem et al. (2006b) with relational substitutions (van Benthem, 2004) for preference change, and show that this does not alter its properties. Next we move to a more constrained context where belief and knowledge can be defined from preferences (Grove, 1988; Board, 2002; Baltag and Smets, 2006, 2008b), prove completeness of a very expressive logic of belief revision, and define a mechanism for updating belief revision models using a combination of action priority update (Baltag and Smets, 2008b) and preference substitution (van Benthem, 2004).

1 Reconstructing AGM style belief revision

Bonanno's paper offers a rational reconstruction of Alchourrón Gärdenfors Makinson style belief revision (AGM belief revision) (Alchourrón et al., 1985; see also Gärdenfors, 1988 and Gärdenfors and Rott, 1995), in a framework where modalities B for single agent belief and I for being informed are mixed with a next time operator $\bigcirc$ and its inverse $\bigcirc^{-1}$.

Both the AGM framework and Bonanno's reconstruction of it do not explicitly represent the triggers that cause belief change in the first place. $I\varphi$ expresses that the agent is informed that φ, but the communicative action that causes this change in information state is not represented. Also, φ is restricted to purely propositional formulas. Another limitation that Bonanno's reconstruction shares with AGM is that it restricts attention to a single agent: changes of the beliefs of agents about the beliefs of other agents are not analyzed. **In** these respects (Bonanno, 2008) is close to Dynamic Doxastic Logic (DDL), as developed by Segerberg (1995, 1999).

AGM style belief revision was proposed more than twenty years ago, and has grown into a paradigm in its own right in artificial intelligence. **In** the

Krzysztof R. Apt, Robert van Rooij (eds.). *New Perspectives on Games and Interaction.* Texts in Logic and Games 4, Amsterdam University Press 2008, pp. 81-104.

meanwhile rieh frameworks of dynamic epistemic logic have emerged that are quite a bit more ambitious in their goals than AGM was when it was first proposed. AGM analyzes operations $+\varphi$ for expanding with φ, $-\varphi$ for retracting φ and $*\varphi$ for revising with φ. **It** is formulated in a purely syntactic way, it hardly addresses issues of semantics, it does not propose sound and complete axiomatisations. **It** did shine in 1985, and it still shines now, but perhaps in a more modest way.

Bonanno's paper creates a nice link between this style of belief revision and epistemicjdoxastic logic. While similar in spirit to Segerberg's work, it addresses the question of the rational reconstruction of AGM style belief revision more explicitly. This does add quite a lot to that framework: deal' semantics, and a sound and complete axiomatisation. Still, it is fair to say that this rational reconstruction, nice as it is, also inherits the limitations of the original design.

2 A broader perspective

Meanwhile, epistemic logic has entered a different phase, with a new focus on the epistemic and doxastic effects of information updates such as public announcements (Plaza, 1989; Gerbrandy, 1999). Public announcements are interesting because they create common knowiedge, so the new focus on information updating fostered an interest in the evolution of multi-agent knowledge and belief under acts of communication.

Public armouncement was generalized to updates with 'action modeis' that can express a wide range of communications (private announcements, group announcements, secret sharing, lies, and so on) in (Baltag et al., 1998) and (Baltag and Moss, 2004). A further generalization to a complete logic of communication and change, with enriched actions that allow changing the facts of the world, was provided by van Benthem et al. (2006b). The textbook treatment of dynamic epistemic logic in (van Ditmarsch et al., 2007) bears witness to the fact that this approach is by now weil established.

The above systems of dynamic epistemic logic do provide an account of knowledge or belief update, but they do not analyse belief revision in the sense of AGM. Information updating in dynamic epistemic logic is monotonie: facts that are announced to an audience of agents cannot be unlearnt. Van Benthem (2004) calls this 'belief change under hard information' or 'eliminative belief revision'. See also (van Ditmarseh, 2005) for refiection on the distinction between this and belief change under soft information.

Assume a state of the world where p actually is the case, and where you know it, but I do not. Then public armouncement of p will have the effect that I get to know it, but also that you know that I know it, that I know that you know that I know it, in short, that p becomes common knowiedge. But if this announcement is followed by an announcement of $\neg p$, the effect will be inconsistent knowledge states for both of us.

It is clear that AGM deals with belief revision of a different kind: 'belief change under soft information' or 'non-eliminative belief revision'. **In** (van Benthem, 2004) it is sketched how this can be incorporated into dynamic epistemic logic, and in the closely related (Baltag and Smets, 2008b) a theory of 'doxastic actions' is developed that can be seen as a further step in this direction.

Belief revision under soft information can, as Van Benthem observes, be modelled as change in the belief accessibilities of a model. This is different from public announcement, which can be viewed as elimination of worlds while leaving the accessibilities untouched.

Agent i believes that φ in a given world w if it is the case that φ is true in all worlds t that are reachable from wand that are minimal for a suitable plausibility ordering relation $\leq_i$. **In** the dynamic logic of belief revision these accessibilities can get updated in various ways. An example from (Rott, 2006) that is discussed by van Benthem (2004) is $\Uparrow A$ for so-called 'Iexicographic upgrade': all A worlds get promoted past all non-A worlds, while within the A worlds and within the non-A worlds the preference relation stays as before. Clearly this relation upgrade has as effect that it creates belief in A. And the belief upgrade can be undone: a next update with $\Uparrow\neg A$ does not result in inconsistency.

Van Benthem (2004) gives a complete dynamic logic of belief upgrade for the belief upgrade operation $\Uparrow A$, and another one for a variation on it, $\uparrow A$, or 'elite change', that updates a plausibility ordering to a new one where the best A worlds get promoted past all other worlds, and for the rest the old ordering remains unchanged.

This is Laken one step Iurt.her in a general logic Ior changing preferences, in (van Benthem and Liu, 2004), where upgrade as relation change is handled for (reftexive and transitive) preference relations $\leq_i$, by means of a variation on product update called product upgrade. The idea is to keep the domain, the valuation and the epistemic relations the same, but to reset the preferences by means of a substitution of new preorders for the preference relations.

Treating knowledge as an equivalence and preference as a preorder, without constraining the way in which they relate, as is done by van Benthem and **Liu** (2004), has the advantage of generality (one does not have to specify what 'having a preference' means), but it makes it harder to use the preference relation for modelling belief change. If one models 'regret' as preferring a situation that one knows to be false to the cutrent situation, then it follows that one can regret things one cannot even conceive. And using the same preference relation for belief looks strange, for this would allow beliefs that are known to be false. Van Benthem (private communication) advised me not to lose sleep over such philosophical issues. If we follow that

advice, and call 'belief' what is true in all most preferred worlds, we can still take comfort from the fact that this view entails that one can never believe one is in a bad situation: the belief-accessible situations are by definition the best conceivable worlds. Anyhow, proceeding from the assumption that knowledge and preferenee are independent basic relations and then studying possible relations between them has turned out very fruitful: the recent theses by Girard (2008) and Liu (2008) are rieh sourees of insight in what a logical study of the interaction of knowledge and preferenee may reveal.

Here we will explore two avenues, different from the above but related to it. First, we assume nothing at all about the relation between knowledge on one hand and preferenee on the other. We show that the dynamic logic of this (including updating with suitable finite update modeis) is complete and decidabIe: Theorem 3.1 gives an extension of the reducibility result for LCC, the general logic of communication and change proposed and investigated in (van Benthem et al., 2006b).

Next, we move closer to the AGM perspective, by postulating a close conneetion between knowiedge, belief and preference. One takes prefereenees as primary, and imposes minimal conditions to allow a definition of knowledge from preferences. The key to this is the simple observation in Theorem 4.1 that a preorder can be turned into an equivalence by taking its symmetrie closure if and only if it is weakly connected and conversely weakly connected. This means that by starting from weakly and converse weakly connected preorders one can interpret their symmetrie closures as knowledge relations, and use the preferences themselves to define conditional beliefs, in the well known way that was first proposed in (Boutilier, 1994) and (Halpern, 1997). The mulLi-agenL version of this kind of condit.ional belief was further explored in (van Benthem et al., 2006a) and in (Baltag and Smets, 2006, 2008b). We extend this to a complete logic of regular doxastic programs for belief revision models (Theorem 4.3), useful for reasoning about common knowiedge, common conditional belief and their interaction. Finally, we make a formal proposal for a belief change mechanism by means of a combination of action model update in the style of (Baltag and Smets, 2008b) and plausibility substitution in the style of (van Benthem and Liu, 2004).

3 Preference change in Lee

An epistemic preferenee model M for set of agents I is a tuple (W, V, R, P) where W is a non-empty set of worlds, V is a propositional valuation, Ris a function that maps each agent i to a relation $R;$ (the epistemic relation for i), and P is a function that maps each agent i to a preferenee relation $Pi.$ There are no conditions at all on the $R;$ and the Pi (just as there are no constraints on the $R;$ relations in LCC (van Benthem et al., 2006b)).

We fix a **PDL** style language for talking about epistemic preference models (assume *p* ranges over a set of basic propositions *Prop* and i over a set of agents *I):*

$$\begin{array}{ll} \varphi & \mathbf{Tip}\ \mathrm{I}\neg\varphi\ \mathrm{I}\varphi_1 \wedge \varphi_2\ \mathrm{I}[\pi]\varphi \\ tt & \sim_i\ \mathrm{I}\geq_i\ |?\varphi \mid \pi_1;\pi_2 \mid \pi_1 \cup \pi_2 \mid \pi^* \end{array}$$

This is to be interpreted in the usual **PDL** marmer, with $[\![\pi]\!]^{\mathbf{M}}$ giving the relation that interprets relational expression *tt* in M = *(W, V, R, P),* where $\sim_i$ is interpreted as the relation *R;* and $\geq_i$ as the relation *Pi,* and where the complex modalities are handled by the regular operations on relations. We employ the usual abbreviations: $\bot$ is shorthand for $\neg$T, φ_1 V φ_2 is shorthand for -,($\neg\varphi_1 \wedge \neg\varphi_2$), $\varphi_1 \rightarrow \varphi_2$ is shorthand for -,($\varphi_1 \wedge \varphi_2$), $\varphi_1 \leftrightarrow \varphi_2$ is shorthand for $(\varphi_1 \rightarrow \varphi_2) \wedge (\varphi_2 \rightarrow \varphi_1)$, and $\langle\pi\rangle\varphi$ is shorthand for $\neg[\pi]\neg\varphi$.

The formula $[\pi]\varphi$ is true in world *w* of M if for all *v* with *(w, v)* E $[\![\pi]\!]^{\mathbf{M}}$ it holds that φ is true in *v.* This is completely axiomatised by the usual **PDL** rules and axioms (Segerberg, 1982; Kozen and Parikh, 1981):

Modus ponens	and axioms for propositional logic
Modal generalisation	From $\vdash \varphi$ infer $\vdash [\pi]\varphi$

Normality	$\vdash [\pi](\varphi \rightarrow \psi) \rightarrow ([\pi]\varphi \rightarrow [\pi]\psi)$
Test	$\vdash [?\varphi]\psi \leftrightarrow (\varphi \rightarrow \psi)$
Sequence	$\vdash [\pi_1;\pi_2]\varphi \leftrightarrow [\pi_1][\pi_2]\varphi$
Choice	$\vdash [\pi_1 \cup \pi_2]\varphi \leftrightarrow ([\pi_1]\varphi \wedge [\pi_2]\varphi)$
Mix	$\vdash [\pi^*]\varphi \leftrightarrow (\varphi \wedge [\pi][\pi^*]\varphi)$
Induction	$\vdash (\varphi \wedge [\pi^*](\varphi \rightarrow [\pi]\varphi)) \rightarrow [\pi^*]\varphi$

In (van Benthem et al., 2üü6b) it is proved that extending the **PDL** language with an extra modality *[A, e]*φ does not change the expressive power of the language. Interpretation of the new modality: *[A, e]*φ is true in *w* in M if success of the update of M with action model *A* to M $\otimes$ *A* implies that φ is true in *(w,* e) in M $\otimes$ *A.* To see what *that* means one has to grasp the definition of update modeIs *A* and the update product operation $\otimes$, which we will now give for the epistemic preference case.

An action model (for agent set *I)* is like an epistemic preference model for *I,* with the difference that the worlds are now called events, and that the valuation has been replaced by a precondition map **pre** that assigns to each event e a formula of the language called the precondition of e. From now on we call the epistemic preference models static modeis.

Updating a static model M = *(W, V, R, P)* with an action model *A* = *(E,* pre, R, P) succeeds if the set

$$\{(w,e)\ Iw \in W, e \in E, M, w \models \mathrm{pre}(e)\}$$

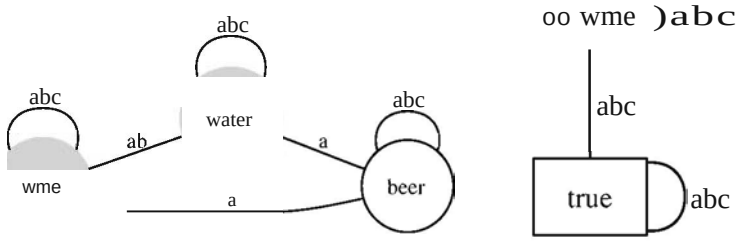


FIGURE 1. Statie model and update.

is non-empty. The result of this update is a new statie model M ⊗ A = (W^{I}, VI, RI, Pi) with

- W^{I} = {(w,e) Iw ∈ W,e ∈ E,M,w ⊨ pre(e)},
- VI(W, e) = V(w),
- R_i' is given by {(w,e),(v,J)) I (w,v) ∈ Ri, (e,J) ∈ $\mathbf{R}_i$},
- PI is given by {(w, e), (v, J)) I (w, v) EPi, (e, J) ∈ $\mathbf{P}_i$}.

If the statie model has a set of distinguished states *Wo* and the action model a set of distinguished events *Eo*, then the distinguished worlds of M ⊗ *A* are the (W, e) with w ∈ *Wo* and e ∈ *Eo*.

Figure 1 gives an example pair of a statie model with an update action. The distinguished worlds of the model and the distinguished event of the action model are shaded grey. Only the *R* relations are drawn, for three agents *a*, *b*, c. The result of the update is shown in Figure 2, on the left. This result can be reduced to the bisimilar model on the right in the same figure, with the bisimulation linking the distinguished worlds. The result of the update is that the distinguished "wine" world has disappeared, without any of *a*, *b*, c being aware of the change.

In LCC, action update is extended with factual change, which is handled by propositional substitution. Here we will consider another extension, with preferenee change, handled by preferenee substitution (first proposed by van Benthem and Liu, 2004). A preferenee substitution is a map from agents to PDL program expressions *tt* represented by a finite set of bindings

$$\{i_1 \mapsto \pi_1, \ldots, i_n \mapsto \pi_n\}$$

where the *ij* are agents, all different, and where the π_i are program expressions from our PDL language. It is assumed that each i that does

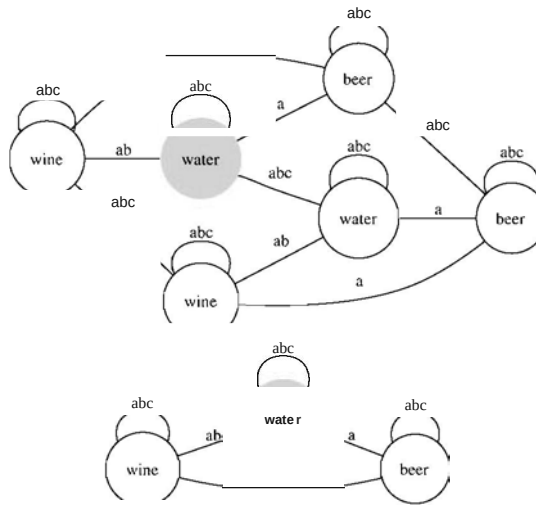


FIGURE 2. Update result, before and aftel' reduction under bisimulation.

not occur in the lefthand side of a binding is mapped to $\geq_i$. Call the set $\{\mathrm{i} \in I \mid p(\mathrm{i}) \neq \geq_i\}$ the *domain* of *p*. If M = *(W, V, R, P)* is a preference model and *p* is a preference substitution, then *MP* is the result of changing the preference map *P* of M to *P"* given by:

$$\begin{array}{ll} PP(i) & Pi \text{ for i not in the domain of } p, \\ PP(i) & [\![\rho(i)]\!]^{\mathrm{M}} \text{ for i = } ij \text{ in the domain of } p. \end{array}$$

Now extend the **PDL** language with a modality $[\rho]\varphi$ for preference change, with the following interpretation:

$$M,w \models [\rho]\varphi \quad :\Longleftrightarrow \quad MP,w \models \varphi.$$

Then we get a complete logic for preference change:

Theorem 3.1. The logic of epistemic preference **PDL** with preference change modalities is complete.

Proo]. The preference change effects of *lp]* can be captured by a set of reduetion axioms for *lp]* that commute with all sententiallanguage constructs, and that handle formulas of the form *[pi* $[\pi]\varphi$ by means of reduction axioms of the form

$$[\rho][\pi]\varphi \quad \leftrightarrow \quad [F_\rho(\pi)][\rho]\varphi,$$

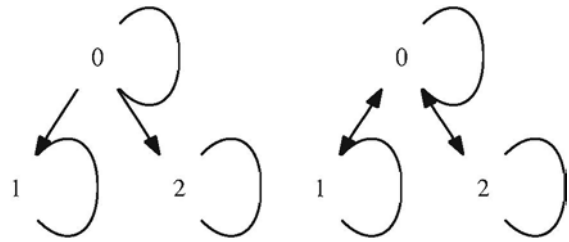


FIGURE 3. Preorder with a non-transitive symmetrie closure.

with F_p given by:

$F_p(\sim_i)$	$\sim_i$
$F_\rho(\geq_i)$	$\begin{cases} p(\mathrm{i}) & \text{if i in the domain of } p, \\ \geq_i & \text{otherwise,} \end{cases}$
$Fp(?cp)$	$?[p]cp,$
$Fp(\pi_1;\pi_2)$	$F_\rho(\pi_1);F_\rho(\pi_2),$
$F_\rho(\pi_1 \cup \pi_2)$	$F_\rho(\pi_1) \cup F_\rho(\pi_2),$
$Fp(\pi^*)$	$(F_\rho(\pi))^*.$

It is easily checked that these reduction axioms are sound, and that for each formula of the extended language the axioms yield an equivalent formula in which $[P]$ occurs with lower complexity, which means that the reduction axioms can be used to translate formulas of the extended language to **PDL** formulas. Completeness then fol!ows from the completeness of PDL. Q.E.D.

4 Vet another logic...

In this section we look at a more constrained case, by replacing the epistemic preferenee models by 'belief revision modeis' in the style of Grove (1988), Board (2002), and Baltag and Smets (2006, 2008b) (who cal! them 'multi-agent plausibility frames'). There is almost complete agreement that preferenee relations should be transitive and refiexive (preorders). But transitivity plus refiexivity of a binary relation R do not together imply that $Ru\ R^\smile$ is an equivalence. Figure 3 gives a counterexample. The two extra conditions of weak connectedness for Rand for $R^\smile$ remedy this. A binary relation R is weakly connected (terminology of Goldblatt, 1987) if the following holds:

$$\forall x, y, z((xRy \wedge xRz) \rightarrow (yRz \vee y = z \vee zRy)).$$

Theorem 4.1. Assume R is a preorder. Then $Ru\ R^\smile$ is an equivalence iff both $\mathrm{Rand}\ R^\smile$ are weakly connected.

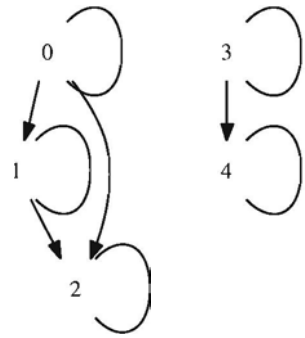


FIGURE 4. Locally connected preorder that is not connected.

Proof. $\Rightarrow$: immediate.

$\Leftarrow$: Let R be a preorder such that both R and $R^{\smile}$ are weakly connected. We have to show that $R \cup R^{\smile}$ is an equivalence. Symmetry and refiexivity are immediate. For the check of transitivity, assume $xR \cup R^{\smile}y$ and $yR \cup R^{\smile}z$. There are four cases. (i) $xRyRz$. Then xRz by transitivity of R, hence $xR \cup tr\ll$ (ii) $xRyR^{\smile}z$. Then $yR^{\smile}x$ and $yR^{\smile}z$, and by weak connectedness of $R^{\smile}$, either $(xR^{\smile}z$ or $zR^{\smile}x)$, hence $xR \cup R{:}z;$ or $x = z$, hence xRz by refiexivity of R. Therefore $xR \cup R^{\smile}z$ in all cases. (iii) $xR^{\smile}yRz$. Similar. (iv) $xR^{\smile}yR^{\smile}z$. Then $zRyRx$, and zRx by transitivity of R. Therefore $xR \cup R^{\smile}z$. Q.E.D.

Call a preorder that is weakly connected and conversely weakly connected locally connected. The example in Figure 4 shows that locally connected preorders need not be connected. Taking the symmetrie closure of this example generates an equivalence with two equivalence classes. More generally, taking the symmetrie closure of a locally connected preorder creates an equivalence that can play the role of a knowledge relation defined from the preferenee order. To interpret the preferenee order as conditional belief, it is convenient to assume that it is also well-founded: this makes for a smooth definition of the notion of a 'best possible world'.

A belief revision model M (again, for a set of agents 1) is a tuple (W, V, P) where W is a non-empty set of worlds, V is a propositional valuation and P is a function that maps each agent i to a preferenee relation $\leq_i$ that is a locally connected well-preorder. That is, $\leq_i$ is a preorder (refiexive and transitive) that is well-founded (in terms of <i for the strict part of $\leq_i$, this is the requirement that there is no infinite sequence of $w_1, w_2, \ldots$ with $\ldots <_i w_2 <_i w_1$), and such that both $\leq_i$ and its converse are weakly connected.

In what follows we will use <i with the meaning explained above, $\geq_i$ for the converse of $\leq_i$, '>: for the converse of <i, and $\sim_i$ for $\leq_i \cup \geq_i$.

The locally connected well-preorders $\leq_i$ can be used to induce accessibility relations $\rightarrow_i^P$ for each subset P of the domain, by means of the following standard definition:

$$\rightarrow_i^P := \{(x, y) \mid x \sim_i y \wedge y \in \mathrm{MIN}_{\leq_i} P\},$$

where $\mathrm{MIN}_{\leq_i} P$, the set of minimal elements of P under $\leq_i$, is defined as

$$\{s \in P : \forall s' \in P(s' \leq s \Rightarrow s \leq s')\}.$$

This picks out the minimal worlds linked to the current world, according to $\leq_i$, within the set of worlds satisfying $[\![\varphi]\!]^{\mathbf{M}}$. The requirement of wellfoundedness ensures that $\mathrm{MIN}_{\leq} P$ will be non-empty for non-empty P. Investigating these $\rightarrow^P$ relations, we see that they have plausible properties for belief:

Proposition 4.2. Let $\leq$ be a locally connected well-preorder on S and let P be a non-empty subset of S. Then $\rightarrow^P$ is transitive, euclidean and serial.

Proo]. Transitivity: if $x \rightarrow^P y$ then y r-» x and y E $\mathrm{MIN}_{\leq} P$. **If** $y \rightarrow^P z$ then z r-» y and z E MIN <Po **It** follows by local connectedness of $\leq$ that z r-» x and by the definition of $\rightarrow^P$ that $x \rightarrow^P z$.

Euclideanness: let $x \rightarrow^P y$ and $x \rightarrow^P z$. We have to show $y \rightarrow^P z$. From $x \rightarrow^P y$, Y r-» x and y E MIN≤P. From $x \rightarrow^P z$, z r-» x and z E MIN<P. From local connectedness, $y \sim z$. Hence $y \rightarrow^P z$. -

Seriality: Let x E P. Since $\leq$ is a preorder there are y E P with $y \leq x$. The wellfoundedness of $\leq$ guarantees that there are $\leq$ minimal such y. Q.E.D.

Transitivity, euclideanness and seriality are the frame properties corresponding to positively and negatively introspective consistent belief (KD45 belief, Chellas, 1984).

Figure 5 gives an example with both the $\leq$ relation (shown as solid arrows in the direction of more preferred worlds, i.e., with an arrow from x to y for $x \geq y$) and the induced $\rightarrow$ relation on the whole domain (shown as dotted arrows). The above gives us in fact knowledge relations $\sim_i$ together with for each knowledge cell a Lewis-style (1973) counterfactual relation: a connected well-preorder, which can be viewed as a set of nested spheres, with the minimal elements as the innermost sphere. Compare also the conditional models of Burgess (1981) and Veltman (1985) (linked to Dynamic Doxastic Logic in Girard, 2007).

Baltag and Smets (2008b,a) present logics of individual multi-agent belief and knowledge for belief revision modeis, and define belief update for

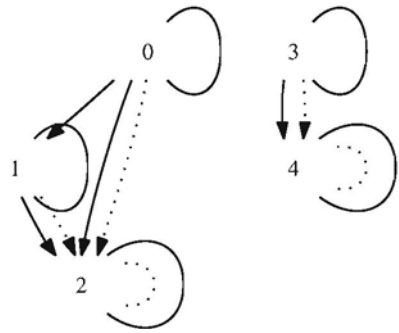


FIGURE 5. Preference (solid arrows) and belief (dotted arrows).

this as a particular kind of action update in the style of (Baltag et al., 1998), called action priority update. Here we sketch the extension to a system that also handles common knowledge and common conditional belief, and where the action update has belief change incorporated in it by means of relational substitution.

The set-up will be less general than in the logic LCe: in LCC no assumptions are made about the update actions, and so the accessibility relations could easily deteriorate, e.g., as a result of updating with a lie. Since in the present set-up we make assumptions about the accessibilities (to wit, that they are locally connected well-preorders), we have to ensure that our update actioris preserve thpsp relational properties.

Consider the following slight modification of the **PDL** language (again assume *p* ranges over a set of basic propositions *Prop* and i over a set of agents 1):

$$\varphi \qquad \mathbf{Tip} \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid [\pi]\varphi$$
$$tt \qquad \sim_i \mid \leq_i \mid \geq_i \mid \rightarrow_i^{\varphi} \mid \leftarrow_i^{\varphi} \mid \mathsf{IC} \mid ?\varphi \mid \pi_1;\pi_2 \mid \pi_1 \cup \pi_2 \mid \text{zr"}$$

Call this language Lpref. This time, we treat $\sim_i$ as a derived notion, by putting in an axiom that defines $\sim_i$ as $\leq_i \mathrm{U} \geq_i$. The intention is to let $\sim_i$ be interpreted as the knowledge relation for agent i, $\leq_i$ as the preference relation for i, $\geq_i$ as the converse preference relation for i, $\rightarrow_i^{\varphi}$ as the conditional belief relation defined from $\leq_i$ as explained above, $\leftarrow_i^{\varphi}$ as its converse, and Cas global accessibility. We use $\rightarrow_i$ as shorthand for $\rightarrow_i^{\top}$.

We have added agiobal modality C, and we will set up things in such way that $[G]\varphi$ expresses that everywhere in the model φ holds, and that $\langle G\rangle\varphi$ expresses that φ holds somewhere. It is well-known that adding agiobal modality and converses to **PDL** does not change its properties: the logic re-

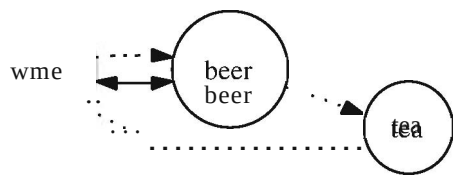
mains decidabie, and satisfiability remains EXPTIME-complete (Blackburn et al., 2001).

The semantics of Lpref is given relative to belief revision models as indicated above. Formula meaning $[\![\varphi]\!]^{\mathbf{M}}$ and relational meaning $[\![\pi]\!]^{\mathbf{M}}$ are handled in the usual way. The interpretation of the knowledge relation of agent i is given by $[\![\sim_i]\!]^{\mathbf{M}} := \leq_i^{\mathbf{M}} \cup \geq_i^{\mathbf{M}}$, that for the preference relation of agent i by $[\![\leq_i]\!]^{\mathbf{M}} := \leq_i^{\mathbf{M}}$, that for the converse preference relation of agent i by its converse, that for the conditional belief of agent i by $[\![\rightarrow_i^{\varphi}]\!]^{\mathbf{M}} := \rightarrow_i^{[\![\varphi]\!]^{\mathbf{M}}}$, that for $\leftarrow_i^{\varphi}$ by its converse. The global modality is interpreted as the universal relation, and test, sequential composition, choice and Kleene star are interpreted as usual.

The interplay between the modalities $[\sim_i]$ (knowiedge) and $[\geq_i]$ (safe belief) is analysed by Baltag and Smets (2008b), where they remark that the converse preference modality $[\geq_i]$ in belief revision models behaves like an S4.3 modality (reftexive, transitive and not forward branching), and lives happily together with the S5 modality for $[\sim_i]$.

To see how this all works out, let us have a look at the truth conditions for $[\rightarrow_i^{\varphi}]\psi$. This is true in a world *w* in model M if in all worlds *v* with *v* $\sim_i$ wand *v* minimal in $[\![\varphi]\!]^{\mathbf{M}}$ under $\leq_i$ it holds that ψ is true. This is indeed conditional belief, relative to φ. Compare this with $[\geq_i]\psi$. This is true in a world *w* if in all worlds that are at least as preferred, ψ is true. Finally, $[\sim_i]\psi$ is true in *w* if ψ is true in all worlds, preferred or not, that i can access from *w*.

As a further example, consider a situation where Alexandru is drinking wine, while Jan does not know whether he is drinking wine or beer, and Sonja thinks that he is drinking tea. The actual situation is shaded grey, Jan's preference relation has solid lines, that of Sonja dotted lines. Reftexive arrows are not drawn, so Alexandru's preferences are not visible in the picture.



In the actual world it is true that Jan knows that Alexandru knows what Alexandru is drinking: $[\sim_j]([\sim_a]w \vee [\sim_a]b)$, and that Sonja believes Alexandru is drinking tea and that Alexandru knows it: $[s][\sim_a]t$. Under condition ***,t,*** however, Sonja has the belief in the actual world that Alexandru is drinking beer: $[\rightarrow_s^{\neg t}]b$. Moreover, Jan and Sonja have a common belief under condition ***,t*** that Alexandru is drinking wine or beer: $[\rightarrow_j^{\neg t} \cup \rightarrow_s^{\neg t}; (\rightarrow_j^{\neg t} \cup \rightarrow_s^{\neg t})^*](w \vee b)$. As a final illustration, note that [+--8J-/- is true

in a world if this is *noi* among Sonja's most preferred worlds. Notice that if Sonja conditionalizes her belief to these worlds, she would believe that Alexandru is drinking beer: $[\rightarrow_s^{[s^\smile]\perp}]b$ is true in the actual world.

It should be clear from the example that this language is very expressive. To get at a complete logic for it, we need axioms and rules for propositional logic, S5 axioms for the global modality (Goranko and Passy, 1992), axioms for forward connectedness of $\geq$ and of $\leq$ (see Goldblatt, 1987), axioms for converse, relating $\leq$ to $\geq$ and $\rightarrow$ to $\leftarrow$, as in temporal logic (Prior, 1967), and the general axioms and rules for PDL (Segerberg, 1982). Finally, add the following definition of conditional belief in terms of knowledge and safe belief that can already be found in (Boutilier, 1994) as an axiom:

$$\langle\sim_i\rangle\varphi \rightarrow \langle\sim_i\rangle(\varphi \wedge [\geq_i](\varphi \rightarrow \psi)).$$

This definition (also used in Baltag and Smets, 2008b) states that conditional to φ, i believes in ψ if either there are no accessible φ worlds, or there is an accessible φ world in which the belief in $\varphi \rightarrow \psi$ is safe. The full calculus for Lpref is given in Figure 6.

Theorem 4.3. The axiom system for Lpref is complete for belief revision modeis; Lpref has the finite model property and is decidabie.

Proo]. Modify the canonieal model construction for modallogic for the case of PDL, by means of Fischer-Ladner closures (Fischer and Ladner, 1979) (also see Blackburn et al., 2001). This gives a finite canonieal model with the properties for $\leq_i$ and $\geq_i$ corresponding to the axioms (since the axioms for $\leq_i$ and $\geq_i$ are canonieal). In particular, each $\geq_i$ relation will be reftexive, transitive and weakly connected, each relation $\leq_i$ will be weakly connected, and the $\leq_i$ and $\geq_i$ relations will be converses of each other. Together this gives (Theorem 4.1) that the $\leq_i \cup \geq_i$ are equivalences. Since the canonical model has a finite set of nodes, each $\leq_i$ relation is also well-founded. Thus, the canonical model is in fact a belief revision model. AIso, the $\rightarrow_i$ and $\leftarrow_i$ relations are converses of each other, and related to the $\geq_i$ relations in the correct way. The canonical model construction gives us for each consistent formula φ a belief revision model satisfying φ with a finite set of nodes. Only finitely many of the relations in that model are relevant to the satisfiability of φ, so this gives a finite model (see Blackburn et al., 2001, for further details). Since the logic has the finite model property it is decidabie. Q.E.D.

Since the axiomatisation is complete, the S5 properties of $\sim_i$ are derivable, as weil as the principle that knowledge implies safe belief:

$$[\sim_i]\varphi \rightarrow [\geq_i]\varphi.$$

Modus ponens	and axioms for propositional logic
Modal generalisation	From $\vdash \varphi$ infer $\vdash [\pi]\varphi$
Normality	$\vdash [\pi](\varphi \to \psi) \to ([\pi]\varphi \to [\pi]\psi)$
Inclusion of everything in G	$\vdash [G]\varphi \to [\pi]\varphi$
Refiexivity of G	$\vdash [G]\varphi \to \varphi$
Transitivity of G	$\vdash [G]\varphi \to$ *[C][C]ep*
Symmetry of G	$\vdash \varphi \to$ *[C]* $\langle G\rangle\varphi$
Knowledge definition	$\vdash [\sim_i]\varphi \leftrightarrow [\leq_i \cup \geq_i]\varphi$
Truthfulness of safe belief	$\vdash [\geq_i]\varphi \to \varphi$
Transitivity of safe belief	$\vdash [\geq_i]\varphi \to [\geq_i][\geq_i]\varphi$
$\geq$ included in $\leq\check{}$	$\vdash \varphi \to [\geq_i]\langle\leq_i\rangle\varphi$
$\leq$ included in $\geq\check{}$	$\vdash \varphi \to [\leq_i]\langle\geq_i\rangle\varphi$
Weak connectedness of $<$	$\vdash [\leq_i]((\varphi \wedge [\leq_i]\varphi) \to \psi) \vee [\leq_i]((\psi \wedge [\leq_i]\psi) \to \varphi)$
Weak connectedness of $\geq$	$\vdash [\geq_i]((\varphi \wedge [\geq_i]\varphi) \to \psi) \vee [\geq_i]((\psi \wedge [\geq_i]\psi) \to \varphi)$
Conditional belief definition	$\vdash [\to_i^{\varphi}]\psi \leftrightarrow (\langle\sim_i\rangle\varphi \to \langle\sim_i\rangle(\varphi \wedge [\geq_i](\varphi \to \psi)))$
$\to$ included in $\leftarrow\check{}$	$\vdash \varphi \to [\to_i^{\psi}]\langle\leftarrow_i^{\psi}\rangle\varphi$
$\leftarrow$ included in $\to\check{}$	$\vdash \varphi \to [\leftarrow_i^{\psi}]\langle\to_i^{\psi}\rangle\varphi$
Test	$\vdash [?\varphi]\psi \leftrightarrow (\varphi \to \psi)$
Sequence	$\vdash [\pi_1;\pi_2]\varphi \leftrightarrow [\pi_1][\pi_2]\varphi$
Choice	$\vdash [\pi_1 \cup \pi_2]\varphi \leftrightarrow ([\pi_1]\varphi \wedge [\pi_2]\varphi)$
Mix	$\vdash [\pi^*]\varphi \leftrightarrow (\varphi \wedge [\pi][\pi^*]\varphi)$
Induction	$\vdash (\varphi \wedge [\pi^*](\varphi \to [\pi]\varphi)) \to [\pi^*]\varphi$

FIGURE 6. Axiom system for Lpref.

The same holds for the following principles for conditional belief given in (Board, 2002):

Safe belief implies belief	$\vdash [\geq_i]\varphi \to [\to_i^{\psi}]\varphi$
Positive introspection	$\vdash [\to_i^{\psi}]\varphi \to [\sim_i][\to\psi_i]\varphi$
Negative introspection	$\vdash \neg[\to_i^{\psi}]\varphi \to [\sim_i]\neg[\to_i^{\psi}]\varphi$
Successful revision	$\vdash [\to_i^{\varphi}]\varphi$
Minimality of revision	$\vdash \langle\to_i^{\varphi}\rangle\psi \to ([i^{\varphi\wedge\psi}]\chi \leftrightarrow [\to_i^{\varphi}](\psi \to X))$

We end with an open question: is $\to_i^{\varphi}$ definable from $\geq_i$ and $\leq_i$ using only test, sequence, choice and Kleene star?

5 Combining update and upgrade

The way we composed knowledge and belief by means of regular operations may have adynamie fiavour, but appearance is deceptive. The resulting doxastic and epistemic 'programs' still describe what goes on in a statie model. Real communicative action is changing old belief revision models into new ones. These actions should represent new hard information that cannot be undone, but also soft information like belief changes that can be reversed again later on. For this we can use update action by means of action modeis, with soft information update handled by means of action priority update (Baltag and Smets, 2008b,a), or preferenee substitution as in (van Benthem and Liu, 2004). Here we will propose a combination of these two.

Action models for belief revision are like belief revision modeis, but with the valnation replaced by a precondition map. Wp add two extra ingredients. First, we add to each event a propositional substitution, to be used, as in LCC, for making factual changes to statie modeis. Propositional substitutions are maps represented as sets of bindings

$$\{\mathrm{Pl} \mapsto \varphi_1, \ldots, Pn \mapsto \varphi_n\}$$

where all the *Pi* are different. It is assumed that each *P* that does not occur in the lefthand side of a binding is mapped to *p*. The domain of a propositional substitution σ is the set $\{p \in Prop \mid \sigma(p) \neq p\}$. If M = (W, V, F) is a belief revision model and σ is an [Pref propositional substitution, then V_{M}^{σ} is the valuation given by $\lambda w \lambda p \cdot w \in [\![p^{\sigma}]\!]^{\mathrm{M}}$. In other words, V_{M}^{σ} assigns to *w* the set of basic propositions *p* such that p^{σ} is true in world *w* in model M. M^{σ} is the model with its valuation changed by σ as indicated. Next, we add relational substitutions, as defined in Section 3, one to each event. Thus, an action model for belief revision is a tuple A = (E,pre,P,psub,rsub) with *E* a non-empty finite set of events, **psub** and **rsub** maps from *E* to propositional substitutions and relational substitutions, respectively, and with **rsub** subject to the following constraint:

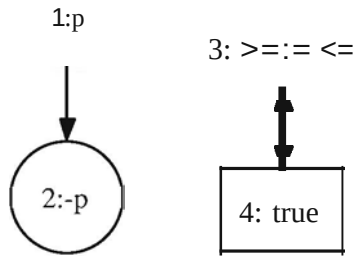


FIGURE 7. Unconstrained relational substitution creates havoc.

If $e \sim_i f$ in the action model, then rsub(e) and rsub(f) have the same binding for i.

This ensures a coherent definition of the effect of relational substitution on a belief structure. The example in Figure 7 il!ustrates this. But note that the substitutions are subject to further constraints. In the action model in the example, a single agent has a preference for -'*p* over *p*. In the update model, a substitution reverses the agent's preferences, but the agent cannot distinguish this from an action where nothing happens. What should the result of the update look like7 E.g., is there a preference arrow from (2,3) to (1,4)7 This is impossible to answer, as action 3 asks us to reverse the preference and action 4 demands that we keep the initial preference. The constraint on substitutions rules out such dilemmas.

The relational substitution p = rsub(e) at event e in action model A is meant to to be interpreted 'Iocal!y' at each world w in input model M. **If** P is the preference map of M, then let P_w^ρ be given by:

$$P_w^\rho(i) \qquad Pi \cap |w|^2_{[\![\sim_i]\!]^{\mathbf{M}}} \text{ for i not in the domain of } p,$$

$$P_w^\rho(i) \qquad [\![\rho(i)]\!]^{\mathbf{M}} \cap |w|^2_{[\![\sim_i]\!]^{\mathbf{M}}} \text{ for i = } ij \text{ in the domain of } p.$$

Thus, P_w^ρ is the result of making a change only to the local knowledge cel! at world w of agent i (which is given by the equivalence class $|w|_{[\![\sim_i]\!]^{\mathbf{M}}}$). Let

$$\bigcup_{w \in W} P_w^\rho(i)$$

Then P''(i) gives the result of the substitution *pon P(*i), for each knowledge cel! $|w|_{[\![\sim_i]\!]^{\mathbf{M}}}$ for i, and P'' gives the result of the substitution *pon P*, for each agent i.

Now the result of updating belief revision model M = *(W, V, P)* with action model A = *(E,* pre, P, psub, rsub) is given by M $\otimes$ A = *(W^{I}, VI, Pi),* where

- W^{I} = *{(w,e) Iw* ∈ *W,e* ∈ *E,M,w* $\models$ pre(e)},
- *VI(W,* e) = $V^{\sigma}(w)$,
- *PI(i)* is given by the anti-lexicographical order defined from *PP(i)* and P(i) (see Baltag and Smets, 2008b,a).

With these definitions in place, what are reasonable substitutions? A possible general form for a preference change could be a binding like this:

$$\geq_i \ \mapsto \ [\varphi_1, \varphi_2, \ldots, \varphi_n].$$

This is to be interpreted as an instruction to replace the belief preferences of i in the local knowledge cells by the new preference relation that prefers the φ_1 states above everything else, the $\neg\varphi_1 \wedge \varphi_2$ above the $\neg\varphi_1 \wedge \neg\varphi_2$ states, and so on, and the $\neg\varphi_1 \wedge \neg\varphi_2 \wedge \ldots \wedge \neg\varphi_{n-1} \wedge \varphi_n$ states above the $\neg\varphi_1 \wedge \neg\varphi_2 \wedge \ldots \wedge \neg\varphi_n$ states. Such relations are indeed connected well-preorders.

most preferred .... least preferred

Note that we can take $[\varphi_1, \varphi_2, \ldots, \varphi_n]$ as an abbreviation for the following doxastic program:

$$\begin{array}{lll} (\sim_i; ?\varphi_1) & \mathrm{U} & (?\neg\varphi_1; \sim_i; ?\neg\varphi_1; ?\varphi_2) \\ & \mathrm{U} & (?\neg\varphi_1; ?\neg\varphi_2; \sim_i; ?\neg\varphi_1; ?\neg\varphi_2; ?\varphi_3) \\ & \mathrm{U} & \\ & \cup & (?\neg\varphi_1; \cdots; ?\neg\varphi_n; \sim_i; ?\neg\varphi_1; \cdots; ?\neg\varphi_{n-1}; ?\varphi_n) \end{array}$$

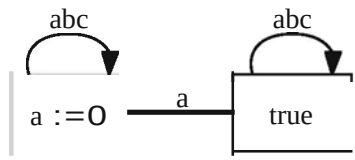
In general we have to be careful (as is also observed in (van Benthem and Liu, 2004)). If we have a connected well-preorder then adding arrows to it in the same knowledge cell may spoil its properties. Also, the union of two connected well-preorders need not be connected. So here is a question: what is the maximal sublanguage of doxastic programs that still guarantees that the defined relations are suitable preference relations? Or should belief revision models be further constrained to guarantee felicitous preference change? And if so: how?

6 Examples

Global amnesia: the event of agent *a* (Jason Bourne) forgetting all his beliefs, with everyone (including himself) being aware of this, is represented by the following action model (for the case of three agents *a, b, c):*

a:=G)abc

Alzheimer: the event of agent *a* forgetting everything, with the ot hers being aware of this, while *a* wrongly believes that nothing has happened. It is tempting to model this with the following update model:



Note however that this does not satisfy the constraint on relation update (the two actions are connected, but the substitution for *a* is not the same), so it may result in incoherent modeis.

Lacular amnesia (specific forgetting): forgetting everything about *p.* One way to model this is by means of an action model with a single action, accessible to all, with the relational substitution

$$\geq_i \;\mapsto\; (\geq_i \cup \sim_i, ?\neg p)^*$$

This will effectively add best-world arrows from everywhere in the knowledge cell to all $\neg p$ worlds.

Confession of faith in *p,* or publicly accepting *p:* an action model with a single action, accessible to all, with the relational substitution

$$\geq_i \;\mapsto\; (\geq_i \cup (?\neg p; \sim_i; ?p))^*.$$

This will make the *p* worlds better than the $\neg p$ worlds everywhere.

Submission to a guru: the act of adopting the belief of someone else, visible to all. A problem here is that the guru may know more than I do, so that the guru's preferences within my knowledge cell may not be connected. This means that the substitution $\leq_i \mapsto \leq_j$—the binding that expresses that i takes over *j's* beliefs-may involve growth or loss of knowledge for i. Consider the example of the wine-drinking Alexandru again: if Jan were to take over Sonja's beliefs, he would lose the information that Alexandru is drinking an alcoholic beverage.

Conformism: adopting the common beliefs of a certain group, visible to all: an action model with a single action accessible to all, with the following substitution for conformist agent i:

$$\geq_i \ \mapsto \ (\geq_i \cup \geq_j);(\geq_i \cup \geq_j)^*.$$

Belief coarsening: the most preferred worlds remain the most preferred, the next preferred remain the next preferred, and all further distinctions are erased. An action model with a single action accessible to all, and the following substitution for agent i:

$$\geq_i \ \mapsto \ \ \rightarrow_i \cup ?\top.$$

The union with the relation ?T has the effect of adding all refiexive arrows, to ensure that the result is refiexive again.

Van Benthem's $\Uparrow\varphi$ is handled by a substitution consisting of bindings like this:

$$\geq_i \ \mapsto \ \ (?\varphi;\sim_i?\neg\varphi) \ \cup \ (?\varphi;\geq_i?\varphi) \ \cup \ (?\neg\varphi;\geq_i?\neg\varphi).$$

This is an alternative for an update with an action model that has $\neg\varphi <B \varphi$. The example shows that conservative upgrade is handled equally weil by action priority updating and by belief change via substitution. But belief change by substitution seems more appropriate for 'elite change'. For this we need a test for being in the best φ world that i can conceive, by means of the Panglossian formula $\langle\leftarrow_i^{\varphi}\rangle\top$. The negation of this allows us to define eliLe change like this:

$$\geq_i \ \mapsto \rightarrow_i^{\varphi} \cup (\geq_i;?[\leftarrow_i^{\varphi}]\bot).$$

This promotes the best φ worlds past all other worlds, while leaving the rest of the ordering unchanged. Admittedly, such an operation could also be performed using action priority updating, but it would be much more cumbersome.

7 Further connections

To conneet up to the work of Bonanno again, what about time? Note that perceiving the ticking of a clock can be viewed as information update. A clock tick constitutes a change in the world, and agents can be aware or unaware of the change. This can be modelled within the framework introduced above. Let $t_i, \ldots, t_n$ be the clock bits for counting ticks in binary, and let C := C + 1 be shorthand for the propositional substitution that is needed to increment the binary number $t_1, \ldots, t_n$ by 1. Then public awareness of the clock tick is modelled by:

$$C := C\text{-}t \;)abc$$

Thus, perception of the ticking of a clock can be modelled as 'being in tune with change in the world'. Still, this is not quite the same as the 'next instance' operator O, for the **DEL** framework is specific about what happens during the clock tick, while O existentially quantifies over the change that takes place, rather in the spirit of (Baibiani et al., 2007).

In belief revision there is the AGM tradition, and its rational reconstruction in dynamic doxastic logic à la Segerberg. Now there also is a modal version in Bonanno style using temporallogic. **It** is shown in (van Benthem and Pacuit, 2006) that temporal logic has greater expressive power than DEL, which could be put to use in a temporal logic of belief revision (although Bonanno's present version does not seem to harness this power). As an independent development there is dynamic epistemic logic in the AmsterdamjOxford tradition, which was inspired by the logic of public announcement, and by the epistemic turn in game theory, à la Aumann. Next to this, and not quite integrated with it, there is an abundance of dynamic logics for belief change based on preference relations (Spohn, Shoham, Lewis), and again the Amsterdam and Oxford traditions. I hope this contribution has made clear that an elegant fusion of dynamic epistemic logic and dynamic logics for belief change is possible, and that this fusion allows to analyze AGM style belief revision in a multi-agent setting, and integrated within a powerful logic of communication and change.

Acknowledgements

Thanks to Alexandru Baltag, Johan van Benthem and Fiool' Sietsma for helpful comments on several earlier versions of the paper, and for constructive discussions. Hans van Ditmarsch sent me his last minute remarks. I am grateful to the editors of this volume for their display of the right mix of patience, tact, flexibility and firmness (with Robert van Rooij exhibiting the first three of these qualities and Krzysztof Apt the fourth). The support of the NWO Cognition Programme to the 'Games, Action and Social Software' Advanced Studies project is gratefully acknowledged.

References

Alchourrón, C.E., Gärdenfors, P. & Makinson, D. (1985). On the Logic of Theory Change: Partial Meet Contraction and Revision Functions. *The Journal of Symbolic Logic, 50(2):510-530.*

Balbiani, P., Baltag, A., van Ditmarsch, H., Herzig, A., Hoshi, T. & Santos De Lima, T. (2007). What can we achieve by arbitrary announcements?

a dynamie take on Fitch's knowability. **In** Samet, D., ed., *Proceedinqs of the l1th Conference on Theoretical Aspects of Rationality and Knowledge (TARK-2007), Brussels, Belgium, June* 25-27, *2007.*

Baltag, A. & Moss, L.S. (2004). Logies for Epistemie Programs. *Synthese,* 139(2):165-224.

Baltag, A., Moss, L.S. & Soleeki, S. (1998). The Logie of Common Knowledge, Public Annouueements, and Private Suspieions. **In** Gilboa, 1., ed., *Proceedinqs of the 'lth Conference on Theoreiical Aspects of Rationality and Knowiedge, (TARK'98),* pp. 43-56.

Baltag, A. & Smets, S. (2006). Conditional Doxastie Modeis: A Qualitative Approach to Dynamie Belief Revision. *Electroruc Notes in Theoreiical Computer Science, 165:5-21.*

Baltag, A. & Smets, S. (2008a). The logie of eonditional doxastie aetions. This volume.

Baltag, A. & Smets, S. (2008b). A qualitative theory of dynamic interaetive belief revision. **In** Bonanno, *G.,* van der Hoek, W. & Wooldridge, M., eds., *Logic and the Foundations of Game and Decision Theoru (LOFT* 7), Vol. 3 of *Texts in Logic and Games,* pp. 11-58. Amsterdam University Press.

van Benthem, J. (2004). Dynamie Logie for Belief Revision. *Journalof Applied Non-Classical Logies, 14(2):129-155.*

van Benthem, .1. & Liu, F. (2004). Dynamic Logie of Preferenee Upgrade. *Journalof Applied Non-Classical Logies, 14:157-182.*

van Benthem, .1., van Otterloo, S. & Roy, O. (2006a). Preferenee logic, eonditionals, and solution concepts in games. **In** Lagerlund, H., Lindström, S. & Sliwinski, R., eds., *Modality Matters: Twenty-Five Essays in HonOUT of Krister Seqerberq,* Vol. 53 of *Uppsala Fhilosophical Studies,* pp. 61-76, Uppsala.

van Benthem, J. & Paeuit, E. (2006). The tree of knowledge in action: Towards a eommon perspeetive. **In** Governatori, *G.,* Hodkinson, I. & Venema, Y., eds., *Advances in Modal Logic, Vol.* 6, pp. 87-106. College Publieations.

van Benthem, .1., van Eijek, .1. & Kooi, B. (2006b). LogiesofCommunieation and Change. *Information and Computation, 204(99):1620-1666.*

Blaekburn, P., de Rijke, M. & Venema, Y. (2001). *Modal Logic,* Vol. 53 of *Cambridqe Tracis in Theoreiical Computer Science.* Cambridge University Press.

Board, O. (2002). Dynamic Interactive Epistemology. *Games and Economie Behaviour, 49(1):49-80.*

Bonanno, G. (2008). Belief revision in a temporal framework. This volume.

Boutilier, C. (1994). Toward a logic of qualitative decision theory. In Doyle, .1., Sandewall, E. & Torasso, P., eds., *Proceedinqs of the 4th International Conference on Principle of Knowledge Representation and Reasoning (KR-*94), pp. 75-86. Morgan Kaufmann.

Burgess, .J.P. (1981). Quick completeness proofs for some logics of conditionals. *Noire Dame Journalof Formol Logic, 22(1):76-84.*

Chellas, B. (1984). *Modal Logic: an Introduetion.* Cambridge University Press.

van Ditmarsch, H. (2005). Belief change and dynamic logic. In Delgrande, .1., Lang, .1., Rott, H. & Tallon, .1.-M., eds., *Belief Change in Rational Agents: Ferspectives [rom Artificial Intelligence, Fhilosophy, and Economics,* number 05321 in Dagstuhl Seminar Proceedings, Dagstuhl. IBFI, Schloss Dagstuhl.

van Ditmarsch, H., van der Hoek, W. & Kooi, B. (2007). *Dynamic Epistemic Logic,* Vol. 337 of *Synthese Library.* Springer.

Fischer, M.J. & Ladner, R.E. (1979). Propositional dynamic logic of regular programs. *Journalof Computer and System Sciences,* 18(2):194-21l.

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

Gärdenfors, P. & Rott, H. (1995). Belief revision. In Gabbay, D.M., Hogger, Ci.I. & Robinson, .1.A., eds., *Handbook of Logic in Artificial Intelligence and Logic Frogmmming, Vol. 4.* Oxford University Press, Oxford.

Gerbrandy, J. (1999). Dynamic Epistemic Logic. In Moss, L.S., Ginzburg, .1. & de Rijke, M., eds., *Logic, Language and Computation, Vol.* 2. CSLI Publications, Stanford University.

Girard, P. (2007). From onions to broccoli: Generalizing Lewis's counterfactuallogic. *Journalof Applied Non-Classical Logies,* 17(3):213-229. Special Issue on Belief Revision and Dynamic Logic.

Girard, P. (2008). *Modal Logic [or Preferenee and Belief Change.* PhD thesis, Universiteit van Amsterdam and Stanford University. *ILLC Fublication* DS-2008-04.

Goldblatt, R. (1992 (first edition 1987)). *Logies of Time and Computation, Seeond Edition, Revised and Expanded,* Vol. 7 of *CSLI Leciure Notes.* CSLI, Stanford. Distributed by University of Chicago Press.

Goranko, V. & Passy, S. (1992). Using the Universal Modality: Gains and Questions. *Journalof Logie and Computation, 2(1):5-30.*

Grove, A. (1988). Two Modellings for Theory Change. *Journalof Philosophieal Logie, 17(2):157-170.*

Halpern, .l.Y. (1997). Defining relative likelihood in partially-ordered preferential structures. J*ournal of Artificial Intelligenee Research.* 7:1-24.

Kozen, D. & Parikh, R. (1981). An elementary proof of the completeness of PDL. *Theoretical Computer Seienee, 14(1):113-118.*

Lewis, D. (1973). *Counterfactuals.* Blackwell Publishing, Oxford.

Liu, F. (2008). *Changing fOT the Better. Preferenee Dynamies and Agent Dioersitu.* **PhD** thesis, University of Amsterdam. *ILLC Publieations* DS-2008-02.

Plaza, .J.A. (1989). Logies of public communications. **In** Emrich, M.L., Pfeifer, M.S., Hadzikadic, M. & Ras, Z.W., eds., *Proceedinçs of the 4th International Symposium on Methodologies fOT Intelligent Systems,* pp. 201-216.

Prior, A. (1967). *Past, Present and Future.* Clarendon Press, Oxford.

Rott, H. (2006). Shifting Priorities: Simple Representations for Twenty-seven Iterated Theory Change Operators. **In** Lagerlund, H., Lindström, S. & Sliwinski, R., eds., *Modality Matters: Twenty-Five Essays in HonOUT of Krister Seqerberq,* Uppsala Philosophical Studies Volume 53, pp. 359-384. Uppsala Universitet.

Segerberg, K. (1982). A completeness theorem in the modal logic of programs. **In** Traczyck, T., ed., *Urnoersol Algebm and Applieations,* pp. 36-46. Polish Science Publications.

Segerberg, K. (1995). Belief Revision from the Point of View of Doxastic Logic. *Bulletin of the IGPL, 3(4):535-553.*

Segerberg, K. (1999). Two Traditions in the Logic of Belief: Bringing Them Together. **In** Ohlbach, H..l. & Reyle, U., eds., *Logie, Language and Reasoning: Essays in HonOUT of Dov Gabbay,* Vol. 5 of *Trends in Logie,* pp. 135-147. Klmver Academie Publishers.

Veltman, F. (1985). *Logies fOT Conditionals*. **PhD** thesis, University of Amsterdam.

Meaningful Talk

Yossi Feinberg

Graduate School of Business
Stanford University
518 Memorial Way
Stanford, CA 94305-5015, United States of America
yossi@gsb.stanford.edu

Abstract

Recent developments in the interface of Economics and Logic yield the promise of capturing phenomena in strategie interaction that was previously beyond theoretical economie modeling. We consider one such application in the case of strategie situations that players may perceive differently. We show that the content of messages sent in this setting carries strategie implications without resorting to prior conventions and beyond the observations made in the literature dealing with cheap talk communications. The content of the message becomes strategically meaningful since it reveals perceptions. Various forms of interaction between meaningful statements and strategie behavior are discussed.

1 Introduction

The modeling of information in strategie settings has been the centerpieee of economie theory for almost four decades now. Models of signaling, bargaining, auctions, contracts, mechanism design and more, are grounded in the premise that the decision makers posses private information relevant for outcomes, that their behavior is conditioned on their information, that they form beliefs about the information held by others and that they revise these beliefs based on observed actions. **In** addition, these fields of research add the possibility of communication between decision makers. However, formal theoretical economics modeling tends to overlook the *content*, or meaning of the messages that the decision makers may exchange, and, in particular, how this meaning may inftuence strategie behavior.

In his seminal work, Spenee (1973) introduced the job market signaling model. **In** this model an employer wishes to hire qualified employees, however only the potential candidates know their private degree of qualification. Spenee shows that if the employees are able to signal their abilities then an equilibrium of the game may be able to separate the high quality candidates from the lower quality ones. This signal is a potentially costly action

Krzysztof R. Apt, Robert van Rooij (eds.). *New Perspectives on Games and Interaction.* Texts in Logic and Games 4, Amsterdam University Press 2008, pp. 105-119.

taken by the candidates, for example obtaining a college degree. The important feature for separation by a signaling action is that it is more costly to a candidate that is less qualified for the job. The signaling action itself may have no impact on the employee's underlined qualities as they pertain to the job offered, it only serves as a way for the more qualified workers to separate themselves from the others. The information-the employee's qualification-is being deduced indirectly from his action in equilibrium. Such inferred information from an observed action is what economists usually refer to as informational content.[1] The content is the indirect deduction about private information derived from the information dependent strategie incentives associated with various actions.

This "(costly) action speaks louder than words" approach resonates weil with the economie perspective that a message that can be sent without incurring costs,[2] has to be stripped of all meaning other than its strategie intention when analyzed by the receiver. With that in mind, the ingenious work by Crawford and Sobel (1982) demonstrated that cheap talk, as it became to be known, can have profound strategie impact. By reacting to different messages differently, the receiver can create an incentive for the sender to send different messages depending on the sender's private information which also makes the receiver's choices best responses and all this without incurring any cost of signaling. The strategie reaction in such cheap talk games distinguishes *between* various messages, but it need not, and does not, depend in any way on the *content* of a message-there is no relevanee to *what* is being said only to saying one arbitrary thing rather than the other.

While cheap talk inlluences strategie outcornes, iL was recognized in Lhe refinement literature" that the content of the messages plays no role in the solution (see Cho and Kreps, 1987, for an illuminating explanation). One of the few direct efforts to confront the issue of meaningful messages in Economics is due to Farrell (1993). He suggests a delicate solution concept termed *Neoloqism-proof equilibrium* where an equilibrium defines the expected messages (that must agree with equilibrium behavior of the various sender types) and has to be robust to messages that are new-that will surprise the receiver given the expected equilibrium play. **In** this sense the language is endogenously dictated by some equilibrium behavior that forces the meaning of the message to coincide with that behavior. The work of Matthews et al. (1991) confronts the problem by introducing *armounce-*

1 See Spence's (2001) Nobel Prize lecture for an insightful description of this type of informational content.

2 In particular no indirect casts in the form of a contractual or reputational commitment.

3 The literature dealing with providing alternative solution concepts to Nash equilibrium that will confront the multiplicity of equilibria and remove seemingly "unwanted" behavior.

ment proof equilibria as weil as additional weak and strong versions. **In** these solutions the sender can make more delicate announcements of deviation from the given equilibrium and they ask for behavior with various levels of robustness against the deviating announcements when it is assumed that deviations are interpreted correctly by the receiver. **In** the background there is a notion of convention that not only assumes asender will do what he says as long as there is no incentive to deviate, but that the receiver interprets the messages in the marmer intended by the sender.

On the other hand, philosophers and, in particular, logicians studied both language, communication and interpretation long before it entered economie discourse. Moreover, there have been studies relating these to actions and mounting interest in the incorporation of strategie reasoning, as weil as the use of logic games (see van Benthem, 2007) which in turn is advancing game theoretical results. **In** relation to pre-play communication in games the work of Lewis (1969) sets the stage. With Gricean maxims (cf. Grice, 1957) shedding light on methodical principles for formulating foundations for refinement conditions as described above. As in similar cases of intersecting disciplines, the interaction works both ways with game theoretical principles utilized on the linguistic side of communication as seen in (Parikh, 2001), (van Rooij, 2003, 2004) and (Jäger, 2006), with dynamics-evolutionary and learning-e-entering the scene and bridging semantics and pragmatics.

Here we take a different approach to the strategie impact of content. We consider communication that may alter how decision makers perceive the strategie situation, or how they view its perception by others. **It** is the game Iorrn that rnay change as a result of cornrnunication. Using a recenLly developed framework for games with unawareness in (Feinberg, 2004, 2005b) we consider games where players may initially model the game differently, or view others as modeling it differently. **In** this paper we consider the impact of adding a communication stage to such games. The players can now reveal their perception. It turns out that some statements are inherently credible, without resorting to prior conventions or to credibility derived from the definition of the solution. As such, we find that the content of the message can inftuence the strategie behavior of the decision makers without resorting to a refinement or an arbitrary assignment of interpretations to statements.

Our aim is to link the modeling of strategie communication in economics with an almost pragmatic interpretation of what constitutes a game which allows us to insert meaning to messages. We find this to be a continuation of previous explorations along the interface between economics and language" with emphasis on applicability to economie settings. The next

4 We see the model here as putting pragmatism mueh closer to mainstream eeonomies when initially Rubinstein saw it as "the topie furthest from the traditional economie

section provides a game theoretical analysis of some games with unawareness to which we add a pre-play communication stage. We then discuss the aspects of strategie choice of messages in aspecific example of cal' sales where a dialogue sets the bargaining game.

2 **Meaningful** message

Our starting point is a game with unawareness-a game where the players may have a partial description of the strategie situation, or where they attribute such limited perception to others, view others as perceiving them to be restricted and so on. As a leading example consider the strategie situation modeled in the game depicted in Figure 1 and Figure 2 below taken from (Feinberg, 2üü5a). The game in Figure 1 represents all the actions available to the two players and the payoffs associated with each action profile. Assume that Alice and Bob are both aware of all the actions available in the game. However, Alice's perception is that Bob is not taking into account *all* her actions. **In** particular she is confident that Bob models the game as depicted in Figure 2. Alice's perception can be due to her observing that action *a3* was never taken in similar situations in the past, which would be supported by the observation that *a3* is not part of a Nash equilibrium in the standard normal form game in Figure 1, or she may assume that Bob is relatively a novice in this strategie situation, or that this is her "secret" action, whichever the reason, Alice views Bob as unaware of this action. Hence, Alice's view is that Bob's view of the game includes only the actions $\{a1,a2,b1,b2,b3\}$ and that the strategie situation he models is the game depicted in Figure 2. We further assume, as in (Feinberg, 2üü5a), that Bob actually views the game as Figure 1. He correctly recognizes that Alice considers all the actions. Furthermore, we assume that he recognizes that Alice is unaware that he is considering *a3,* i.e., Bob views Alice's view of his view of the game to coincide with Figure 2. Similarly, all other higher order iterations correspond to the game in Figure 2-Alice's view of Bob's view of Alice's view, Bob's view of Alice's view of Bob's view, Alice's and Bob's view of these views, and so on.

The game depicted in Figure 1 has a unique Nash equilibrium $(a2, b_1)$ obtained by iteratively eliminating strictly dominated strategies, hence a unique rationalizable outcorne". However, since Alice perceives Bob's view of the situation to be as depicted in Figure 2 she may believe that Bob is considering playing according to the Nash equilibrium $(a1, b2)$ of the game in Figure 2-the pareto optimal equilibrium.

issues" in his book (Rubinstein, 2000).

5 As noted above this may justify Alice's perception that *a3* may be overlooked by others.

		Bob		
		b_1	b_2	b_3
Alice	$a1$	0,2	3,3	0,2
	$a2$	2,2	2,1	2,1
	$a3$	1,0	4,0	0,1

FIGURE 1.

		Bob		
		b_1	b_2	b_3
Alice	$a1$	0,2	3,3	0,2
	$a2$	2,2	2,1	2,1

FIGURE 2.

Alice will be inclined to choose her best response to b_2 which is $a3$– her "secret" action. Bob can make the exact same deduction as we, the modelers, just did, since he is aware of all the actions and correctly perceives that Alice is unaware that he is considering that she may take the action $a3$. Hence, Bob can deduce that Alice assumes he plays b_2 and she chooses $a3$. Bob's best response is to play bz. The outcome will be a best response based on a Nash equilibrium of the restricted game where Alice plays $a3$ and Bob plays bz. The decision makers' reasoning (incorporating Nash equilibria reasoning) may lead them to actions that are not part of an equilibrium neither in the game Figure 1 nor in Figure 2. We end up with the worst possible payoff for Alice and a low payoff for Bob, although both are aware of all possible actions in the game, both are commonly aware of the action profile of the unique Nash equilibrium $(a2, b_1)$ and both act rationally given their perceived view of the game.

We wish to add pre-play communication to this game and consider the impact of messages on the strategie behavior. More precisely, assume that Bob can send a message to Alice. If Alice does not consider the possibility that Bob is actually aware of $a3$ there is no impact of sending messages that have no meaning as in cheap-talk communications since there are no differing types that Alice is considering and for which updated beliefs could result from messages sent by Bob. However, Bob can send the message "do not use action $a3$" or "I know about action $a3$" or even "don't think that 1 am not taking $a3$ into account". The specific reference to this action's *name*[6] *requires* Alice to reconsider her view of Bob's view of the game. The

6 The reader ean replaee the narnes of aetions with an appropriate deseription for a specific strategie interaetion. lt is here that Grieean pragmaties are invoked, see (Griee,

interpretation is that Bob has just related the action a_3 to the strategie interaction. She has modeled the situation with Bob being unable, or, practically unable, or even unwilling, to reason about her "secret" action, and here Bob is explicitly mentioning this action in the context of this game. Can Alice ignore this statement? We believe the answer is negative. What forces Alice to revise her view is not the fact that *a* message was sent, nor is it some preconceived interpretation of the statement, it is that the content of the message indicates that Bob can reason about aspecific action a_3 that she initially deemed relevant to the specific strategie interaction.

A formal framework for Alice's revision due to such a message follows from the representation of a game as a formal language as in (Feinberg, 2005b) when extended to restricted languages as in (Feinberg, 2004). Hence, the interpretation of a message can be a demonstration of the cognitive ability of the sender, in the sense of the extent of his ability (or his modeling choice) when reasoning about the situation at hand. With this interpretation, Bob can send a statement that includes the term a_3 only if a_3 is part of the language he employs in reasoning about the strategie situation at hand. Hence, Alice who already possesses a_3 in *her* description of the situation, must deduce that a_3 is part of Bob's language for the game which is a revision of her initial perception.

What follows aftel' Bob mentions a_3 and Alice revises her perception of Bob is the play of the unique Nash equilibrium of the game in Figure 1. **In** this specific example, Alice is motivated to choose a_2 even if she believes that Bob would not believe that she will revise her perception of his perception, i.e., even if Bob would still choose b_3 expecting her to choose a_3 she now realizes that this is possible, making a_2 a best response. WhaLever Lhe update of perception is, our main claim is that the content of a message can change the scope of reasoning that decision makers attribute to each other and hence dramatically inftuence strategie behavior and outcomes.

In this example we assumed that a_3 is not some generic statement. When Bob mentions a_3 Alice must find it is highly unlikely that he was able to guess it. iVIoreover, even if Alice suspects that Bob is lucky she would realize that he might consider such actions possi bie, leading her to the (a_2, b_1) solution once more. Our observation relies on the message containing a_3 and not on any promised action, or intention Bob expresses. If anything, when Alice hears Bob warning her not to take action a_3 she might even deduce something about Bob's reasoning about high order perceptions in this situation. She might realize that Bob is able to deduce that she is contemplating a_3 hence he might actually realize that she initially did not realize he is reasoning about a_3. Even without any warning attached, Bob's incentive for mentioning a_3 may lead Alice to deduce that Bob is not only

1957) and in the context of game see (van Rooij, 2003).

aware of *a3* but that he realizes that she perceived him to be unaware of it. Since, otherwise, why would he choose to mention *a3* of all things? While this transmission of higher order perception will not impact the outcome in this case, assuming Alice realizes that Bob considers *a3* to begin with, it could be of strategie impact in other settings. It also sheds some light on the various notions of credibility associated with a message and the derived high order perceptions.

There are two different notions of credibility that can be attributed to a statement made by Bob. The first is the one that Farrell (1993) (see also Myerson, 1989) suggested which corresponds to a weil developed body of work in logic-see (Grice, 1957), (Lewis, 1969) and more recent results in (van Rooij, 2003). It corresponds to the case where Bob makes the statement "I am going to play *bs"*. Alice is assumed to interpret it according to the convention that Bob is indeed intending to play the action with the same name and credibility is generated when it survives the economie incentive criteria for Bob once Alice acts rationally and assumes that he speaks truthfully-robustness to deviation. The second notion of credibility of a statement we introduced here corresponds to a statement whose content reveals perception, or ability, that may not have been assumed to begin with. When Bob mentions action *a3* he expressed awareness that was beyond Alice's perception of Bob. This may lead her to assume that he is modeling the game with action *a3* and the exact marmer in which *a3* is stated, e.g., as a promise or a threat, matters less. To distinguish from the first notion of credibility we call a statement *meaningful* if it conveys an aspect of the perception of a strategie situation.

InLermediaLe notions between credible and meaningful talk also arise. For example, a player may want to describe their own action when they perceive that ot hers do not incorporate it into their view of the game. But here, there must be some evidence convincing the other side. Such evidence maybe exogenous (e.g., demonstrating a proof of the existence of the action), as in a convention, or be supported from intrinsic strategie motivation. In the example above, when Bob mentioned *a3* the statement was more meaningful due to its credibility, i.e., Alice had a good reason to believe that Bob was not just guessing, because strategically it is beneficial for Bob to make such a statement if he is aware of *a3,* while if he thinks the game is as in Figure 2 then even if he convinces Alice to believe in a fictitious *a3* he would, from the limited point of view, lose by leading to the lower payoff equilibrium in Figure 2. Alice should correctly retrieve its meaning-update her view of Bob's perception of the game---since this deduction is also supported when considering Bob's incentives.

In general, the interaction of credibility and meaning may be more intricate and sometimes conftieting. For example, consider the game depicted

		Bob		
		b_1	b_2	b_3
Alice	a_1	0,0	3,5	1,1
	a_2	5,2	2,1	O,x

FIGURE 3.

in Figure 3. Assume that in this game Alice is unaware of b_3 and perceives the game as in Figure 4 where two pure strategy Nash equilibria compete for coordination. The strategy profile (a_2, b_1) is much more attractive to Alice in Figure 4. Assume that $5 > x > 2$ then if Bob says "do not play a_2 because I will never play b: since I have a strategy b_3 that dominates it" (or simply says "I have b_3") then Alice may consider a revision of the perception of the game. The description of b_3 mayor may not convince her. But the situation becomes puzzling if $x > 5$. **In** this case Bob has no incentive telling Alice about this action-he would rather have her choose a_2. The problem is what should Alice deduce if he does describe action b_2. Should she revise her perception, or should she just assume that he is trying to lead her to the solution of Figure 4 which benefits him? How should she weigh any evidence of the existence of b_3 that Bob may produce? Her problem is that if she accepts his statement as meaningful it would be irrational for him to make it, if she doesn't, then stating it is no longer irrationa1. Hence, the level of verifiability of the content may mix with the notion of credibility of statements and with the meaningfulness of a statement which in turn determines its credibility.

To complicate things even further, we point out that credibility and meaning can interact with the incentives for making any statement at all. **If** one assumes credibility to begin with[7] it might inftuence deduction about the perception of others. Bob would rather not reveal b_3 when $x > 5$, but the strategic situation might force him to make a statement about actions. **If** he believes that Alice will interpret his statements as truthful unless they are not credible, then he has to choose between saying "I will play b_1" and "I will play b_2". **If** he says the second then Alice will find it credible since she views the game as Figure 4, but Bob would lose the opportunity of gaining more than 5. However, if he says he will play b: she would find it to be non-credible since having the opportunity to determine the focal point according to the convention, Bob should be strictly better off declaring b_2 in Figure 4. **In** this case Bob would like to have the right to be silent, but only if Alice does not realize that Bob had the opportunity to make a

7 In the traditional sense that a stated action is played if there is an incentive to foIlow through if it is believed.

		Bob	
		b_1	b_2
Alice	$a1$	0,0	3,5
	$a2$	5,2	2,1

FIGURE 4.

statement. **In** this case, the perception of *uilietlier* statements can be made may inftuence the strategie behavior.

This example leads to an additional twist which concerns with how credibility may trigger reasoning. **If** Alice reasons by assuming credibility then when Bob armounces "I will play b_1" he might trigger her to wonder if she is missing something in the strategie environment. This in turn may lead her to discover the missing action or develop beliefs about the consequences of its existence. She might very weil conclude that she is better off playing $a1$ (maybe even declaring it). But if this is the case, we may ask whether we could have a game with no unawareness at all where Bob would deliberately make an announcement that may cause Alice to believe that she is missing some aspect of the game, leading her to behavior that Bob could not induce otherwise. The conceptual answer seems negative since if the armouncement eventually leads Alice to behave in a way that benefits Bob beyond any believed announcement, then it is actually a credible declaration by Bob and Alice's reasoning should unravel it. However, we lack the formal machinery for discovery in this setting.

We note that in a dynamic setting meaningful indirect deductions can be made not only from statements but also from observing actions. Consider the first example where Alice assumes that Bob views the game as Figure 2. Assume the game is repeated twice but that only Alice can observe the choices made in the first round before playing the second round, i.e. even if Alice chooses *a3* in the first round, Bob will not become aware of *a3* in the second round if he was previously unaware of it. **In** this case, even without communication, if Alice observes Bob choosing b_3 in the first round she can either assume that Bob is irrational since b_3 is dominated in Figure 2, or she can deduce that Bob is aware of *a3* and followed the reasoning described in the beginning of this section.

We reiterate that the use of a variety of messages types in these examples such as "do not choose *a3",* "I know about *a3",* "I realize you are considering *a3"* and "I am going to do... because of *a3"* all have the same impact in the sense of meaningful talk. **It** is the mentioning of *a3* in any form which conveys an ability to reason about it in the specific context and impacts the perception of the game leading to strategie deductions. The messages

become reliable indications of the extent of the language that the decision maker employs in the situation. **In** general, both messages and actions can inftuence perceptions, deductions about others perceptions and so on. The phrases in the language which describe the situation from an economie view point-the game that is being played-convey meaning because they reveal the perception of the game which may not be commonly known.

3 Negotiating a bargain

The interaction of meaningful talk and strategie behavior is not limited to pre-play communication. As indicated above the strategie choice of statements takes into account the derived meaning by the receiver. Furthermore, since meaningful statements change the perception of the strategie situation, they could be highly sensitive to the choices made in adynamie setting-a dialogue.

To illustrate some of the strategie features of a dialogue consider the case of a new cal' purchase. Alice is an inexperienced buyer who is facing the experienced cal' salesperson Bob. Before explicit bargaining over the price begins (or while price discussions are held), there is a dialogue in which statements are made, questions are asked and answers are given. The strategie decision making involved in this process is far from trivia!. We focus only on a partial number of dimensions of the cal' purchasing case with emphasis on the items that set the beliefs over reservation prices, most notably the dealer's reservation prices. Throughout this story we assume that Bob is fully aware of all possible dimensions of the interaction, he will not be informed of Alice's private information and may possess uncertainty as to Alice's perception of the situation, but he has full command of what Alice *might* be reasoning about.

We simplify the situation further by assuming that Alice made her choice of vehicle she wishes to purchase", the cal' is available at the dealership and Alice has no trade-in. Alice may be unaware of the following dimensions of the transaction:

1. The sticker price on the cal' (the MSRP) has no immediate relation to the dealer's cost or relationship with the manufacturer.

2. There is an "invoice price" which is the basis for the cost of the cal' to the dealer.

3. There are adjustments to the dealer's invoice, such as quota incentives, year end bonuses, opportunity costs (new modeis, lot space).

8 This is astrong assumption as in many cases salespeople have an extensive range of inftuence on how the buyer perceives and compares potential substitutes.

4. The actual cost of various additional options (such as fabricjpaint protection treatments, extended warranties).

5. The existence of a nearby competing dealership selling an identical cal'.

6. The possibility of purchasing the cal' over the Internet.

7. The possibility of financing a cal' purchase in favorable terms compared to unsecured consumer credit such as credit card purchases.

8. The information the salesperson has regarding the dealership's costs.

We use this partial list of concrete attributes of the specific strategie situation to highlight some aspects of a strategie dialogue that incorporates meaningful content being revealed via statements.

Being the more experienced party, the salesperson's problem is to discover not only as much as possible about Alice's reservation prices (assuming her taste was completely revealed by the choice of cal' and model), but also to inftuence her perception of the strategie situation in a marmer that puts her at a disadvantage in the bargaining stage and eventually leads to the highest profits to the dealership''. This may amount to things as obvious as trying to sell options as in 4 aftel' the price has been agreed upon, but it also relates to inftuencing how Alice models the bargaining game itself, in particular, what bounds she sets on the dealership costs. Alice, on the other hand, must realize she is in a situation where there are potential aspects she might be missing, aftel' all she does not buy cars very often and should assume that Bob is experienced.l''

We list various aspects which are relevant to a strategie dialogue without formulating a detailed decision tree, in which the choice of statements and questions are presented and the two parties perception of the game (including higher order perceptions) evolves according to these statements. Instead, we focus on various considerations that inftuence the optimal choice of statements, this allows us to capture a variety of scenarios without detailing the full scope of the game for each one. We refer to (Feinberg, 2004) for an illustration of how these dynamic games with unawareness can be constructed.

The first aspect is the timing of bargaining and the importance of having the dialogue at all. As we will see, Bob can strictly benefit from obtaining information from Alice and hence has an incentive to engage in a dialogue

9 The salespersen commission is increasing in the dealership profit from the sale.

10 This is a very particular notion of experience, as it relates to the magnitude of the set of attributes of the situation that a decision maker is aware of, as weil as the ability to consider the limited perception of the buyers.

before discussion of prices occurs. For example, if Alice asks what the price is, Bob would refer her to the sticker price, or may even delay the answer by stating it depends on other factors, such as financing, options etc. We conclude that the ability to lead the other party into the dialogue has value on its own. In particular, there will be use of statements that might try and convince Alice as to why a lengthier discussion is required.

Much of the information on Alice's perception can be gained by Bob from indirect questions. Indirect questions are preferabie for Bob since they are less likely to reveal the very same aspect of which Alice may be unaware. For example, to adjust his beliefs on whether Alice is aware of 5 Bob may prefer to ask "have you driven this cal' before?" or "have you seen feature X in this car?" rather than "have you been at another dealership?" . On the other hand, to find out whether she considered 6 Bob may need to ask Alice if she has seen the cal' manufacturer web site demonstration of some feature, or offer to e-mail her the link, and by doing that she might be moved to consider Internet pricing. Hence, the statements/questions that Bob uses to elicit information about awareness may require a trade-off since they could provide Alice with valuable aspects of the situation and sometimes the exact aspect that Bob was hoping to exploit.

A third aspect of the strategie dialogue is the order of statements made. We can illustrate this with Alice's strategie choices. If she is aware of both 6 and 7 she would prefer to reveal 6 before the price negotiations begin since she could provide a price quote which sets the upper bound to her reservation price. But in doing so she would indicate that she is likely to have obtained, or be aware of, various financing opportunities. She may want Lo hide this inforrnal.ion unLil aſLel' Lhe price is set and befere Lhe financing arrangement is made, the reason being that the salesperson may be willing to lower the price if he believes there is a high probability he can have a substantial gain from the financing agreement. Alice would like to make both statements, however their relative piace along the dialogue may inftuence the final outcome.

The next aspect involves perception of attributes for which there is littie information. Alice might be aware that the dealership may have some incentive from the manufacturer, i.e., of 3. However, she may not know the size of this incentive. Bob would probably not be willing to supply this information, or may not have access to it (see the discussion below). If Alice claims that she is aware of such incentives for the dealership yet is unable to provide any description of their magnitude, or some supporting argument, her statement is less credible and Bob may rightly deduce that she is only aware of the possibility but has no reason to assign high probability to a substantial impact on the dealership's costs. If credibility of one statement affects that of another, there is a cost associated with statements that are meaningful but not specific enough.

We note that Bob may not be informed of the dealership's actual costs if there is a concern that he might reveal it to potential buyers. In many cases cal' salespeople need the authorization of supervisors to make any significant price reductions along the bargaining process, this procedure allows managers to supervise the bargaining, but also to make pricing decisions based on information that may not be readily available to the salesperson. iVIoreover, Alice's awareness of 8 would lead her to treat the game differently, as the solutions to bargaining with an intermediary could differ from direct negotiations. This also applies to Alice's perception of the incentives of the salesperson and could explain why firms tend to prominently state when their salespeople are not werking for commissions.

As we noted in the previous section, an action can also yield a change of perception from both operational reasons and by deduction. For example, Bob may ask for Alice's driver license before a test drive, this might allow him to retrieve information about Alice that she is not aware he might be able to obtain. On the other hand, it could also alert Alice that her financial background could have impact on the financing offered and lead her to consider other lenders-7.

Finally, a combination of higher order reasoning about perception may interact with strategic meaningful talk. For example, Alice may consider 3 and even assume that stating it could be reliable, for example, she might have read that dealers are about to obtain delivery of a new model and have an incentive to deal' lot space. She mayalso be aware of 2-the existence of a formal invoice price, but she may not know its exact value. Bob may not realize that Alice is aware of 3. As such, Bob might be inclined to reveal the invoice price Lo Allee if he percel ves that she will model Lhe reserval.ion price of the dealership as no less than this invoice price. If Alice perceives that Bob is modeling her as unaware of 3 then Alice may decide to mention that she knows about invoice and try to induce Bob to show her documentation of this price, obviously making no mention of 3. Hence, Alice can obtain valuable information by reasoning about how Bob models her awareness and choosing her statements in the corresponding manner. Once she has the invoice price, Alice can more accurately estimate the dealership's true costs.

4 Summary

We have set out to present communication in games where the content of messages impacts strategic behavior without resorting to ex-ante assumed conventions. Our main observation is that in the case of games with unawareness the content of a message can change the perception of the game being played leading to changes in strategic behavior. Potentially the strongest form of meaningful talk we identified was when Bob told Alice

something about the situation that she already thought about, but that she did not realize Bob is also able to consider. This had led us to various new strategic considerations when the interpretation of the message is not only driven from a convention, or an arbitrary assignment of interpretation as in cheap talk, but is also derived from the expected change in perception of what game is being played.

We used an illustrative example to show how aspects such as the timing and order of statements, indirect questions, deliberate withholding of information from agents, trade-offs between revelation and discovery of perceptions, and the discovery of high order perceptions, inftuence strategic choices of meaningful statements in a dialogue which determines the bargaining game perceptions and eventually its outcome.

We conclude that in a strategic setting a statement becomes meaningful as it describes the extent of the speaker's reasoning. **In** addition, statements become tools for molding perceptions and questions become tools for discovery of perceptions by generating answers. The implication is that dialogues are strategically intricate exchanges, since, to adopt a pragmatic approach, a statement does not only depend on the context but defines the context and determines the relevant strategic situation.

References

van Benthem, J. (2007). Logic in Games. Lecture notes. Institute for Logic, Language and Computation.

Cho, I. & Kreps, D.M. (1987). Signaling Games and Stabie Equilibria. *The Quarterly Journal of Economics,* 102(2):179-221.

Crawford, V.P. & Sobel, J. (1982). Strategic Information Transmission. *Econometrica, 50(6):1431-1451.*

Farrell, J. (1993). Meaning and Credibility in Cheap-Talk Games. *Games and Economie Behavior, 5(4):514-531.*

Feinberg, Y. (2004). Subjective Reasoning - Games with Unawareness. Research Paper 1875, Stanford, Graduate School of Business.

Feinberg, Y. (2005a). Games with unawareness. Research Paper 1894, Stanford, Graduate School of Business.

Feinberg, Y. (2005b). Subjective reasoning-dynamic games. *Games and Economie Behavior, 52(1):54-93.*

Grice, H.P. (1957). Meaning. *Philosophical Review, 66(3):377-388.*

Jäger, G. (2006). Game Dynamics Connects Semantics and Pragmatics. Manuscript, University of Bielefeld.

Lewis, D. (1969). *Convention.* Harvard University Press.

Matthews, S.A., Okuno-Fujiwara, M. & Postlewaite, A. (1991). Refining cheap-talk equilibria. *Joumal of Economie Theorsj, 55(2):247-273.*

Myerson, R.B. (1989). Credible Negotiation Statements and Coherent Plans. *Joumal of Economie Theory, 48(1):264-303.*

Parikh, P. (2001). *The Use of Language.* CSLI Publications.

van Rooij, R. (2003). Quality and Quantity of Information Exchange. *Journal of Logic, Language and Information; 12(4):423-451.*

van Rooij, R. (2004). Signaling Games Select Hom Strategies. *Linguistics and Philosophy,* 27(4):493-527.

Rubinstein, A. (2000). *Economics and Language.* Cambridge University Press, Cambridge.

Spence, M. (1973). Job market signaling. *The Quarterly Joumal of Economics, 87(3):355-374.*

Spence, M. (2001). Signaling in retrospect and the informational structure of markets. http://nobelprize.org/nobel_prizes/economics/laureates/2001/spence-lecture.pdf. Nobel Prize Lecture.

A Study In the Pragmatics of Persuasion: A Game Theoretical Approach*

Jacob Glazer[1,2]
Ariel Rubinsteiriv"

1 The Faculty of Management
Tel Aviv University
Tel Aviv, 69978, Israel

2 Department of Economics
Boston University
270 Bay State Road
Boston MA 02215, United States of America

3 The School of Economics
Tel Aviv University
Tel Aviv, 69978, Israel

4 Economics Department
New York University
19 W. 4th Street, 6FL
New York NY 10012, United States of America

{glazer,rariel}@post.tau.ac.il

Abstract

A speaker wishes to persuade a listener to take a certain action. The conditions under which the request is justified, from the listener's point of view, depend on the state of the world, which is known only to the speaker. Each state is characterized by a set of statements from which the speaker chooses. A persuasion mie specifies which statements the listener finds persuasive. We study persuasion mies that maximize the probability that the listener accepts the request if and only if it is justified, given that the speaker maximizes the probability that his request is accepted. We prove that there always exists a persuasion mie involving no randomization and that all optimal persuasion mies are ex-post optima!. We relate our analysis to the field of pragmatics.

1 Introduction

A persuasion situation involves an agent *(the speakers* who attempts to persuade another agent *(the listeneT)* to take a certain action. Whether

* Originally appeared in Theoretical Economics 1(4), pp. 395-410, 2006.

Krzysztof R. Apt, Robert van Rooij (eds.). *New Perspectives on Games and Interaction.* Texts in Logic and Games 4, Amsterdam University Press 2008, pp. 121-139.

or not the listener should accept the speaker's suggestion depends on information possessed by the speaker. **In** such a situation, the speaker often presents hard evidence to support his position, but is restricted as to how many pieces of evidence he can present. This restrietion may be due either to time constraints or to limitations on the listener's capability to process information. Our purpose in this paper is to shed light on the rules that determine which of the facts, presented by the speaker, the listener will find persuasive.

The topic of this paper is related to a field in linguistics called pragmatics, which explores the rules that determine how people interpret an utterance, made in the course of a conversation, beyond its literal content (see Grice, 1989). Grice suggested that the leading principle in the interpretation of utterances is what he termed the "cooperative principle", according to which the interpretation of utterances in a regular conversation can be made on the assumption that the speaker and the listener have common interests. However, the cooperative principle does not appeal' to be relevant in a persuasion situation in which the agents may have confticting interests.

The following example ciarifies the distinction between the pragmatics of conversation and the pragmatics of persuasion: You are discussing the chances of each of two candidates in an upcoming election. The electorate consists of ten voters. Assume that the other person has access to the views of these ten voters. Imagine that he has just informed you that *a, d,* and *9* support candidate *A*. **If** it is a friendly conversation, then you are most likely to think that he has selected three people who represent the views of the majority of the voters. Thus, you are likely to be persuaded that *A* is likely Lo win Lhe election. H, on Lhe oLher hand, independenLly of Lhe truth, the other person is trying to persuade you that *A* will win, you will find this very same statement to be a weak argument since you will suspect that he has intentionally selected three supporters of *A*.

What governs the pragmatic rules of persuasion? We propose an approach analogous to Grice's cooperative principle in which the pragmatic rules of persuasion are determined by a fictitious designer before the discourse begins. These rules govern the speaker's choice of facts to present in the knowledge that the listener will interpret his statements according to these rules. The rules are structured by the designer to maximize the probability that the listener will make the "right" dec ision (from his point of view and given the "true" situation) on the basis of the information provided to him by a self-interested speaker and subject to constraints on the amount of information that can be submitted to him by the speaker.

We conduct our investigation within the narrow boundaries of a particular model in which several assumptions admittedly play a critical role. Our analysis is faithful to economie tradition rather than to the method-

ology of Pragmatics. Nevertheless, we believe that the study conducted here demonstrates the potentialof research to find a uniform principle that guides individuals in interpreting statements in persuasion situations.

This paper belongs to a research program in which we apply a game theoretical approach to issues in Pragmatics. **In** (Glazer and Rubinstein, 2001) we study an example of a debate situation involving two parties each of whorn tries to persuade a third party to accept his position. Even closer to this paper is (Glazer and Rubinstein, 2004), which analyzes a persuasion situation in which aftel' the speaker makes his case the listener can obtain partial information about the state ofthe world. Aftel' specifying our current model, we will compare it to the one in (Glazer and Rubinstein, 2004).

2 The model

A speaker wishes to persuade a listener to take a certain action. The listener can either accept or reject the speaker's suggestion (there is no partial acceptance). Whether or not the listener should be persuaded depends on the *state,* which is an element in a set X. A set $A \subset X$ consists of all the states in which the listener would wish to be persuaded (i.e. to accept the speaker's suggestion) if he knew the state, and the set $R = X \setminus A$ consists of all the states in which the listener would wish to reject the speaker's request. The listener's initial beliefs about the state are given by a probability measure p over X. Denote by p_x the probability of state x.

We assume that for every state x, there is a set of statements $\sigma(x)$ that the speaker can make. Let $S = \cup_{x \in X} \sigma(x)$. The meaning of "making statement s" is to present proofthat the event $\sigma^{-1}(s) = \{x \mid s \in \sigma(x)\}$ has occurred.

In state x the speaker can make one and only one of the statements in $\sigma(x)$. Thus, for example, if the speaker can choose between remaining silent, making the statement α, making the statement β, or making both statements, the set $\sigma(x)$ consists of the four elements *silence,* α, β, and $\alpha \wedge \beta$.

To summarize, we model a *persuacion problem* as a four-tuple (X, A, p, σ). We say that the persuasion problem is finite if X is finite. We refer to the pair $\langle X, \sigma \rangle$ as a *signal struciure.*

Comment. We say that a signal structure (Y, e) is *oecioric* if Y is a product set, i.e. $Y = \times_{k \in K} Y_k$ for some set K and some sets Y_k, $k \in K$, and the speaker in state x can make a statement concerning the value of one of the components ofx, that is, $e(x) = \{(k,v) \mid k \in K \text{ and } v = x_k\}$.

One might think that we could make do by analyzing only vectoric signal structures. To see that this is not the case, let (X, σ) be a signal structure. Let (Y, e) be the vectoric signal structure with $Y = \{0, 1\}^S$. Every state $x \in X$ can be represented by the vector $\varphi(x) \in Y$, which indicates the statements available at x, that is, $cp(x)(s) = 1$ if $s \in \sigma(x)$ and Ootherwise.

However, the two structures are not equivalent. First, we allow for the possibility that two states have the same set of feasible statements. Second, and more importantly, in the corresponding vectoric structure the speaker in any state is able to show the value of the component that corresponds to any statement s. **In** other words, he is always able to prove whether s is available or not. **In** contrast, in our framework the fact that the speaker can make the statement s does not necessarily mean that he can make a statement that proves that s is not available.

We have in mind a situation in which the speaker makes a statement and the listener must then either take the action *a,* thus accepting the speaker's position, or the action r, thus rejecting it. A persuasion rule determines how the listener responds to each of the speaker's possible statements. We define a *persuasioti rule* f as a function $f : S \to [0,1]$. The function f specifies the speaker's beliefs about how the listener will interpret each of his possible statements. The meaning of $f(s) = q$ is that following a statement s, with probability q the listener is "persuaded" and chooses *a,* the speaker's favored action. We call a persuasion rule f *determinisiic* if $f(s) \in \{0, 1\}$ for all $s \in S$.

We assume that the speaker wishes to maximize the probability that the listener is persuaded. Thus, given a state *x,* the speaker solves the problem $\max_{s \in \sigma(x)} f(s)$. The value of the solution, denoted by $\alpha(f, x)$, is the maximal probability of acceptance that the speaker can induce in state *x.* For the case in which $\sigma(x)$ is infinite, the solution can be approached but is not attainable and therefore we define $\alpha(f, x) = \sup_{s \in \sigma(x)} f(s)$.

Given the assumption that the speaker maximizes the probability of acceptance, we define the (listener's) error probability $\mu_x(J)$ in state x as follows: If $x \in A$, then $\mu_x(f) = 1 - \alpha(f, x)$, and if $x \in R$, then $\mu_x(f) = \alpha(J,x)$. The *error pTobability* induced by the persuasion rule f is $m(J) = \sum_{x \in X} p_x \mu_x(J)$. Given a problem $\langle X, A, p, \sigma \rangle$, an *optimal* persuasion rule is one that minimizes $m(J)$.

Note that persuasion rules are evaluated according to the listener's interests while those of the speaker are ignored. **In** addition, we assume that all errors are treated symmetrically. Our analysis remains the same if we add a variabie c_x for the (listener's) "costs" of an error in state x and define the objective function to minimize $\sum_{x \in X} p_x c_x \mu_x(f)$.

Example 2.1 ("The majority of the facts supports my position"). There are five independent random variables, each of which takes the values 1 and 0 each with probability 0.5. A realization of 1 means that the random variabie supports the speaker's position. The listener would like to accept the speaker's position if and only if at least three random variables take the value 1. **In** the process of persuasion, the speaker can present the realization of at most m random variables that support his position.

Formally, $X = \{(X_1, \ldots, X_S) \mid X_k \in \{0, 1\} \text{ for all } k\}$, $A = \{x \mid n(x) \geq 3\}$ where $n(x) = \sum_k x_k$, $P_x = \frac{1}{32}$ for all $x \in X$, and $\sigma(x) = \{\kappa \mid \kappa \subseteq \{k \mid x_k = 1\} \text{ and } |\kappa| \leq m\}$.

If m = 3, the optimal persuasion rule states that the listener is persuaded if the speaker presents any three random variables that take the value 1. The more interesting case is m = 2. If the listener is persuaded by the presentation of any two random variables that support the speaker's position, then the error probability is $\frac{10}{32}$. The persuasion rule according to which the listener is persuaded only by the speaker presenting a set of two "neighboring" random variables ({1,2}, {2,3}, {3,4}, or {4,5}) with the value 1 reduces the error probability to $\frac{5}{32}$ (an error in favor of the speaker occurs in the four states in which exactly two neighboring random variables support the speaker's position and in the state (1,0,1,0,1) in which the speaker is not able to persuade the listener to support him even though he should).

The two mechanisms above do not use lotteries. Can the listener do better by applying a random mechanism? What is the optimal mechanism in that case? We return to this example aftel' presenting some additional results.

Comment. At this point, we wish to compare the cutrent model with the one studied in (Glazer and Rubinstein, 2004). Both models deal with a persuasion situation in which (a) the speaker attempts to persuade the listener to take a particular action and (b) only the speaker knows the state of the world and therefore whether or not the listener should accept the speaker's request.

Unlike the cutrent model, the speaker in the previous model could first send an arbitrary message (cheap talk) to the listener. Aftel' receiving the message, the listener could ask the speaker to present some hard evidence to support his request. The state of the world in that model is a realization of two random variables and the listener is able to ask the speaker to reveal at most one of them. Thus, unlike the cutrent model, in which the speaker simply decides which hard evidence to present, in the previous model the speaker has to "follow the listener's instructions" and the listener can apply a random device to determine which hard evidence he asks the speaker to present. That randomization was shown to often be a critical element in the listener's optimal persuasion rule (a point further discussed below). On the other hand, in the previous model we do not allow randomization during the stage in which the listener finally decides whether or not to accept the speaker's request, which we do allow in the current model. Allowing for such randomization in the previous model, however, is not beneficial to the listener, as we show to be the case in the current paper as weil.

The randomization in the previous paper is employed during the stage in which the listener has to decide which hard evidence to request from the speaker. Note that if in that model we restriet attention to deterministic persuasion rules, then it is a special case of the cutrent model. Eliminating randomization on the part of the listener in order to verify the information presented by the speaker, allows us to think about the persuasion situation in the previous model as one in which the speaker chooses which hard evidence to present rather than one in which the listener chooses which hard evidence to request.

Randomization plays such an important role in the previous model because it is, in fact, employed as a verification device. Without randomization, there is no value to the speaker's message since he could be Iying. The listener uses randomization to induce the speaker to transfer more information than the information that is eventual!y verified.

Although the cutrent model draws some inspiration from the previous one, the two papers relate to different persuasion situations and the results of the current paper cannot be derived from those of the previous one.

3 Two lemmas

We now present two lemmas that are useful in deriving an optimal persuasion rule.

3.1 A finite number of persuasive statements is sufficient

Our first observation is rather technical though simpie. We show that if the set of states X is finite then even if the set of statements S is infinite there is an optimal persuasion rule in which at most $|X|$ statements are persuasive with positive probability.

Lemma 3.1. Let $\langle X, A, p, \sigma \rangle$ be a finite persuasion problem.

1. An optimal persuasion rule exists.
2. There is an optimal persuasion rule in which $\{s \mid f(s) > 0\}$ does not contain more than $|X|$ elements.

Proo]. Consider a partition of S such that s and s_i are in the same cel! of the partition if $\sigma^{-1}(s) = \sigma^{-1}(s')$. This partition is finite. Let T be a set of statements consisting of one statement from each cel! of the partition. We now show that for every persuasion rule i, there is a persuasion rule 9 that takes a positive value only on T, such that $\alpha(g,x) = \alpha(f,x)$ for al! x and thus $m(g) = m(J)$.

For every $s \in T$ let S_s be the cel! in the partition of S that contains s. Define $g(s) = \text{SUP}_{S/ES_s} f(SI)$. For every $s \notin T$ define $g(s) = 0$.

For every state x,

$$\alpha(g,x) = \max_{s\in T\cap\sigma(x)} g(s) = \max_{s\in T\cap\sigma(x)} \sup_{s'\in S_s} f(s') = \sup_{s'\in\sigma(x)} f(s') = \alpha(f,x).$$

Thus, we can confine ourselves to persuasion rules that take the value 0 for any statement besides those in the finite set T. Any such persuasion rule is characterized by a vector in the compact set $[0,1]^T$. The error probability is a continuous function on this space and thus there is an optimal persuasion rule j^* with $j^*(s) = 0$ for all $s \notin T$.

For every $x \in X$ let $s(x) \in \sigma(x)$ be a solution of $\max_{s\in\sigma(x)} j^*(s)$. Let g^* be a persuasion rule such that

$$g^*(s) = \begin{cases} f^*(s) & \text{if } s = s(x) \text{ for some } x \\ 0 & \text{otherwise.} \end{cases}$$

The persuasion rule g^* is optimal as weil since $\alpha(g^*,x) = \alpha(f^*,x)$ for all x and thus $m(g^*) = m(J^*)$. Thus, we can confine ourselves to persuasion rules for which the number of statements that persuade the listener with positive probability is no larger than the size of the state space. Q.E.D.

3.2 The "L-principle"

The following result is based on an idea discussed in (Glazer and Rubinstein, 2004).

Let $\langle X, A, p, \sigma\rangle$ be a persuasion problem such that for all $x \in X$, $\sigma(x)$ is finite. We say that a pair (x, T), where $x \in A$ and $T \subseteq R$, is an L if for any $s \in \sigma(x)$ there is $t \in T$ such that $s \in \sigma(t)$. That is, an L consists of an element x in A and a set T of elements in R such that every statement that can be made by x can also be made by some mernber of T. An L, (x, T) is minimal if there is no $T' \subset T$ such that (x, T') is an L.

Lemma 3.2 (The L-Principle). Let (x, T) be an L in the persuasion problem $\langle X, A, p, \sigma\rangle$ and let f be a persuasion rule. Then $\sum_{t\in\{x\}\cup T} \mu_t(f) \geq 1$.

Proof. Recall that $\mu_x(J) = 1 - \alpha(f,x)$ and for every $t \in T$, $\mu_t(J) = \alpha(f,t)$. Therefore,

$$\begin{aligned} \sum_{t\in\{x\}\cup T} \mu_t(f) &\geq \mu_x(f) + \max_{t\in T} \mu_t(f) \\ &\geq \mu_x(f) + \max_{s\in\sigma(x)} f(8) \\ &= \mu_x(f) + oi\ ],x) = 1. \end{aligned}$$

Q.E.D.

The following example demonstrates how the L-principle can be used to verify that a certain persuasion rule is optima!. For any persuasion problem, the L-principle provides a lower bound on the probability of error that can be induced by a persuasion rule. Thus, if a particular persuasion rule induces a probability of error equal to a Iower bound derived from the L-principle, then one can conclude that this persuasion rule is optima!.

Example 3.3 ("I have outperformed the population average"). Consider a situation in which a speaker wishes to persuade a listener that his average performance in two previous tasks was above the average performance of the population. Denote by x_1 the proportion of the population that performed worse than the speaker in the first task and by x_2 the proportion of the population that performed worse than the speaker in the second task. The speaker wishes to persuade the listener that $x_1 + x_2 \geq 1$. The speaker knows his relative performance in the two tasks (that is, he knows x_1 and x_2) but can present details of his performance in only one of the tasks. We assume that the speaker's performances in the two tasks are uncorrelated. Formally, the signal structure is vectoric with $X = [0,1] \times [0,1]$; the probability measure p is uniform on X; and $A = \{(x_1, x_2) \mid x_1 + x_2 \geq 1\}$.

Note that if a statement is interpreted by the listener based only on its content, i.e. by stating that his performance was above $\frac{1}{2}$ in one of the tasks, the speaker persuades the listener and the probability of error is $\frac{1}{4}$.

The following argument (borrowed from Glazer and Rubinstein, 2004) shows that there exists an optimal persuasion rule according to which the listener is persuaded by the speaker if and only if the speaker can show that his performance in one of the two tasks was above $\frac{2}{3}$. Furthermore, the minimal probability of error is $\frac{1}{6}$.

A minimal L in this case is any pair $(x, \{y, z\})$ where $x \in A$, $y, z \in R$, $x_1 = y_1$, and $x_2 = z_2$.

The set $T_1 = \{(x_1,x_2) \in A \mid x_1 \leq \frac{2}{3}$ and $x_2 \leq \frac{2}{3}\}$ is one of the three triangles denoted in 1 by the number 1. Any three points $x = (x_1, x_2) \in T_1$, $Y = (x_1 - \frac{1}{3}, x_2) \in$ Rand $z = (x_1, x_2 - \frac{1}{3}) \in R$ establish an L. Ey the L-principle, for any persuasion rule f we have $\mu_x(f) + \mu_y(f) + \mu_z(f) \geq 1$. The collection of all these L's is a set of disjoint sets whose union is the three triangles denoted in the figure by the number 1. Therefore, the integral of $\mu_x(J)$ over these three triangles must be at least the size of T_1, namely $\frac{1}{18}$. Similar considerations regarding the three triangles denoted by the number 2 and the three triangles denoted by the number 3 imply that the minimal error probability is at least $\frac{1}{6}$. This error probability is attained by the persuasion rule according to which the listener is persuaded if and only if the speaker shows that either x_1 or x_2 take a value of at least $\frac{2}{3}$.

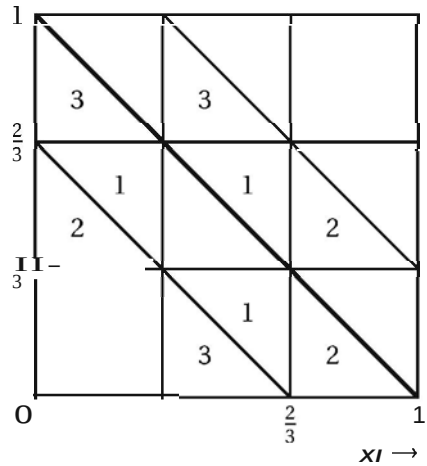


FIGURE 1. An optimal persuasion rule for 3.3.

4 Randomization is not needed

The next question to be addressed is whether randomization has any role in the design of the optimal persuasion rule. **In** other words, can the listener ever do better by making the speaker uncertain about the consequences of his statement? Glazer and Rubinstein (2004) show that in persuasion situations in which the listener can acquire partial information about the state of the world, uncertainty regarding what information he will acquire can be a useful device to the listener. However, as stated in 4.1 below, uncertainty is not useful to the listener in the present context.

Proposition 4.1.

1. For every finite persuasion problem $(X, A, p, \sigma\rangle$, there exists an optimal persuasion rule f that is deterministic.

2. For every persuasion problem $(X, A, p, \sigma\rangle$ and every $c > 0$, there exists a deterministic persuasion rule f^* such that $m(J^*) < \inf_f m(J) + s$.

Proo]. (1) Ey 3.1, there exists an optimal persuasion rule with a finite number of statements that induces acceptance with positive probability. Consider an optimal persuasion rule f with the fewest non-integer values. Let $0 < \alpha_1 < \ldots < o\!\!\!\alpha < 1$ be the values of f that are not 0 or 1. We show that $K = 0$. **If** not, consider the set $T = \{s \mid f(s) = \alpha_1\}$. Let Y be the set of all states in which it is optimal for the speaker to make a statement from *T,* that is, $Y = \{x \mid \alpha(f, x) = \alpha_1\}$.

If the probability of Y ∩ A is at least that of Y ∩ R, then consider $j+$ which is a revision of j :

$$j+(8) = \alpha_2 \text{ for all } 8 \in T \text{ and } j+(8) = j(8) \text{ for } 8 \notin T.$$

Thus, $a(j+,x) = \alpha_2$ for $x \in Y$ and $a(j+,x) = \alpha(f,x)$ for $x \notin Y$. It follows that $m(j+) \leq m(j)$.

If the probability of Y ∩ A is at most that of Y ∩ R, then consider r: which is a revision of j:

$$j-(8) = 0 \text{ for all } 8 \in T \text{ and } j-(8) = j(8) \text{ for } 8 \notin T.$$

Thus, $a(j-,x) = 0$ for $x \in Y$ and $a(j-,x) = \alpha(f,x)$ for $x \notin Y$. It follows that $m(j-) \leq m(j)$.

The number of non-integer values used by either $j+$ or r: is reduced by 1, which contradiets the assumption that j uses the the minimal number of non-integer values.

(2) Let f' be a persuasion rule such that $m(f') < \text{infj } m(j) + \varepsilon/2$. Let n be an integer such that $l/n < \varepsilon/2$. Let f'' be the persuasion rule defined by $f''(s) = \max\{\text{m/n I } mln \leq f'(s)\}$. Obviously $m(f'') < m(f') + l/n$. The persuasion rule f'' involves a finite number of values. By the proof of 4.1 there is a deterministic persuasion rule j^* with $m(j^*) \leq m(j'')$. Thus, $m(j^*) < m(f') + c/2 < \text{inf} I\, m(j) + \text{s}$. Q.E.D.

Example 4.2 (Example 2.1 revisited: a Solution). We return now to example 2.1 and show that no persuasion rule induces a probability of error less than $\frac{4}{32}$. Consider an optimal persuasion rule that is deterministic. Thus, μ_x is either 0 or 1 for any state x. Ey the L-principle, $\mu_{(1,1,1,0,0)} + \mu_{(1,1,0,0,0)} + \mu_{(1,0,1,0,0)} + \mu_{(0,1,1,0,0)} \geq 1$ and similar inequalities hold for any of the other 9 states in which exactly three aspects support the speaker. Summing up over these 10 inequalities yields

$$\sum_{n(x)=3} \mu_x + 3 \sum_{n(x)=2} u\text{«> } 10.$$

Using the fact that μ_x is either 0 or 1 implies that $\sum_{n(x)=3} \mu_x + \sum_{n(x)=2} \mu_x \geq 4$ and thus $\sum_x p_x \mu_x \geq \frac{4}{32}$.

Let us now describe an optimal persuasion rule for this case. Partition the set of random variables into t he two sets {I, 2, 3} and {4, 5}. The listener is persuaded only if the speaker can show that two random variables from the same cell of the partition support him. In states in which there are at least three random variables in favor of the speaker, at least two of them must belong to the same cell and, thus, the speaker is justifiably able to persuade the listener. However, in the four states in which exactly two

random variables belonging to the same cell support the speaker's position, the speaker is able to persuade the listener even though he should not be able to. Thus, the probability of error under this persuasion rule is $\frac{4}{32}$.

This persuasion rule seems to be attractive when the partition of the random variables is prominent. For example, if the random variables are associated with Alice, Beth, Christina, Dan, and Edward, they can naturally be divided into two groups by gender. Given the constraint that the speaker cannot refer to more than two individuals, we have found an optimal persuasion rule whereby referring to two individuals of the same gender is more persuasive than referring to two individuals of different genders.

Example 4.3 (Persuading someone that the median is above the expected value). A speaker wishes to persuade the listener that the median of the values of three independent random variables uniformly distributed over the interval [0,1] is above 0.5. The speaker can reveal the value of only one of the three random variables. Is it more persuasive to present a random variabie with a realization of 0.9 or one with a realization of 0.67

Formally, let X = [0,1] x [0,1] x [0,1] with a uniform distribution and A = {(Xl, X2, X3) I two of the values are above 0.5}. Let $x_i = t_i$ denote the statement "the realization of the variabIe Xi is t_i" and $S(t_1, t2, t3)$ = {Xl = t_i, X2 = t2, X3 = t3}' **In** other words (X, S) is vectoric.

The persuasion rule according to which the listener is persuaded only by the statement $Xl = t_i$ where $t_i > \frac{1}{2}$ yields a probability of error of $\frac{1}{4}$. We will employ the L-principle to show that this persuasion rule is optima!.

Note that the space X is isomorphic to the probabilistic space Y x Z with a uniform distribution, where Y = $[0, \frac{1}{2}]$ x $[0, \frac{1}{2}]$ x $[0, \frac{1}{2}]$ and Z = {-I, I} x {-I, I} x {-I, I}, by identifying a pair (y, z) with $X = (\frac{1}{2} + YiZi)i=1,2,3$.

As aresuit, every $(y, (1, 1, -1)) \in A$ is part of an L with $(y, (-1, 1, -1)) \in$ Rand $(y, (1, -1, -1)) \in R$.

Thus we obtain the following inequalities:

$$\mu_{(y,(1,1,-1)} + \mu_{(y,(-1,1,-1))} + \mu_{(y,(1,-1,-1))} \geq 1$$
$$\mu_{(y,(1,-1,1)} + \mu_{(y,(-1,-1,1))} + \mu_{(y,(1,-1,-1))} \geq 1$$
$$\mu_{(y,(-1,1,1)} + \mu_{(y,(-1,1,-1))} + \mu_{(y,(-1,-1,1))} \geq 1.$$

Hence

$$\mu_{(y,(1,1,-1))} + \mu_{(y,(1,-1,1))} + \mu_{(y,(-1,1,1))} + 2\mu_{(y,(-1,1,-1))} + 2\mu_{(y,(1,-1,-1))} + 2\mu_{(y,(-1,-1,1))} \geq 3.$$

For deterministic persuasion rules it must be that at least two of the variables $\mu_{(y,z)}$ take the value 1 and, thus, for all y, we have $\sum_z p_{(y,z)}\mu_{(y,z)} \geq \frac{2}{8} = \frac{1}{4}$. **If** there exists a persuasion rule that yields an error probability

strictly less than $\frac{1}{4}$, then by 4.1(ii) there is also a deterministic persuasion rule that yields an error probability less than $\frac{1}{4}$. Thus, the persuasion rule described above (which yields an error probability of exactly $\frac{1}{4}$) is optima!.

5 A procedure for finding an optimal persuasion rule

We are now able to prove a proposition that reduces the task of finding an optimal persuasion rule to a simple optimization problem.

Proposition 5.1. Let $(X, A, p, \sigma\rangle$ be a finite persuasion problem. Let $(\mu_x^*)_{x\in X}$ be a solution to the optimization problem

$$\min_{\{\mu_x\}_{x\in X}} \sum_{x\in X} p_x\mu_x \text{ s.t. } \mu_x \in \{0, 1\} \text{ for all } x \in X, \text{ and}$$

$$\sum_{t\in\{x\}\cup T} \mu_t \geq 1 \text{ for any minimal } L, (x, T).$$

Then there is an optimal persuasion rule that induces the probabilities of errors $(\mu_x^*)_{x\in X}$.

Proof. By 4.1 we can restriet ourselves to deterministic mechanisms. By 3.2 any persuasion rule satisfies the constraints (regarding the L's), so it is sufheient to construct a persuasion rule f that induces the optimal error probabilities vector $(\mu_x^*)_{x\in X}$.

Define $f(s) = 1$ for any signal s such that there exist $x \in A$ with $s \in \sigma(x)$ so that $\mu_x^* = 0$ and $\mu_y^* = 1$ for all $Y \in R$ with $s \in \sigma(y)$. Define $f(s) = 0$ for any other signal s.

It is sufficient to show that for all x, the induced probability $\mu_x(f) \leq \mu_x^*$. Let $x \in A$ and $\mu_x^* = 0$. There is a statement $Sx \in \sigma(x)$ so that $\mu_y^* = 1$ for all $Y \in R$ such that $Sx \in \sigma(y)$. Otherwise, there is an L, (x, T) such that $\sum_{t\in\{x\}\cup T} \mu_t^* = 0$. Thus $f(sx) = 1$ and $\mu_x(f) = 0$

Let $x \in$ Rand $\mu_x^* = 0$. Then there is no $s \in \sigma(x)$ such that $f(s) = 1$ and thus $\alpha(f, x) = 0$ and $\mu_x(f) = \mu_x^*$. Q.E.D.

6 Ex-post optimality

So far we have assumed that the listener is committed to a persuasion rule. In what follows, we address the question of whether the listener's optimal persuasion rule is one that he would indeed follow were he able to reconsider his commitment aftel' the speaker has made his statement.

To motivate this analysis consider the following example.

Example 6.1. The listener wishes to choose a guest for a TV news program. He is looking for a person with strong views about the issues of the day. There is a potential candidate who the listener knows is one of four

types: "hawk" *(H),* "dove" *(D),* a "pretender" (M) who can pretend to be either a hawk or a dove, or "ignorant" *(J).* The listener is not interested in the candidate's political views, but only in whether he has clear views one way or the other, i.e., if he is type *H* or *D.* The probabilities of the types are $p(H) = p(D) = 0.2$ and $p(M) = p(J) = 0.3$.

The listener can interview the candidate, aftel' which he must decide whether or not to invite him onto the show. During the interview the listener plans to ask the speaker to make a statement regarding his views on cutrent issues. Assume that apart from remaining silent (action 0), type *H* can make only the statement *h; D* can make only the statement *d;* and M can make either statement *hor d.* Type I can only remain silent. Thus, $\sigma(H) = \{h,O\}$, $\sigma(D) = \{d,O\}$, $\sigma(M) = \{h,d,O\}$, and $\sigma(I) = \{O\}$.

A "naïve" approach to this problem is the following: Given the statement *s,* the listener excludes the types that cannot make the statement *s* and makes the optimal decision given the probabilities. For example, the message *d* excludes types land Hand therefore implies that the conditional probability that the speaker is of type *D* is 0.4. The listener thus rejects the speaker. This approach yields a probability of error of 0.4.

Suppose that the listener can commit to how he will respond to the speaker's statement. It is easy to see that, in this example, the listener can reduce the probability of error to 0.3. The best persuasion rule is to invite the speaker to the show if and only if he makes the statement *d* or *h.* (This avoids the possibility that *I* is invited to the show but leaves the possibility that, in addition to Hand *D, M* might be invited.)

Assume now that the listener is released from his commitment once a statement has been made. If he believes that M's strategy is Lu uLLer *d,* then the listener, upon hearing the statement *d,* should attribute a higher probability to the possibility that he is facing *M* than to the possibility that he is facing *D.* Therefore, in this case he should not follow the optimal persuasion rule and should reject the speaker if he makes the statement *d.* H, however, the listener believes that M randomizes with equal probability between uttering *d* and *h,* then the listener, upon hearing the message *d (h),* should attribute the probability $\frac{4}{7}$ to the possibility that he is facing type *D (H)* and, thus, should not deviate from the optimal persuasion rule.

Note that the ex-post optimality of the optimal persuasion rule in this example hinges on the knife-edge condition that the speaker of type *M* randomizes with equal probability between hand *d.* This observation hints at the possibility that a persuasion problem might exist in which the listener's optimal persuasion rule is not ex-post optima!. However, as the analysis below demonstrates, this is never the case for finite persuasion problems .

For a given persuasion problem $(X, A, p, \sigma\rangle$, consider the corresponding extensive persuasion game $\Gamma(X, A, p, o)$. First, nature chooses the state

according to p; the speaker is then informed of the state x and makes a statement from the set $\sigma(x)$; and finally, aftel' hearing the speaker's statement, the listener chooses between a and τ. The payoff for the speaker is 1 if the listener takes the action a and 0 otherwise. The payoff for the listener is 1 if $x \in A$ and the action a is taken or if $x \in \text{Rand}$ the action r is taken, and 0 otherwise. We say that a certain persuasion rule f is *credsble* if there exists a sequential equilibrium of $\Gamma(X, A, p, \sigma)$, such that the listener's strategy is f.

Example 6.2 (Example 3.3 revisited). The optimal persuasion rule described above is credible. The speaker's strategy of arguing in state *(tl, t2)* that $x1 = t_1$ if $t_1 \geq t_2$ and that $x2 = t_1$ if $t_1 > t_2$ is optima!. The set of types that use the argument $x1 = t_1$ is $\{(t1, x2) \mid x2 \leq t_1\}$. Conditionalon this set, the probability that *(tl, X2)* is in A is greater than $\frac{1}{2}$ if and only if $t_1 > \frac{2}{3}$ and is less than $\frac{1}{2}$ if and only if $t_1 < \frac{2}{3}$.

Proposition 6.3. If the persuasion problem is finite, then any optimal persuasion rule is credible.

This proposition follows from solving the auxiliary problem presented in the next section.

Comment. The problem studied here can be viewed as a special case of a leader-follower problem in which the leader can commit to his future moves. As is weil known, it is generally not true that the solution to such an optimization problem is credible. We are not aware, however, of any general theorem or principle that addresses this issue and that can explain why it is the case that in our model the listener's optimal strategy is credible. This question remains for future research.

We should emphasize, however, that 6.3 does not hold in case the listener has three actions, the speaker holds a fixed ordering over the actions, and the listener's preferences depend on the state. Consider, for example, the case in which the set of states is $X = \{1, 2\}$, the probability measure over X is $p_1 = 0.4$ and $p_2 = 0.6$, the signal function is $\sigma(1) = \{1\}$, $\sigma(2) = \{1, 2\}$, and the listener's set of actions is $\{1, 2, 3\}$. The speaker always prefers 1 over 2 and 2 over 3 and the listener's utility from the state x and action a is $u(1, 1) = u(2, 2) = 1$, $u(1, 2) = u(2, 1) = -1$, and $u(1, 3) = u(2, 3) = 0$. The optimal persuasion rule for the listener is to respond to signal 2 with action 2 and to signall with action 3. However, once he observes signal 1 it is better for the listener to take action 1.

7 The bridges problem

A group of individuals is partitioned into a finite number of types, which are members of a set X. The mass of type x is p_x. Let S be a set of bridges

spanning a river. The individuals are loeated on one side of the river and would like to eross to the other side. Individuals of type $x \in X$ ean use only the bridges in the set $\sigma(x) \neq \emptyset$. The set X is partitioned into two subsets, A whose members are welcome on the ot her side and R whose members are not. A deeision maker has to deeide, for eaeh bridge, the probability that that bridge will be open. The deeision maker eannot diseriminate between the individuals in A and R. Eaeh individual of type x ehooses a bridge in $\sigma(x)$ with the highest probability of being open from among the ones he ean use. The deeision maker's objeetive is to maximize the "net flow", i.e., the differenee in size between the group of type A's and the group of type R's erossing the river.

A bridge poliey determines the probability with whieh eaeh bridge is open. A bridge poliey is eredible if there exists an assignment of types to bridges whereby: *(i)* eaeh type is assigned only to a bridge he ean use, *(ii)* within the set of bridges he ean use, eaeh type is assigned only to bridges with the highest probability of being open, and *(iii)* the mass of types in A who are assigned to a bridge that is open (closed) with strietly positive probability is at least as high (low) as the mass of types in R who are assigned to that bridge. We show that any optimal bridge poliey is eredible.

Formally, a *brulçe policy* is a function $O : S \to [0, 1]$ with the interpretation that O(s) is the probability that bridge s is open. Let $\alpha(O, x) = \max\{O(s) \mid s \in \sigma(x)\}$, that is the maximal probability of erossing the bridges that type x ean aehieve given the bridge poliey O. Let $N(O) = \sum_{x \in A} p_x \alpha(O, x) - \sum_{x \in R} p_x \alpha(O, x)$ be ealled the net flow. A bridge poliey is *optimal* ifiL maxiruizes $N(O)$. Given a bridge policy O, a *mtionalfeasible bridqe assignment* β is a function that assigns to eaeh type x a probability measure on $\sigma(x)$, sueh that $\beta(x)(s) > 0$ only for values of s that maximize $O(s)$ in $\sigma(x)$. Given an assignment β, the *net assignment* to bridge s is $n(s, \beta) = \sum_{x \in A} p_x \beta(x)(s) - \sum_{x \in R} p_x \beta(x)(s)$. A bridge poliey O is *credible* if there is a rational feasible assignment β sueh that for every s, $O(s) > 0$ implies $n(s, \beta) \geq 0$ and $O(s) < 1$ implies $n(s, \beta) \leq 0$.

Claim 7.1. All optimal bridge polieies are eredible.

Proof. Let O^* be an optimal bridge poliey. For any assignment β, let

$$\delta(\beta) = \sum_{s \in \{s \mid n(s,\beta) < O\}} \text{In}(s, \beta) O^*(s) + \sum_{s \in \{s \mid n(s,\beta) > O\}} n(s, \beta)(1 - O^*(s)).$$

Let β^* be a minimizer of $\delta(\beta)$ over all rational feasible assignments. We show that $\delta(\beta^*) = 0$ and thus for all s sueh that $O^*(s) > 0$ we have $n(s, \beta) \geq 0$ and for all sueh s that $O^*(s) < 1$ we have $n(s, \beta) \leq 0$.

Assume, for the purpose of contradiction, that $\delta(\beta^*) > 0$. Assume that there is a bridge S for which $0^*(s) > 0$ and $n(s,;3^*) < 0$ (an analogous argument applies to the case in which there is a bridge s for which $0^*(s) < 1$ and $n(s,;3^*) > 0$).

Let α be the minimum of $O^*(s)$ over $\{s \in O^*(s) > 0$ and $n(s,;3^*) < 0\}$. Let $S(\alpha) = \{s \in O^*(s) = \alpha\}$. Let $X(\alpha) = \{x \in ;3^*(x)(s) > 0$ for a bridge s such that $s \in S(\alpha)\}$, that is, $X(\alpha)$ is the set of types who are assigned by;3* to the bridges whose probability of being open is α. Note that types in $X(\alpha)$ cannot do better than trying to cross a bridge in $S(\alpha)$ and are indifferent between all bridges in $S(\alpha)$. Let $Sa = \{s \in S(\alpha) \in n(s,;3^*) < 0\}$. The set Sa is not empty and contains all bridges that are open with probability α and for which the net assignment is negative.

Let $Y1, \ldots, in$ be the longest sequence of distinct bridges in $S(\alpha) - Sa$ such that for every Yt,

(i) $n(Yt, \beta^*) = 0$

(ii) there exist $x \in R$ and $Yo \in Sa \cup \{Y1, \ldots, y_{t-1}\}$ such that $\beta^*(x)(yo) > 0$ and $Yt \in \sigma(x)$.

In other words, under β^* each Yt is a bridge with a zero net transfer such that there is a positive mass of types in R that can cross Yt and is assigned by β^* either to cross a bridge that precedes Yt in the sequence or to cross a bridge in Sa.

Denote $Z = Sa \cup \{Y1, \ldots, Yr\}$. There are two possibilities:

(i) There is no $s \in S(\alpha) - Z$, $x \in R$, and $z \in Z$ such that $s \in \sigma(x)$ and $\beta^*(x)(z) > 0$. That is, there is no bridge s outside Z that is opened with probability α and that can be crossed by a type in R who can cross the river with probability α. The net transfer in Z is negative. Reducing the probability of transfer to all bridges in Z will increase the total net flow, thus violating the optimally of 0*.

(ii) There is $s \in S(\alpha) - Z$, $x \in R$, and $z \in Z$ such that $s \in \sigma(x)$ and $;3^*(x)(z) > 0$. By the definition of $(Y1, \ldots, Yr)$ it must be that $n(s, \beta^*) > 0$. It follows that there are sequences of distinct bridges $SO,Sl, \ldots, SK = s$ and types $io, \ldots, i_{K-1} \in R$ such that $sa \in Sa$, $;3^*(ik)(Sk) > 0$, and $Sk+1 \in \sigma(i_k)$ (for $k = 0, \ldots, K - 1$). This allows us to construct a new rational assignment β by shifting a positive mass of types in R from sa to $s1$, from $s1$ to $s2$, and so on, such that $\delta(\beta) < \delta(\beta^*)$. Formally, let ε be a positive number such that for $k = 0, \ldots, K - 1$ we have $c < ;3^*(ik)(Sk)$, $c < n(sK,;3^*)$, and $c < In(so,;3^*)I$. Define;3 as an assignment that is obtained from;3* by successively shifting to $Sk+1$ a mass c of individuals of type ik assigned

by;3* to cross s_k. For all bridges with the exception of s_a and s_K we have $n(s,;3) = n(s,;3^*)$. Furthermore, $n(s_K,;3) = n(s_K,;3^*) - c > 0$ and $n(s_o,;3) = n(s_o, \beta^*) + c < 0$. Thus, $\delta(\beta) = \delta(\beta^*) - \alpha\varepsilon - (1 - \alpha)\varepsilon$, contradicting the choice of β^*.

Thus, it follows that there exists a rational feasible assignment with nonnegative net flow on all open bridges and nonpositive net flow on all closed bridges. Q.E.D.

8 Concluding remarks

This paper has attempted to make a modest contribution to the growing literature linking economie theory to linguistics. Our purpose is not to suggest a general theory for the pragmatics of persuasion but rather to demonstrate a rationale for inferences in persuasion situations.

One of our main findings is that any optimal persuasion rule is also ex-post optimal. It is quite rare that in a principal-agent problem the optimal incentive scheme is one that the principal would wish to obey even aftel' the agent has made his move. The bridge problem described in 7 provides an example of a principal-agent problem that in fact does have this property. The problem discussed in (Glazer and Rubinstein, 2004) is shown there to have this property as weil. The generalizability of this result is still an open question.

Our work is related to several areas of research in linguistics and economics. **In** the linguistics literature, our paper belongs to the emerging field that tries to explain pragmatic rules by employing game theoretical methods. **In** our approach, pragmatic rules determine a game between the participants in the discourse. Whatever the process that created these rules, it is of interest to compare them with the rules that would have been chosen by a rational designer seeking to maximize the functionality of the discourse. Such an approach is suggested in (Glazer and Rubinstein, 2001, 2004) and discussed in (Rubinstein, 2000). A recent collection of articles in (Benz et al., 2006) presents various ideas that explain pragmatics phenomena using game theoretical tools.

Within the economie literature our paper relates to two areas of research.

The first investigates sender-receiver games (see Crawford and Sobel, 1982) in which one agent (the sender) sends a costless message to the other (the receiver). The receiver cannot verify any ofthe information sent by the sender and the interests of the sender and the receiver do not necessarily coincide. The typical question in this literature is whether an informative sequential equilibrium exists.

The second (and closer) area of research studies models where a principal tries to elicit verifiabie information from the agent(s). The agent however

can choose which pieces of information to convey. Among the early papers on this topic are (Townsend, 1979), (Green and Laffont, 1986), and (Milgrom and Roberts, 1986), and among the more recent are (Bull and Watson, 2004), (Deneekere and Severinov, 2003), (Fishman and Hagerty, 1990), (Forges and Koessier, 2005), (Lipman and Seppi, 1995), and (Shin, 1994).

References

Benz, A., Jäger, G. & van Rooij, R., eds. (2006). *Game Theoru and Pragmatics.* Palgrave Macmillan, Basingstoke.

Bull, J. & Watson, J. (2004). Evidence Disclosure and Verifiability. *Journol of Economie Theoru, 118(1):1-31.*

Crawford, V.P. & Sobel, J. (1982). Strategie Information Transmission. *Econometrica, 50(6):1431-1451.*

Deneekere, R. & Severinov, S. (2003). Mechanism Design and Communication Costs. Working paper, Fuqua School of Business, Duke University.

Fishman, M.J. & Hagerty, K.M. (1990). The Optimal Amount of Discretion to Allow in Disclosure. *Quarterly Journalof Economics, 105:427-444.*

Forges, F. & Koessier, F. (2005). Communication Equilibria with Partially Verifiabie Types. *Journalof Mathematical Economics, 41(7):793-811.*

Glazer, J. & Rubinstein, A. (2001). Debates and Decisions: On a Rationale of Argumentation Rules. *Games and Economie Behaoior, 36(2):158-173.*

Glazer, J. & Rubinstein, A. (2004). On Optimal Rules ofPersuasion. *Econometrica,* 72(6):1715-1736.

Green, J.R. & Laffont, J.-J. (1986). Partially Verifiabie Information and Mechanism Design. *Review of Economie Studies, 53:447-456.*

Grice, H.P. (1989). *Studies in the Wayof WOTds.* Harvard University Press, Cambridge, Mass.

Lipman, B.L. & Seppi, D.J. (1995). Robust Inference in Communication Games with Partial Provability. *Journalof Economie Theoru, 66(2):370-*405.

Milgrom, P. & Roberts, J. (1986). Relying on the Information of Interested Parties. *Rand Journalof Economics, 17:18-32.*

Rubinstein, A. (2000). *Economics and Language.* Cambridge University Press, Cambridge.

Shin, H.S. (1994). The Burden of Proof in a Game of Persuasion. *Journol of Economie Theory, 64(1):253-264.*

Townsend, RM. (1979). Optimal Contracts and Competitive Markets with Costly State Verification. *Journalof Economie Theoru, 21(2):265-293.*

On Glazer and Rubinstein on Persuasion

Boudewijn de Bruin

Faculty of Philosophy
University of Groningen
9712 GL Groningen, The Netherlands
b.p.de.bruin@rug.nl

> These were her internal persuasions: "Old fashioned notions; country hospitality; we do not profess to give dinners; few people in Bath do; Lady Alicia never does; did not even ask her own sister's family, though they were here a month: and I dare say it would be very inconvenient to Mrs Musgrove; put her quite out of her way. I am sure she would rather not come; she cannot feel easy with us. I will ask them al! for an evening; that will be much better; that will be a novelty and a treat. They have not seen two such drawing rooms before. They will be delighted to come to-morrow evening. It shal! be a regular party, smalI, but most elegant."
>
> -Jane Austen, *Persuasion*

Jacob Glazer and Ariel Rubinstein profter an exciting new approach to analyze persuasion. Perhaps even without being aware of it, and at least not acknowledged in the bibliography, their paper addresses questions that argumentation theorists, logicians, and cognitive and social psychologists have been interested in since Aristotle's *Rhetoric*. Traditionally, argumentation was thought of as an activity involving knowiedge, beliefs, opinions, and it was contrasted with bargaining, negotiation and other strategie activities involving coereion, threats, deception, and what have you. More recently, however, several theorists have argued that strict boundaries are conceptually indefensible and undesirable methodologically, separating as they do researchers who would more fruitfully combine efforts. Katia Sycara, for instanee, writes that "persuasive argumentation lies at the heart of negotiation" (Sycara, 1990), identifying various argumentation and negotiation techniques on the basis of careful empirical research on labor organizations. Simon Parsons, Catles Sierra, and Nick Jennings, by contrast, develop models of argumentation-based negotiation (Parsons et al., 1998) with a high level of logical formality. Chris Provis, to mention a third, perhaps more skeptical representative, gives a systematic account of the distinction between argumentation and negotiation suggesting to locate persuasion right

Krzysztof R. Apt, Robert van Rooij (eds.). *New Perspectives on Games and Interaction.* Texts in Logic and Games 4, Amsterdam University Press 2008, pp. 141-150.

in the middle (Provis, 2004).1 Glazer and Rubinstein's work enriches this literature with an analysis of persuasion. Using machinery from a formal theory of negotiation *par excellence,* economie theory, they develop a model of persuasion problems in which a speaker desires a listener to perform a certain action. Informing the listener about the state of the world is the only thing the speaker can do, but he can do it in many ways. By strategically making a statement that maximizes the likelihood that the listener decides to perform the action, the speaker exploits a peculiar feature of the model, namely, that the speaker does not need to teil the listener the *whole* truth; the truth alone suffices. Persuasion, for Glazer and Rubinstein, is telling the truth strategically, and phrased in the framework of a novel methodology this new approach merits close attention.

A speaker, a listener, and a tuple (X, A, p, σ) with X a set of worlds (not necessarily finite), $A \subseteq X$, p a probability measure over X, and $\sigma: X \to S$ a function mapping worlds to sets of proposition (symbols?) in S, that is all there is to a "persuasion problem." There is a certain action that the speaker wants the listener to perform. The listener wants to perform it just in case the actual world is an element of A. The speaker, by contrast, wants the listener to perform the action even if the actual world is a mernber of the complement R of A. Since the listener does not have full information about the world but only "initial beliefs ... given by a probability measure pover X" (Glazer and Rubenstein, 2008, p. 123), he is partly dependent on the speaker who has full knowledge of the world. Vet the speaker is under no obligation to report his full knowledge to the listener. The rules fixed by σ allow the speaker to make all of the statements contained in $\sigma(x)$, if x is the acl.ual world. Clazer and Rubinstein wriLe that "Lhe rneaning of 'rnaking statement s' is to present proof that the event $\sigma^{-1}(s) = \{x | s \in \sigma(x)\}$ has occurred" (page 5). Strategically picking such an s characterizes persuasion: an s with a large $\sigma^{-1}(s)$ wiII generally do better than a small one. The "persuasion function" $f: S \to [0,1]$, moreover, is intended to capture "the speaker's beliefs about how the listener will interpret each of his possible statements" (p. 124). The statement $f(s) = q$ means that "following a statement s, there is a probability of q that the listener will be 'persuaded' and choose ... the speaker's favored action" (p. 124). The speaker solves the maximization problem

$$\max_{s \in \sigma(x)} f(s),$$

where x is the actual world.é From the perspective of the listener, if the speaker makes a statement t such that $f(t) = \max_{s \in \sigma(x)} f(s)$ there is a probability $\mu_x(f)$ that by using f he makes an error at x to perform an

1 [owe mueh to Chris Provis' exposition in (Provis, 2004).

2 [f $\sigma(x)$ is infinite, the supremum of the expression can be approached.

action while $x \notin A$, or not to perform an action at x while $x \in A$. These probabilities are given by

$$\mu_x(f) = \begin{cases} 1 - \max_{s \in \sigma(x)} t_s & \text{if } x \in A \\ \max_{s \in \sigma(x)} t_s & \text{otherwise.} \end{cases}$$

The listener chooses a persuasion rule that solves the minimization problem

$$\min_{f:\, S \to [0,1]} \sum_{x \in X} p(x)\mu_x(f).$$

given his beliefs p.[3]

If this is a model, what does it model? Glazer and Rubinstein give several examples of persuasion problems. They bear names suggesting rather concrete applications ("The Majority of the Facts Support My Position," "I Have Outperformed the Population Average,") as weil as technical, perhaps contrived, ones ("Persuading Someone that the Median is Above the Expected Value"). In the first example, the speaker tosses a coin five times in a row, and wants the listener to perform a certain action the listener wants to perform just in case the coin landed heads at least three times. If the persuasion problem is such that the speaker can show how the coin landed in three of the five cases, he will of course succeed in persuading the listener, provided the coin landed heads at least three times. More interestingly, the speaker may only be able to reveal the outcomes of two coin tosses. Given that the listener only wants to perform the action in case the coin landed heads at least three times, there is always a risk involved in acting on the basis of the information the speaker provides to him. The listener may consider the persuasion rule according to which he performs the action just in case the speaker demonstrates that the coin landed heads twice. Among the total of 32 possible outcomes of the experiment (HHHHH, THHHH, and so on), there are 10 in which the coin landed heads twice, not thrice, and this makes the error probability of this rule $\frac{10}{32}$. The listener can improve if he adopts the persuasion rule to accept only if the coin landed heads twice in a row. This persuasion rules has error probability $\frac{5}{32}$:

> An error in favor of the speaker will occur in the four states in which exactly two neighboring random variables [two successive coin tosses] support the speaker's position and in the state [HTHTH] in which the speaker will not be able to persuade the listener to support him even though he should. (p. 125)

The example reveals a number of epistemic presuppositions behind the model Glazer and Rubinstein propose. Speaker and listener, for instance,

3 Lemma 1 says that whatever the cardinality of S, there is always a solution to this minimization problem if the persuasion problem is finite.

have to know exactly what the rules of the game are. If the speaker does not know that he can reveal the outcomes of at most two successive coin tosses he will not consider the sequence in which the coin landed **HTHTH** as a problematic situation for him, and if the listener believes that the speaker may so be misinformed about the structure of the game he will also evaluate differently what is the optimal persuasion rule. One might even conjecture that as long as there is no *common* knowledge of the structure of the persuasion problem, game play is impossible. In addition, for the listener to calculate the error probabilities of the various persuasion rules he has to agree on the probability distribution of the relevant random variables. In the description of the formal model, Rubinstein and Glazer to that end insert a probability measure pover possible worlds with the initial beliefs of the listener as intended interpretation. The example, however, suggests that this probability is rather derived from the objective characteristics of set of possible worlds X, available to listener and speaker alike. Not only does the speaker, then, know what the actual world is in a situation in which the listener only has probabilistic beliefs concerning that issue, he also knows exactly what the listener believes about the world.

Nor is this all. While strictly speaking no condition of possibility for an application of the model, Glazer and Rubinstein suggest that the speaker not only knows the listener's beliefs, but also the listener's prospective choice of strategy. Given the fact that the speaker has access to the listener's beliefs p, it is routine for him to calculate an optimal persuasion rule, and assuming that the listener is in some sense rational, the speaker is quite justified in believing that the listener will choose that rule. There is an interesting proviso, Lhough, Ior Lhe meaningfulness of Lhe definiLion of error probability depends not only on the fact that p expresses the listener's probabilistic beliefs concerning possible worlds, but also on the fact that the listener assumes that the speaker wants to maximize the likelihood that the listener perform the action. If the speaker did not want so to maximize, the listener would be unwise to build his risk estimation on the basis of the value of the solution to $\max_{s \in \sigma(x)} f(s)$. The speaker, for his derivation ofthe persuasion rule, needs to believe that the listener believes the speaker to be rational.

For the speaker, it is quite deal' how to motivate the rule of rationality embodied in his maximizing the probability of acceptance. If the speaker has non-probabilistic beliefs concerning the persuasion rule f adopted by the listener, the only thing he needs to do is to piek a statement s, and it makes much sense to choose one that maximizes expected acceptance. Conceptions of rationality such as maximin or minimax regret are out of place here. For the listener this may be a bit different. The listener wants to piek a persuasion rule $f: S \rightarrow [0,1]$, and the most direct constraint is that f favors assigning high probability in cases in which the actual world is

an A-world, and low probability in cases in which it is an R-world. Without indication about how the elements from S relate to the A-ing or *R-ing* of the actual world, there is only one thing the listener could use to determine his strategy: his beliefs. If he believes with probability one that the world is in R, areasonabie persuasion rule assigns the value of zero (non-performance) to any statement made by the speaker. But the speaker knows what the actual world is, and the listener knows that the speaker knows it, so if the speaker makes a statement s with $\sigma^{-1}(s) \subset A$ to the effect that the world is definitely an A-world, then what should the listener do? This is a deal' case of belief revision the model may not fully capture by assuming that the probabilities are objectively induced by the random variables determining X. For a rational choice of a persuasion rule the listener may not have enough information about the relation between the state of the world and the statement the speaker makes.

A *conditional* statement can be made, though. If the listener believes that the speaker knows what persuasion rule f the listener chooses, and the listener believes that the speaker is rational, then the listener believes that in his calculation of the optimal persuasion rule he can use the $\mu_x(f)$ to assess the risk of making errors and solve min!: $_{S\to[0,1]} \sum_{x\in X} p(x)\mu_x(f)$. Such a conditional statement, however, may delineate the applicability of the model in ways analogous to what we learn from the epistemic characterization of the Nash equilibrium (Aumann and Brandenburger, 1995). To constitute a Nash equilibrium, knowledge of strategies is presupposed. If I know what you are playing, and I am rational, and if you know what I am playing, and you are rational, then we will end up in a Nash equilibrium. Such assurnptions are HoL always problernal.ic, Ior sure, but Lo justify making them requires in any case additional argumentation about, for instance, evolutionary (learning) mechanisrns or repeated game play. As the turn to iterative solution concepts constitutes to some extent an answer to the epistemic problerns with the Nash equilibrium, it may be interesting to investigate whether the model Glazer and Rubinstein put forward can similarly turn into the direction of common knowledge of game structure and rationality, especially if this can be accomplished in an extensive game framework. For consider the following dialog (election time in Italy):

POLITICIAN: Vote for me.

CITIZEN: Why?

POLITICIAN: If you vote for me, 1'11 create one million new jobs.

CITIZEN: That's unpersuasive.

POLITICIAN: If you vote for me, 1'11 light for your freedom.

CITIZEN: Persuaded.

This dialog, due to Isabella Poggi, illustrates a theory of persuasion in terrns of framing developed by Frederic Schick, among ot hers (Poggi, 2005; Schick, 1988). While argumentation is about beliefs, and negotiation and bargaining are about desires and interests, persuasion, for Schick, involves the framing of options. A persuader persuades a persuadee by describing in novel and attractive terrns an action the persuadee found unattractive under previous description:

> We [may] want something under one description and... not want it under another. We may even want a proposition true and want a coreportive proposition false ...
>
> Persuasion is the attempt to change a person's understanding of something, to get him to see it in some way that prompts him to act as he would not have done. (Schick, 1988, p. 368)

At first sight it may be too much to ask Rubinstein and Glazer to incorporate this insight, if an insight it is, in their formalism. Vet once we closely consider the way they set up the rules of the game, and in particular, the function they assign to function σ, there are in fact two ways to recommend.

The set $\sigma(x)$ contains exactly the statements that the speaker can make if the actual world happens to be *x,* and making a statement *s* amounts to demonstrating that the event $\sigma^{-1}(s) = \{x \text{is } \in \sigma(x)\}$ has occurred. As aresuit, there is no room for the speaker to provide false information, but there is quite some room to provide true information tactically and strategically. A bit informally put, if $\sigma^{-1}(s) = \{x \text{is } \in \sigma(x)\}$ contains many A-states and few R-states, then the speaker has good reasons to make the statement *s,* rather than another statement with less fortunate division between *A* and *R.*[4] From an extensional point of view, it suffices if σ maps worlds to sets of worlds. Propositions, under this extensional perspective, are nothing more than sets of worlds. Extensionally speaking, modeling framing seerns pretty hopeless, though: a glass half full is the same as a glass half empty. From an intensional point of view, however, distinctions can be made between coextensive statements, and it is here that there is room for Glazer and Rubinstein to incorporate framing in their framework. The recipe is this. On the basis of a formal language, a set of statements *S* is defined from which the speaker may choose, and to make this set truly interesting, it has to be larger than $\wp(X)$, the set of all statements possible with respect to *X* in purely extensional terms. To the description of the persuasion problem a relation of extensionality is added over the statements such that $s \equiv t$ iff $\sigma^{-1}(s) = \sigma^{-1}(t)$. Define a preference relation inside the resulting equivalence classes to express the listener's preferences for differing

4 This is rough since it ignores the probabilistic judgments *p* the listener will invoke to calculate his error probability.

descriptions, and make the persuasion rule dependent on the statements in a systematic way described in terrns of the extensionality relation and the preferences of the speaker.

A rather different approach to framing is possible, too, one that is much closer to the actual model Glazer and Rubinstein put forward. The speaker, in the persuasion problem as the authors define it, has a lot of freedom to choose a statement s to give the listener information about the actual world. The only restriction is that the actual world be part of the set of worlds $\sigma^{-1}(s)$. Instead of locating framing in intensionally different but extensionally equivalent such sets, framing can also be modeled fully extensionally. Different statements s and *t*, each with the actual world in their σ inverse image, frame the actual world differently, and one could very weil maintain that when the speaker selects what statement to make in Glazer and Rubinstein's model, he is already engaged in framing decisions. While in the intensional solution to framing the speaker would choose between making a statement in terrns of the morning star and one in terms of the evening star, and opt for the latter because he knows that the listener is a night owl, in the alternative solution the speaker would describe the weather in terms of one of two extensionally different statements such as "the weather is good for sailing" and "the weather is good for kite surfing," depending on whether the listener likes sailing or kite surfing.

The simple substitution of freedom for jobs in the dialog. and of drawing rooms for country hospitality in the quotation from *Persuasion;* is an example of persuasion by framing, however simple or simplistic such a switch may beo The dialog, and Austen's stream of consciousness *avant la letire,* point al, another important aspect of persuasiori, Loo: its temporal and sequenLial character. Argumentation theorists and logicians alike have noticed that the persuasive force one can exercise on ot hers often depends on the order in which one presents one's arguments, offers, opinions, and threats. **In** dialogical logic, for instance, a proponent defends a proposition against an opponent who may attack according to clearly described rules. **In** its early days, dialogicallogic was used to promote intuitionist logic *à la* Heyting, or even to give it firm conceptual grounds. Contemporary dialogical logicians, however, see themselves engaged in building a "Third Way" alternative to syntactic and semantic investigations of logical consequence; or in one word, pragmatics.

To get some feel for the kind of models used here, this is a dialog argument to the effect that $(\varphi \to \psi) \wedge \varphi) \to \psi$ is a tautology, due to Rückert (2001):

PROPONENT: $((\varphi \to \psi) \wedge \varphi) \to \psi$

OPPONENT: Weil, what if $(\varphi \to \psi) \wedge \varphi$?

PROPONENT: 1'll show you ψ in a minute. But wait, if you grant

$(\varphi \rightarrow \psi) \wedge ip$, then I ask you to grant the left conjunct.

OPPONENT: No problem, you get your $ip \rightarrow \psi$

PROPONENT: And what about the right conjunct?

OPPONENT: That one you get, toa, ip.

PROPONENT: Weil, if you say ip, 1 may say φ to question you assuming the implication $ip \rightarrow \psi$.

OPPONENT: Right, I see what you're aiming at: you want me to say ψ, and 1'11 admit that ψ.

PROPONENT: Perfect, that means I have shown you ψ in response to your initial query: *ipse dixisti!*

Glazer and Rubinstein's approach to persuasion is decidedly static as it stands, but I believe that it can be turned dynamic at relatively low costs. A first step to consider is to take the probability distribution p as an expression of the truly subjective beliefs of the listener. This has the advantage that belief revision policies can be described to deal with cases in which the speaker comes up with new information, contradicting the listener's beliefs. **In** general, the listener may stubbornly stick to his p, but in more interesting persuasion problems the listener will revise his beliefs because, as it may be assumed, he knows that, however tactically and strategically the speaker will speak, he will at least speak the truth. **In** a dynamic setting, furthermore, there may be more room for less heavy epistemic assumptions. To put it bluntly, my guess is that once persuasion games are represented as extensive games, common knowledge of game structure and rationality suffices to derive optimal persuasion rules. To my mind, this would constitute an increase in realism.

An additional advantage is that extensive models can also take care of Aristotelian analyses of persuasion. **In** the *Rheioric* Aristotle distinguished three ways in which speakers can persuade their listeners. The rational structure of what the speaker says, the *logos*, first of all contributes to the persuasive force. Then the character of the speaker, his *ethos*, determines how credible and trustworthy the listener will judge the speaker, while, finally, the emotional state of the listener, the *pathos*, plays a role in how a certain speech is received. Compare: a well-organized defense by a lawyer of established reputation in a law court with a serious and objective judge, with: a messy argument by a shabby lawyer directed at a judge involved in the case itself. And Aristotle is still highly popular, even among empirically oriented researchers. Isabella Poggi, for instance, agreeing with Schick about the role of framing in persuasion, sees expressions of rationality, credibility, and emotionality as the modern analogs of Aristotle's tripartite division, and gives them all the force in her theory of persuasion

as hooking the speaker's goals to (higher) goals of the listener. **In** the Italian dialog, for instanee, the speaker's goal that the listener votes for him was, first, hooked to diminishing unemployment. The goal of having a job, however, turned out not to be very important to the listener, and therefore another goal was used, the more general one of freedom, of which the speaker had reason to believe that it would arouse the listener's emotions. At the end, the speaker in fact succeeded persuading (or so the story goes).

Using the suggested extensional way of modeling framing, *pathos* can be captured by the preferenee relations the listener has over various descriptions of the actual world. Speaker's beliefs about such preferences can be included to describe specific persuasion strategies the speaker may wish to follow. The speaker is expected to try to describe the world in a way that makes it most attractive for the listener to perform the action but in order to be able to do that, the speaker needs to have some information concerning the listener's preferences.[5] Assumptions about general human preferences (concerning freedom, recognition, or what have you) make it possible for the speaker to do that without information about the specific listener. *Ethos* is captured by the belief revision policies of the listener. **If** the listener readily revises his beliefs upon hearing statements that contradiet his own opinions, he reveals to trust the speaker showing the character of the speaker as a dependable person are at work. More skeptical belief revision policies, in all kinds of gradations, reveal the speaker's *ethos* to be functioning less than optimally. Extensive games can also model ways in which the speaker iteratively tries out reframing the description of the actual world. He may find out that the listener does not like sailing, so it does noL help hirn Lo describe Lhe world as one that is opt.irnal Ior sailing. **In** several models of persuasion, the listener's preferences play a crucial role. *Logos,* finally, gets modeled once speakers may take clever and less clever steps in iterative persuasion games, and it is especially here that cooperation with game theoretic approaches to logic (of which dialogical logic is only one among many) can be very fruitful (van Benthem, 2007).

References

Aumann, R. & Brandenburger, A. (1995). Epistemic Conditions for Nash Equilibrium. *Econometrica, 63(5):1161-1180.*

van Benthem, .1. (2007). Logic in Games. Lecture notes. Institute for Logic, Language and Computation.

5 Aristotle saw persuasion as directed at judgements rather than at actions, though (Rapp, 2002).

Glazer, J. & Rubenstein, A. (2008). A study in the pragmatics of persuasion: A game theoretical approach. This volume.

Parsons, S., Sierra, C. & Jennings, N. (1998). Agents that Reason and Negotiate by Arguing. *Journal of Logic and Computation, 8(3):261-292.*

Poggi, I. (2005). The Goals of Persuasion. *Pragmatics and Cognition,* 13(2):297-336.

Provis, C. (2004). Negotiation, Persuasion and Argument. *Argumentation,* 18(1):95-112.

Rapp, C. (2002). Aristotle's Rhetoric. **In** Zalta, KN., ed., *The Stanford Encyclopedia of Philosophy.* Summer 2002 edition. http://plato.stanford.edu/archives/sum2002/entries/aristotle-rhetoric/.

Rückert, H. (2001). Why Dialogical Logic? **In** Wansing, H., ed., *Essays on Non-Classical Logic,* Vol. 1 of *Advances in Logic,* pp. 165-185. World Scientific Publishing, New Jersey.

Schick, F. (1988). Coping with Conflict. *The Journal of Philosophy,* 85(7):362-375.

Sycara, K. (1990). Persuasive Argumentation in Negotiation. *Theory and Decision, 28(3):203-242.*

Solution Concepts and Algorithms for Infinite Multiplayer Games

Erich Grädel
Michael Ummels

Mathematische Grundlagen der lnformatik
Rheinisch-Westfälische Technische Hochschule Aachen
52056 Aachen, Germany
{graedel,ummels}@logic.rwth-aachen.de

Abstract

We survey and discuss several solution concepts for infinite turn-based multiplayer games with qualitative (i.e., win-Iose) objectives of the players. These games generalise in a natural way the common model of games in verification which are two-player, zero-sum games with w-regular winning conditions. The generalisation is in two directions: our games may have more than two players, and the objectives of the players need not be completely antagonistic.

The notion of a Nash equilibrium is the classical solution concept in game theory. However, for games that extend over time, in particular for games of infinite duration, Nash equilibria are not always satisfactory as a notion of rational behaviour. We therefore discuss variants of Nash equilibria such as subgame perfect equilibria and secure equilibria. We present criteria for the existence of Nash equilibria and subgame perfect equilibria in the case of arbitrarily many players and for the existence of secure equilibria in the two-player case. In the second part of this paper, we turn to algorithmic questions: For each of the solution concepts that we discuss, we present algorithms that decide the existence of a solution with certain requirements in a game with parity winning conditions. Since arbitrary w-regular winning conditions can be reduced to parity conditions, our algorithms are also applicable to games with arbitrary w-regular winning conditions.

1 Introduction

Infinite games in which two or more players take turns to move a token through a directed graph, tracing out an infinite path, have numerous applications in computer science. The fundamental mathematical questions on such games concern the existence of optimal strategies for the players,

Krzysztof R. Apt, Robert van Rooij (eds.). *New Perspectives on Games and Interaction.* Texts in Logic and Games 4, Amsterdam University Press 2008, pp. 151-178.

the complexity and structural properties of such strategies, and their realisation by efficient algorithms. Which games are determined, in the sense that from each position, one of the players has a winning strategy? How to compute winning positions and optimal strategies? How much knowledge on the past of a play is necessary to determine an optimal next action? Which games are determined by memoryless strategies? And so on.

The case of two-player, zero-sum games with perfect information and ω-regular winning conditions has been extensively studied, since it is the basis of a rich methodology for the synthesis and verification of reactive systems. On the other side, other models of games, and in particular the case of infinite multiplayer games, are less understood and much more complicated than the two-player case.

In this paper we discuss the advantages and disadvantages of several solution concepts for infinite multiplayer games. These are Nash equilibria, subgame perfect equilibria, and secure equilibria. We focus on turn-based games with perfect information and qualitative winning conditions, i.e., for each player, the outcome of a play is either win or lose. The games are not necessarily completely antagonistic, which means that a play may be won by several players or by none of them.

Of course, the world of infinite multiplayer games is much richel' than this class of games, and includes also concurrent games, stochastic games, games with various forms of imperfect or incomplete information, and games with quantitative objectives of the players. However, many of the phenomena that we wish to il!ustrate appeal' already in the setting studied here. To which extent our ideas and solutions can be catried over to other scenarios of infiniLe rnultiplayer gatnes is an interestlng topic of current research.

The outline of this paper is as fol!ows. Aftel' fixing our notation in Section 2, we proceed with the presentation of several solution concepts for infinite multiplayer games in Section 3. For each of the three solution concepts (Nash equilibria, subgame perfect equilibria, and secure equilibria) we discuss, we devise criteria for their existence. **In** particular, we wil! relate the existence of a solution to the determinacy of certain two-player zero-sum games.

In Section 4, we turn to algorithmic questions, where we focus on games with parity winning conditions. We are interested in deciding the existence of a solution with certain requirements on the payoff. For Nash equilibria, it turns out that the problem is NP-complete, in genera!. However, there exists a natural restriction of the problem where the complexity goes down to UPnco-UP (or even P for less complex winning conditions). Unfortunately, for subgame perfect equilibria we can only give an ExpTllVIE upper bound for the complexity of the problem. For secure equilibria, we focus on two-player games. Depending on which requirement we impose on the payoff, we

show that the problem falls into one of the complexity classes **UP** ∩ co-UP, NP, or co-NP.

2 Infinite multiplayer games

We consider here infinite turn-based multiplayer games on graphs with perfect information and qualitative objectives for the players. The definition of such games readily generalises from the two-player case. A game is defined by an arena and by the winning conditions for the players. We usually assume that the winning condition for each player is given by a set of infinite sequences of colours (from a finite set of colours) and that the winning conditions of the players are, a priori, independent.

Definition 2.1. An *infinite (tum-based, qualitative) multiplayeT game* is a tuple $\mathcal{G}$ = (II, *V,* $(V_i)_{i\in\Pi}$, *E, X,* (Wini)iETI) where II is a finite set of *plosjers, (V, E)* is a (finite or infinite) directed graph, $(V_i)_{i\in\Pi}$ is a partition of *V* into the position sets for each player, $X : V \to C$ is a colouring of the position by some set C, which is usually assumed to be finite, and Win, $\subseteq CW$ is the winning condition for player i.

The structure G = $(V,(V_i)_{i\in\Pi},E, X)$ is called the *arena* of *g.* For the sake of simplicity, we assume that $uE := \{v \in V : (u,v) \in E\} \neq 0$ for all *u.* ∈ *V,* i.e., each vertex of G has at least one outgoing edge. We call $\mathcal{G}$ a *zero-sum game* if the sets Win, define a partition of *CW.*

A *play* of $\mathcal{G}$ is an infinite path through the graph *(V, E),* and a *historu* is a finite initial segment of a play. We say that a play *tt* is *won* by player i ∈ II if $\chi(\pi) \in$ Win.. The *payoff* of a play *tt* of $\mathcal{G}$ is the vector payt-r) ∈ {O, 1}TI defined by $\mathrm{pay}(\pi)_i$ = 1 if *tt* is won by player i. A *{pure} strategy of plosier* i *in* $\mathcal{G}$ is a function $\sigma : V^*V_i \to V$ assigning to each sequence *xv* of position ending in a position *v* of player i a next position $\sigma(xv)$ such that $(v,\sigma(xv)) \in E$. We say that a play *tt* = $\pi(0)\pi(1)\ldots$ of $\mathcal{G}$ is *consistent* with a strategy σ of player i if *tt[i]:* + 1) = $\sigma(\pi(0)\ldots\pi(k))$ for all $k < w$ with $\pi(k) \in V_i$. A *strategy profile of* $\mathcal{G}$ is a tuple $(\sigma_i)_{i\in\Pi}$ where *o;* is a strategy of player i.

A strategy σ of player i is called *positional* if σ depends only on the current vertex, i.e., if $\sigma(xv) = \sigma(v)$ for all $x \in V^*$ and $v \in V_i$. More generally, σ is called a *finite-memcru strategy* if the equivalence relation $\sim_\sigma$ on V^* defined by $x \sim_\sigma x'$ if $\sigma(xz) = \sigma(x'z)$ for all $z \in V^*V_i$ has finite index. **In** other words, a finite-memory strategy is a strategy that can be implemented by a finite automaton with output. A strategy profile $(\sigma_i)_{i\in\Pi}$ is called *positional* or a *finite-memcru strategy profile* if each *o,* is positional or a finite-memory strategy, respectively.

It is sometimes convenient to designate an initial vertex *Va* ∈ *V* of the game. We call the tuple $(\mathcal{G}, va)$ an *initialised infinite multiplayeT game.* A *play (histOTY) of* $(\mathcal{G}, va)$ is a play (history) of $\mathcal{G}$ starting with *Va.* A strategy

(strategy profile) of $(\mathcal{G}, va)$ is just a strategy (strategy profile) of Q. A strategy σ of some player i in $(\mathcal{G}, va)$ is *winning* if every play of $(\mathcal{G}, va)$ consistent with σ is won by player i. A strategy profile $(\sigma_i)_{i\in\Pi}$ of $(\mathcal{G}, va)$ determines a unique play of $(\mathcal{G}, va)$ consistent with each σ_i, called the *outcome of* $(\sigma_i)_{i\in\Pi}$ and denoted by $\langle\langle(\sigma_i)_{i\in\Pi}\rangle\rangle$ or, in the case that the initial vertex is not understood from the context, $\langle\langle(\sigma_i)_{i\in\Pi}\rangle\rangle_{v_0}$. **In** the following, we will often use the term *game* to denote an *(initialised) infinite muliiplosjer game* according to Definition 2.1.

We have introduced winning conditions as abstract sets of infinite sequences over the set of colours. **In** verification the winning conditions usually are *ui-requlars sets* specified by formulae of the logic SIS (monadic second-order logic on infinite words) or **LTL** (linear-time temporallogic) referring to unary predicates *Pc* indexed by the set C of colours. Special cases are the following well-studied winning conditions:

- *Büchi* (given by $F \subseteq C$): defines the set of all $\alpha \in C^\omega$ such that $\alpha(k) \in F$ for infinitely many $k < \omega$.
- *co-Büchi* (given by $F \subseteq C$): defines the set of all $\alpha \in C^\omega$ such that $\alpha(k) \in F$ for all but finitely many $k < \omega$.
- *Paritsj* (given by a *prioritij function* $\Omega : C \to \omega$): defines the set of all $\alpha \in C^\omega$ such that the least number occurring infinitely often in $\Omega(\alpha)$ is even.
- *Rabin* (given by a set Ω of *pairs* (G_i, R_i) where $G_i, R_i \subseteq C$): defines the set of all $\alpha \in C^\omega$ such that there exists an index i with $\alpha(k) \in G_i$ for infinitely many $k < \omega$ but $\alpha(k) \in R_i$ only for finitely many $k < \omega$.
- *Sireett* (given by a set Ω of *pairs* (G_i, R_i) where $G_i, R_i \subseteq C$): defines the set of all $\alpha \in C^\omega$ such that for all indices i with $\alpha(k) \in R_i$ for infinitely many $k < \omega$ also $\alpha(k) \in O_i$ for infinitely many $k < \omega$.
- *Muller* (given by a family $\mathcal{F}$ of *accepting sets* $F_i \subseteq C$): defines the set of all $\alpha \in C^\omega$ such that there exists an index i with the set of colours seen infinitely often in α being precisely the set F_i.

Note that (co-)Büchi conditions are a special case of parity conditions with two priorities, and parity conditions are a special case of Rabin and Streett conditions, which are special cases of Muller conditions. Moreover, the complement of a Büchi or Rabin condition is a co-Büchi or Streett condition, respectively, and vice versa, whereas the class of parity conditions and the class of Muller conditions are closed under complement. Finally, any of these conditions is *prefix independent*, i.e., for every $\alpha \in C^\omega$ and $x \in C^*$ it is the case that α satisfies the condition if and only if $x\alpha$ does.

We call a game $\mathcal{G}$ a *multiplayeT ui-requlaT, (co-)Büchi, paTity, Rabin, Streeti, 0T Muller game* if the winning condition of *each* player is of the specified type. This differs somewhat from the usual convention for two-player zero-sum games where a Büchi or Rabin game is a game where the winning condition of the *first* player is a Büchi or Rabin condition, respectively.

Note that we do distinguish between colours and priorities. For two-player zero-sum parity games, one can identify them by choosing a finite subset of w as the set C of colours and defining the parity condition directly on the set C, i.e., the priority function of the first player is the identity function, and the priority function of the second player is the successor function $k \mapsto k + 1$. This gives *paTity games* as considered in the literature (Zielonka, 1998).

The importance of the parity condition sterns from three facts: First, the condition is expressive enough to express any w-regular objective. More precisely, for every w-regular language of infinite words, there exists a deterministic word automaton with a parity acceptance condition that recognises this language. As demonstrated by Thomas (1995), this allows to reduce a two-player zero-sum game with an arbitrary w-regular winning condition to a parity game. (See also Wolfgang Thomas' contribution to this volume.) Second, two-player zero-sum parity games arise as the model-checking games for fixed-point logies, in particular the modal μ-calculus (Grädel, 2007). Third, the condition is simple enough to allow for *positional* winning strategies (see above) (Emerson and Jutla, 1991; Mostowski, 1991), i.e., if one player has a winning strategy in a parity game she also has a posit.ional one. **In** (Umrnels, 2006) iL was shown that Lhe Iirst property extends to the multiplayer case: Any multiplayer game with w-regular winning conditions can be reduced to a game with parity winning conditions. Hence, in the algorithmic part of this paper, we will concentrate on multiplayer parity games.

3 Solution concepts

So far, the infinite games used in verification mostly are two-player games with win-lose conditions, i.e., each play is won by one player and lost by the other. The key concept for such games is *deierminacsj:* a game is determined if, from each initial position, one of the players has a winning strategy.

While it is well-known that, on the basis of (a weak form of) the Axiom of Choice, non-determined games exist, the two-player win-lose games usually encountered in computer science, in particular all w-regular games, are determined. Indeed, this is true for much more general games where the winning conditions are arbitrary (quasi-)Borel sets (Martin, 1975, 1990).

In the case of a determined game, solving the game means to compute the winning regions and winning strategies for the two players. A famous

result due to Büchi and Landweber (1969) says that in the case of games on finite graphs and with w-regular winning conditions, we can effectively compute winning strategies that are realisabie by finite automata.

When we move to multiplayer games andjor non-zero sum games, other solution concepts are needed. We will explain some of these concepts, in particular Nash equilibria, subgame perfect equilibria, and secure equilibria, and relate the existence of these equilibria (for the kind of infinite games studied here) to the determinacy of certain associated two-player games.

3.1 Nash equilibria

The most popular solution concept in classical game theory is the concept of a *Nash equilibrium.* Informally, a Nash equilibrium is a strategy profile from which no player has an incentive to deviate, if the other players stick to their strategies. A celebrated theorem by John Nash (1950) says that in any game where each player only has a finite collection of strategies there is at least one Nash equilibrium provided that the players can randomise over their strategies, i.e., choose *mixed stmtegies* rather than only pure ones. For turn-based (non-stochastic) games with qualitative winning conditions, mixed strategies play no relevant role. We define Nash equilibria just in the form needed here.

Definition 3.1. A strategy profile $(\sigma_i)_{i\in\Pi}$ of a game $(\mathcal{G}, va)$ is called a *Nash equilibriurn* if for every player i **EIland** all her possible strategies σ' in $(\mathcal{G}, v_0)$ the play $\langle \sigma', (\sigma_j)_{j\in\Pi\setminus\{i\}} \rangle$ is won by player i only if the play $\langle (\sigma_j)_{j\in\Pi} \rangle$ is also won by her.

It has been shown by Chatterjee et al. (2004b) that every multiplayer game with Borel winning conditions has a Nash equilibrium. We will prove a more general result below.

Despite the importance and popularity of Nash equilibria, there are several probiems with this solution concept, in particular for games that extend over time. This is due to the fact that Nash equilibria do not take into account the sequential nature of these games and its consequences. Aftel' any initial segment of a play, the players face a new situation and may change their strategies. Choices made because of a threat by the other players may no longer be rational, because the opponents have lost their power of retaliation in the remaining play.

Example 3.2. Consider a two-player Büchi game with its arena depicted in Figure 1; round vertices are controlled by player 1; boxed vertices are controlled by player 2; each of the two players wins if and only if vertex 3 is visited (infinitely often); the initial vertex is 1. Intuitively, the only rational outcome of this game should be the play 123^w. However, the game has two Nash equilibria:

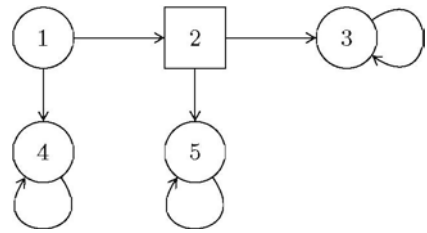


FIGURE 1. A two-player Büchi game.

1. Player 1 moves from vertex 1 to vertex 2, and player 2 moves from vertex 2 to vertex 3. Hence, both players win.

2. Player 1 moves from vertex 1 to vertex 4, and player 2 moves from vertex 2 to vertex 5. Hence, both players lose.

The second equilibrium certainly does not describe rational behaviour. Indeed both players move according to a strategy that is always losing (whatever the other player does), and once player 1 has moved from vertex 1 to vertex 2, then the rational behaviour of player 2 would be to change her strategy and move to vertex 3 instead of vertex 5 as this is then the only way for her to win.

This example can be modified in many ways. Indeed we can construct games with Nash equilibria in which every player moves infinitely often according to a losing strategy, and only has a chance to win if she deviates from the equilibrium strategy. The following is an instructive example with quantitative objectives.

Example 3.3. Let Q_n be an n-player game with positions $0, \ldots, n$. Position n is the initial position, and position 0 is the terminal position. Player i moves at position i and has two options. Either she loops at position i (and stays in control) or moves to position $i-1$ (handing control to the next player). For each player, the value of a play π is $(n+1)/|\pi|$. Hence, for all players, the shortest possible play has value 1, and all infinite plays have value 0. Obviously, the rational behaviour for each player i is to move from i to $i-1$. This strategy profile, which is of course a Nash equilibrium, gives value 1 to all players. However, the 'most stupid' strategy profile, where each player loops forever at his position, i.e., moves forever according to a losing strategy, is also a Nash equilibrium.

3.2 Subgame perfect equilibria

An equilibrium concept that respects the possibility of a player to change her strategy during a play is the notion of a subgame perfect equilibrium (Selten, 1965). For being a subgame perfect equilibrium, a choice of strategies is not only required to be optimal for the initial vertex but for every possible initial history of the game (including histories not reachable in the equilibrium play).

To define subgame perfect equilibria formal!y, we need the notion of a subgame: For a game $\mathcal{G}$ = (II, V, $(V_i)_{i\in\Pi}$, E, X, (Wini)iETI) and a history h of Q, let the game $\mathcal{G}|_h$ = (II, V, $(V_i)_{i\in\Pi}$, E, X, (Win, Ih)iETI) be defined by Win, Ih = $\{\alpha$ E CW : X(h) . α E $\mathrm{Win}_i\}$. For an initialised game $(\mathcal{G}, va)$ and a history hv of $(\mathcal{G}, v_0)$, we cal! the initialised game $(\mathcal{G}|_h, v)$ the *subgame of* $(\mathcal{G}, va)$ *with history* hv. For a strategy σ of player i E II in Q, let $\sigma|_h$: $V^*V_i \to V$ be defined by $\sigma|_h(xv) = \sigma(hxv)$. Obviously, $\sigma|_h$ is a strategy of player i in $\mathcal{G}|_h$.

Definition 3.4. A strategy profile $(\sigma_i)_{i\in\Pi}$ of a game $(\mathcal{G}, va)$ is cal!ed a *subgame perfect equilibrium (SPE)* if $(\sigma_i|_h)_{i\in\Pi}$ is a Nash equilibrium of $(\mathcal{G}|_h, v)$ for every history hv of $(\mathcal{G}, v_0)$.

Example 3.5. Consider again the game described in Example 3.2. The Nash equilibrium where player 1 moves from vertex 1 to vertex 4 and player 2 moves from vertex 2 to vertex 5 is not a subgame perfect equilibrium since moving from vertex 2 to vertex 5 is not optimal for player 2 aftel' the play has reached vertex 2. On the other hand, the Nash equilibrium where player 1 moves from vertex 1 to vertex 2 and player 2 moves from vertex 2 to vertex 3 is also a subgame perfect equilibrium.

It is a classical result due to Kuhn (1953) that every *finite* game (i.e., every game played on a finite tree with payoffs attached to leaves) has a subgame perfect equilibrium. The first step in the analysis of subgame perfect equilibria for *infinite* duration games is the notion of subgame-perfect determinacy. While the notion of subgame perfect equilibrium makes sense for more general classes of infinite games, the notion of subgame-perfect determinacy applies only to games with qualitative winning conditions (which is tacitly assumed from now on).

Definition 3.6. A game $(\mathcal{G}, va)$ is *subgame-perfect deiermined* if there exists a strategy profile $(\sigma_i)_{i\in\Pi}$ such that for each history hv of the game one of the strategies $\sigma_i|_h$ is a winning strategy in $(\mathcal{G}|_h, v)$.

Proposition 3.7. Let $(\mathcal{G}, va)$ be a qualitative zero-sum game such that every subgame is determined. Then $(\mathcal{G}, va)$ is subgame-perfect determined.

Proof. Let $(\mathcal{G}, va)$ be a multiplayer game such that for every history hv there exists a strategy σ_i^h for some player i that is winning in $(\mathcal{G}|_h, v)$. (Note that we can assume that σ_i^h is independent of v.) We have to combine these strategies in an appropriate way to strategies as. (Let us point out that the trivial combination, namely $\sigma_i(hv) := \sigma_i^h(v)$, does not work in genera!.) We say that a decomposition $hv = hl \cdot h_2$ is *good* for player i w.r.t. vertex v if $\sigma_i^{h_1}|_{h_2}$ is winning in $(\mathcal{G}|_h, v)$. If the strategy σ_i^h is winning in $(\mathcal{G}|_h, v)$, then the decomposition $h = h \cdot \varepsilon$ is good w.r.t. v, so a good decomposition exists.

For each history hv, if σ_i^h is winning in $(\mathcal{G}|_h, v)$, we choose the good (w.r.t. vertex v) decomposition $h = h_1 h_2$ with minimal hl, and put

$$\sigma_i(hv) := \sigma_i^{h_1}(h_2 v)\,.$$

Otherwise, we set

$$\sigma_i(hv) := \sigma_i^h(v)\,.$$

It remains to show that for each history hv of $(\mathcal{G}, va)$ the strategy $\sigma_i|_h$ is winning in $(\mathcal{G}|_h, v)$ whenever the strategy σ_i^h is. Hence, assume that σ_i^h is winning in $(\mathcal{G}|_h, v)$, and let $tt = \pi(0)\pi(1)\ldots$ be a play starting in $\pi(0) = v$ and consistent with $\sigma, |h'$ We need to show that tt is won by player i in $(\mathcal{G}|_h, v)$.

First, we claim that for each $k < w$ there exists a decomposition of the form $h\pi(0)\ldots tt[:k-1) = hl'(h_2\pi(0)\ldots tt[:k-1))$ that is good for player i w.r.t. $\pi(k)$. This is obviously true for $k = 0$. Now, for $k > 0$, assume that there exists a decomposition $h\pi(0)\ldots tt[:k-2) = hl'(h_2\pi(0)\ldots tt[:k-2))$ that is good for player i w.r.t. $tt[:k-1)$ and with hl being minima!. Then $\pi(k) = \sigma_i(h\pi(0) \quad \pi(k-1)) = \sigma^{h_1}(h_2\pi(0)\ldots\pi(k-1)$, and $h\pi(0)\ldots\pi(k-1) = h_1(h_2\pi(0) \quad tt[:k-1))$ is a decomposition that is good w.r.t. $\pi(k)$.

Now consider the sequence $h_1^0, h_1^1, \quad$ of prefixes of the good decompositions $h\pi(0)\ldots tt[:k-1) = h_1^k h_2^k \pi(0) \quad tt[:k-1)$ (w.r.t. $\pi(k)$) with each h_1^k being minima!. Then we have $h_1^0 \succeq h_1^1 \succeq \ldots$, since for each $k > 0$ the decomposition $h\pi(0)\ldots tt[:k-1) = h_1^{k-1} h_2^{k-1} \pi(0)\ldots tt[:k-1)$ is also good for player i w.r.t. $\pi(k)$. As $\prec$ is well-founded, there must exist $k < w$ such that $hl := h_1^k = h_1^l$ and $h_2 := h_2^k = h_2^l$ for each $k \leq l < w$. Hence, we have that the play $\pi(k)\pi(k+1)\ldots$ is consistent with $\sigma_i^{h_1}|_{h_2\pi(0)\ldots\pi(k-1)}$, which is a winning strategy in $(\mathcal{G}|_{h\pi(0)\ldots\pi(k-1)}, \pi(k))$. So the play $ha:$ is won by player i in $(\mathcal{G}, va)$, which implies that the play tt is won by player i in $(\mathcal{G}|_h, v)$. Q.E.D.

We say that a class of winning conditions is closed under taking subgames, if for every condition $X \subseteq CW$ in the class, and every $h \in C^*$, also $Xlh := \{x \in CW : hx \in X\}$ belongs to the class. Since Borel winning conditions are closed under taking subgames, it follows that any two-player zero-sum game with Borel winning condition is subgame-perfect determined.

Corollary 3.8. Let $(\mathcal{G}, va)$ be a two-player zero-sum Borel game. Then $(\mathcal{G}, va)$ is subgame-perfect determined.

Multiplayer games are usually not zero-sum games. Indeed when we have many players the assumption that the winning conditions of the players form a partition of the set of plays is very restrictive and unnatura1. We now drop this assumption and establish general conditions under which a multiplayer game admits a subgame perfect equilibrium. **In** fact we will relate the existence of subgame perfect equilibria to the determinacy of associated two-player games. **In** particular, it will follow that every multiplayer game with Borel winning conditions has a subgame perfect equilibrium.

In the rest of this subsection, we are only concerned with the *existence* of equilibria, not with their complexity. Thus, without loss of generality, we assume that the arena of the game under consideration is a tree or a fotest with the initial vertex as one of its roots. The justification for this assumption is that we can always replace the arena of an arbitrary game by its unravelling from the initial vertex, ending up in an equivalent game.

Definition 3.9. Let $\mathcal{G}$ = (11, V, $(V_i)_{i\in\Pi}$, E, X, (Wini)iETI) be a multiplayer game (played on a forest), with winning conditions Win, $\subseteq C'''$, The associated class $\mathrm{ZeroSum}(\mathcal{G})$ of two-player zero-sum games is obtained as follows:

1. For each player i, $\mathrm{ZeroSum}(\mathcal{G})$ contains the game Qi where player i plays Q, with his winning condition Win., against the coalition of all other players, with winning condition CW \ Win..
2. Close the class under taking subgames (i.e., consider plays aftel' initial histories).
3. Close the class under taking subgraphs (i.e., admit deletion of positions and moves).

Note that the order in which the operations (1), (2), and (3) are applied has no effect on the class ZeroSum(Ç).

Theorem 3.10. Let $(\mathcal{G}, va)$ be a multiplayer game such that every game in $\mathrm{ZeroSum}(\mathcal{G})$ is determined. Then $(\mathcal{G}, va)$ has a subgame perfect equilibrium.

Proof. Let $\mathcal{G}$ = (11, V, $(V_i)_{i\in\Pi}$, E, X, (Wini)iETI) be a multiplayer game such that every game in $\mathrm{ZeroSum}(\mathcal{G})$ is determined. For each ordinal α we define a set $E^\alpha \subseteq E$ beginning with $EO = E$ and

$$E^\lambda = \bigcap_{\alpha<\lambda} E^\alpha$$

for limit ordinals λ. To define $E^{\alpha+1}$ from E^α, we consider for each player **i E** 11 the two-player zero-sum game $\mathcal{G}_i^\alpha$ = $(V, V_i, E^\alpha, X,$ Win.) where player i

plays with his winning condition Win, against the coalition of all other players (with winning condition *CW*\ Wiu.). Every subgame of $\mathcal{G}_i^\alpha$ belongs to $\mathrm{ZeroSum}(\mathcal{G})$ and is therefore determined. Hence we can use Proposition 3.7 to fix a subgame perfect equilibrium $(\sigma_i^\alpha, \sigma_{-i}^\alpha)$ of $(\mathcal{G}_i^\alpha, va)$ where σ_i^α is a strategy of player i and σ_{-i}^α is a strategy of the coalition. Moreover, as the arena of $\mathcal{G}^\alpha$ is a forest, these strategies can be assumed to be positiona1. Let X_i^α be the set of all $v \in V$ such that σ_i^α is winning in $(\mathcal{G}_i^\alpha \mathrm{Ih}, v)$ for the unique maximal history h of $\mathcal{G}$ leading to v. For vertices $v \in V_i \cap X_i^\alpha$ we delete all outgoing edges except the one taken by the strategy σ_i^α, i.e., we define

$$E^{\alpha+1} = E^\alpha \setminus \bigcup_{\mathrm{iETI}} \{(u,v) \in E : u. \in V_i \cap X_i^\alpha \text{ and } v \neq \sigma_i^\alpha(u)\}.$$

Obviously, the sequence $(E^\alpha)_{\alpha \in \mathrm{On}}$ is nonincreasing. Thus we can fix the least ordinal ξ with $E^\xi = E^{\xi+1}$ and define $o; = \sigma_i^\xi$ and $a\,i., = \sigma_{-i}^\xi$. Moreover, for each player $j \neq$ i let $\sigma_{j,i}$ be the positional strategy of player j in $\mathcal{G}$ that is induced by σ_{-i}.

Intuitively, Player i's equilibrium strategy τ_i is as follows: Player i plays $o;$ as long as no other player deviates. Whenever some player $j \neq$ i deviates from her equilibrium strategy σ_j, player i switches to $\sigma_{i,j}$. Formally, define for each vertex $v \in V$ the player $p(v)$ who has to be "punished" at vertex v where $p(v) = \bot$ if nobody has to be punished. If the game has just started, no player should be punished. Thus we let

$$p(v) = \bot \text{ if } v \text{ is a root.}$$

At vertex v with predecessor u, the same player has to be punished as at vertex $u.$ as long as the player whose turn it was at vertex $u.$ did not deviate from her prescribed strategy. Thus for $u. \in V_i$ and $v \in uE$ we let

$$p(v) = \begin{cases} \bot & \text{if } p(u) = \bot \text{ and } v = \sigma_i(u), \\ p(u) & \text{if } p(u) \neq \text{i, } p(u) \neq \bot \text{ and } v = \sigma_{i,p(u)}(u), \\ i & \text{otherwise.} \end{cases}$$

Now, for each player i $\in$ II we can define the equilibrium strategy τ_i by setting

$$\tau_i(v) = \begin{cases} \sigma_i(v) & \text{if } p(v) = \bot \text{ or } p(v) = \text{i,} \\ \sigma_{i,p(v)}(v) & \text{otherwise} \end{cases}$$

for each $v \in V$.

It remains to show that (Ti)iETI is a subgame perfect equilibrium of $(\mathcal{G}, v_0)$. First note that $o;$ is winning in $(\mathcal{G}_i^\xi|_h, v)$ if σ_i^α is winning in $(\mathcal{G}_i^\alpha|_h, v)$ for some ordinal α because if σ_i^α is winning in $(\mathcal{G}_i^\alpha \mathrm{Ih}, v)$ every play of

$(\mathcal{G}_i^{\alpha+1}|_h, v)$ is consistent with σ_i^α and therefore won by player i. As $E^\xi \subseteq E^{\alpha+1}$, this also holds for every play of $(\mathcal{G}_i^\xi$ Ih, v). Now let v be any vertex of $\mathcal{G}$ with h the unique maximal history of $\mathcal{G}$ leading to v. We claim that (Tj)jETI is a Nash equilibrium of $(\mathcal{G}|_h, v)$. Towards this, let TI be any strategy of any player i E II in Q; let tt = ((Tj)jETI)v, and let π' = (TI, (Tj)jETI\{i})V' We need to show that ha: is won by player i or that hm' is not won by player i. The claim is trivial if tt = n', Thus assume that $tt \neq \pi'$ and fix the least k < w such that $\pi(k+1) \neq \pi'(k+1)$. Clearly, $\pi(k) \in V_i$ and $\tau'(\pi(k)) \neq \tau_i(\pi(k))$. Without loss of generality, let k = 0. We distinguish the foliowing two cases:

- o; is winning in $(\mathcal{G}_i^\xi|_h, v)$. By the definition of each Tj, tt is a play of $(\mathcal{G}_i^\xi|_h, v)$. We claim that tt is consistent with σ_i; which implies that ha: is won by player i. Otherwise fix the least l < w such that $\pi(l) \in V_i$ and $\sigma_i(\pi(l)) \neq \pi(l+1)$. As o; is winning in $(\mathcal{G}_i^\xi|_h, v)$, o; is also winning in $(\mathcal{G}_i^\xi|_{h\pi(0)\dots\pi(l-1)}, \pi(l))$. But then $(\pi(l), \pi(l+1)) \in E^\xi \setminus E^{\xi+1}$, a contradiction to $E^\xi = E^{\xi+1}$.

- o; is not winning in $(\mathcal{G}_i^\xi|_h, v)$. Hence σ_{-i} is winning in $(\mathcal{G}_i^\xi|_h, v)$. As TI(v) $\neq$ Ti(v), player i has deviated, and it is the case that π' = $(TI, (\sigma_{j,i})_{j\in\Pi\setminus\{i\}}\rangle_v$. We claim that π' is a play of $(\mathcal{G}_i^\xi|_h, v)$. As σ_{-i} is winning in $(\mathcal{G}_i^\xi$ Ih, v), this implies that hn:' is not won by player i. Otherwise fix the least l < w such that $(\pi'(l), \pi'(l+1)) \notin E^\xi$ together with the ordinal α such that $(\pi'(l), \pi'(l+1)) \in E^\alpha \setminus E^{\alpha+1}$. Clearly, $\pi'(l) \in V_i$. Thus σ_i^α is winning in $(\mathcal{G}_i^\alpha|_{h\pi'(0)\dots\pi'(l-1)}, \pi'(l))$, which implies that o, is winning in $(\mathcal{G}_i^\xi|_{h\pi'(0)\dots\pi'(l-1)}, \pi'(l))$. As π' is consistent with σ_{-i}, this means that σ_{-i} is not winning in $(\mathcal{G}_i^\xi|_h, v)$, a contradiction.

It foliows that (Tj)jETI = (Tj Ih)jETI is a Nash equilibri urn of $(\mathcal{G}|_h, v)$ for every history hv of $(\mathcal{G}, v_0)$. Hence, (Tj)jETI is a subgame perfect equilibrium of $(\mathcal{G}, Vo)$. Q.E.D.

Corollary 3.11 (Ummels, 2006). Every multiplayer game with Borel winning conditions has a subgame perfect equilibrium.

This generalises the result in (Chatterjee et al., 2004b) that every multiplayer game with Borel winning conditions has a Nash equilibrium. Indeed, for the existence of Nash equilibria, a slightly weaker condition than the one in Theorem 3.10 suffices. Let ZeroSum(Ç)Nash be defined in the same way as $\mathrm{ZeroSum}(\mathcal{G})$ but without closure under subgraphs.

Corollary 3.12. If every game in ZeroSum(Ç)Nash is determined, then $\mathcal{G}$ has a Nash equilibrium.

3.3 Secure equilibria

The notion of a *secure equilibrium* introduced by Chatterjee et al. (2004a) tries to overcome another deficiency of Nash equilibria: one game may have many Nash equilibria with different payoffs and even several maximal ones w.r.t. to the componentwise partial ordering on payoffs. Hence, for the players it is not obvious which equilibrium to play. The idea of a secure equilibrium is that any rational deviation (i.e., a deviation that does not decrease the payoff of the player who deviates) will not only not increase the payoff of the player who deviates but it will also not decrease the payoff of any other player. Secure equilibria model rational behaviour if players not only attempt to maximise their own payoff but, as a secondary objective, also attempt to minimise their opponents' payoffs.

Definition 3.13. A strategy profile $(\sigma_i)_{i\in\Pi}$ of a game $(\mathcal{G}, va)$ is called *secure* if for all players $i \neq j$ and for each strategy σ' of player j it is the case that

$$\begin{aligned} &\langle(\sigma_i)_{i\in\Pi}\rangle \notin \text{Win}, \text{ or } \langle(\sigma_i)_{i\in\Pi\setminus\{j\}}, \sigma'\rangle \in \text{Win}, \\ \Rightarrow\ &\langle(\sigma_i)_{i\in\Pi}\rangle \notin \text{Win}, \text{ or } \langle(\sigma_i)_{i\in\Pi\setminus\{j\}}, \sigma'\rangle \in \text{Win}, . \end{aligned}$$

A strategy profile $(\sigma_i)_{i\in\Pi}$ is a *secure equilibrium* if it is both a Nash equilibrium and secure.

Example 3.14 (Chatterjee et al., 2004a). Consider another Büchi game played on the game graph depicted in Figure 1 by the two players 1 and 2 where, again, round vertices are controlled by player 1 and square vertices are controlled by player 2. This time player 1 wins if vertex 3 is visited (infinitely often), and player 2 wins if vertex 3 or vertex 5 is visited (infinitely often). Again, the initial vertex is 1.

Up to equivalence, there are two different strategies for each player: Player 1 can choose to go from 1 to either 2 or 4 while player 2 can choose to go from 2 to either 3 or 5. Except for the strategy profile where player 1 moves to 4 and player 2 moves to 3, all of the resulting profiles are Nash equilibria. However, the strategy profile where player 1 moves to 2 and player 2 moves to 3 is not secure: Player 2 can decrease player 1's payoff by moving to 5 instead while her payoffremains the same (namely 1). Similarly, the strategy profile where player 1 moves to 2 and player 2 moves to 5 is not secure: Player 1 can decrease player 2's payoff by moving to 4 instead while her payoff remains the same (namely 0). Hence, the strategy profile where player 1 moves to 4 and player 2 moves to 5 is the only secure equilibrium of the game.

It is an open question whether secure equilibria exist in arbitrary multiplayer games with well-behaved winning conditions. However, for the case of only two players, it is not only known that there always exists a

secure equilibrium for games with well-behaved winning conditions, but a unique maximal secure equilibrium payoff w.r.t. the componentwise ordering $\leq$ on payoffs, i.e., there exists a secure equilibrium (σ, τ) such that $\mathrm{pay}(\langle \sigma', \tau' \rangle) \leq \mathrm{pay}(\langle \sigma, \tau \rangle)$ for every secure equilibrium (σ', TI) of $(\mathcal{G}, v_0)$. Clearly, such an equilibrium is preferabie for both players.

For two winning conditions Win.. Win., $\subseteq V^w$. we say that the pair (Winj, Win-) is *deiermined* if any Boolean combination of Win, and Win, is determined, i.e., any two-player zero-sum game that has a Boolean combination of Win, and Win, as its winning condition is determined.

Definition 3.15. A strategy σ of player 1 (player 2) in a 2-player game $(\mathcal{G}, va)$ is *strongly winning* if it ensures a play with payoff (1, 0) (payoff (0, 1)) against any strategy τ of player 2 (player 1).

The strategy σ is *retaliating* if it ensures a play with payoff (0, 0), (1,0), or (1,1) against any strategy τ of player 2 (player 1).

Note that if $(\mathcal{G}, va)$ is a game with a determined pair $(\mathrm{Win}_1, \mathrm{Win2})$ of winning conditions, then player 1 or 2 has a strongly winning strategy if and only if the other player does not have a retaliating strategy.

Proposition 3.16. Let $(\mathcal{G}, va)$ be a two-player game with a determined pair (Winj, Win-) of winning conditions. Then precisely one of the following four cases holds:

1. Player 1 has a strongly winning strategy;
2. Player 2 has a strongly winning strategy;
3. There is a pair of retaliating strategies with payoff (1, 1);
4. There is a pair of retaliating strategies, and all pairs of retaliating strategies have payoff (0,0).

Proo]. Note that if one player has a strongly winning strategy, then the other player neither has a strongly winning strategy nor a retaliating strategy. Vice versa, if one player has a retaliating strategy, then the other player cannot have a strongly winning strategy. iVIoreover, cases 3 and 4 exclude each other by definition. Hence, at most one of the four cases holds.

Now, assume that neither of the cases 1-3 holds. **In** particular, no player has a strongly winning strategy. By determinacy, this implies that both players have retaliating strategies. Let (σ, T) be any pair of retaliating strategies. As case 3 does not hold, at least one of the two players receives payoff 0. But as both players play retaliating strategies, this implies that both players receive payoff 0, so we are in case 4. Q.E.D.

Theorem 3.17. Let $(\mathcal{G}, va)$ be a two-player game with a determined pair (Win- , Win-) of winning conditions. Then there exists a unique maximal secure equilibrium payoff for $(\mathcal{G}, va)$.

Proo]. We show that the claim holds in any of the four cases stated in Proposition 3.16:

1. **In** the first case, player 1 has a strongly winning strategy σ. Then, for any strategy τ of player 2, the strategy profile (σ, T) is a secure equilibrium with payoff (1, 0). We claim that (1,0) is the unique maximal secure equilibrium payoff. Otherwise, there would exist a secure equilibrium with payoff 1 for player 2. But player 1 could decrease player 2's payoff while not decreasing her own payoff by playing σ, a contradiction.

2. The case that player 2 has a strongly winning strategy is analogous to the first case.

3. **In** the third case, there is a pair (σ, T) of retaliating strategies with payoff (1,1). But then (σ, T) is a secure equilibrium, and (1,1) is the unique maximal secure equilibrium payoff.

4. **In** the fourth case, there is a pair of retaliating strategies, and any pair of retaliating strategies has payoff (0,0). Then there exists a strategy σ of player 1 that guarantees payoff 0 for player 2, since otherwise by determinacy there would exists a strategy for player 2 that guarantees payoff 1 for player 2. This would be a retaliating strategy that guarantees payoff 1 for player 2, a contradiction to the assumption that all pairs of retaliating strategies have payoff (0,0). Symmetrically, there exists a strategy τ of player 2 that guarantees payoff 0 for player 1. By the definition of σ and τ, the strategy profile (σ, τ) is a Nash equilibrium. But it is also secure, since it gives each player the least possible payoff. Hence, (σ, T) is a secure equilibrium. Now assume there exists a secure equilibrium (σ', TI) with payoff (1, 0). Then also (σ', τ) would give payoff 1 to player 1, a contradiction to the fact that (σ, T) is a Nash equilibrium. Symmetrically, there cannot exists a secure equilibrium (σ', TI) with payoff (0,1). Hence, either (0,0) or (1,1) is the unique maximal secure equilibrium payoff. Q.E.D.

Since Borel winning conditions are closed under Boolean combinations, as a corollary we get the result by Chatterjee et al. that any two-player game with Borel winning conditions has a unique maximal secure equilibrium payoff.

Corollary 3.18 (Chatterjee et al., 2004a). Let $(\mathcal{G}, va)$ be two-player game with Borel winning conditions. Then there exists a unique maximal secure equilibrium payoff for $(\mathcal{G}, va)$.

4 Algorithmic problems

Previous research on algorithrns for multiplayer games has focused on computing *some* solution of the game, e.g., *some* Nash equilibrium (Chatterjee et al., *2aa4b).* However, as we have seen, a game may not have a unique solution, so one might be interested not in *any* solution, but in a solution that fulfils certain requirements. For example, one might look for a solution where certain players win while certain ot her players lose. Or one might look for a *maximal* solution, i.e., a solution such that there does not exist another solution with a higher payoff. **In** the context of games with parity winning conditions, this motivation leads us to the following dec ision problem, which can be defined for any solution concept *S:*

> Given a multiplayer parity game *(Q,va)* played on a finite arena and thresholds $\bar{x}, \bar{y} \in \{0, l\}^k$, deeide whether $(\mathcal{G}, va)$ has a solution $(\sigma_i)_{i \in \Pi} \in S(Q,va)$ sueh that $\bar{x} \leq \mathrm{pay}(\langle (\sigma_i)_{i \in \Pi} \rangle) \leq \bar{y}$.

In particular, the solution concepts ofNash equilibria, subgame perfect equilibria, and secure equilibria give rise to the dec ision problerns NE, SPE and SE, respectively. **In** the following three sections, we analyse the complexity of these three probiems.

4.1 Nash equilibria

Let $(\mathcal{G}, va)$ be a game with prefix-independent, determined winning conditions. Assume we have found a Nash equilibrium $(\sigma_i)_{i \in \Pi}$ of $(\mathcal{G}, v_0)$ with payoff $\bar{x}$. Clearly, the play $\langle (\sigma_i)_{i \in \Pi} \rangle$ never hits the winning reg ion W_i of some player i with $x_i = a$ because otherwise player i can improve her payoff by waiting until the token hits W_i and then apply her winning strategy. The crucial observation is that this condition is also sufheient for a play to be induced by a Nash equilibrium, i.e., $(\mathcal{G}, va)$ has a Nash equilibrium with payoff $\bar{x}$ if and only if there exists a play in (Q, va) with payoff $\bar{x}$ that never hits the winning region of some player i with $x_i = a$.

Lemma 4.1. Let $(\mathcal{G}, va)$ be a k-player game with prefix-independent, determined winning conditions, and let W_i be the winning region of player i in Q. There exists a Nash equilibrium of $(\mathcal{G}, va)$ with payoff $\bar{x} \in \{a, l\}^k$ if and only if there exists a play tt of $(\mathcal{G}, va)$ with payoff $\bar{x}$ such that $\{\pi(k) : k < w\} \cap W_i = 0$ for each player i with $x_i = a$.

Proo]. ($\Rightarrow$) This direction follows from the argumentation above.

($\Leftarrow$) Let tt be a play with payoffx such that $\{\pi(k) : k < w\} \cap W_i = 0$ for each player i with $x_i = a$. Moreover, let τ_{-j} be an optimal strategy of the coalition $1 \setminus \{j\}$ in the two-player zero-sum game Qj where player j plays against all other players in Q, and let $\tau_{i,j}$ be the corresponding strategy of player i in $\mathcal{G}$ (where $\tau_{i,i}$ is an arbitrary strategy). For each player $i \in 1$, we

define a strategy $o;$ in $\mathcal{G}$ as follows:

$$\sigma_i(hv) = \begin{cases} \pi(k+1) & \text{if } hv = \pi(0)\dots\pi(k) \prec tt, \\ \tau_{i,j}(h_2 v) & \text{otherwise,} \end{cases}$$

where, in the latter case, $h = h_{1h2}$ such that hl is the longest prefix of h still being a prefix of $tt,$ and j is the player whose turn it was aftel' that prefix (i.e., hl ends in V_j), where j = i if hl = s.

Let us show that $(\sigma_i)_{i\in\Pi}$ is a Nash equilibrium of $(\mathcal{G}, va)$ with payoff $\overline{x}$. First observe that $\langle(o;)_{\text{iETI}}\rangle = tt,$ which has payoff $\overline{x}$, thus it remains to show that $(\sigma_i)_{i\in\Pi}$ is a Nash equilibrium. So let us assume that some player iE II with $x; = 0$ can improve her payoff by playing according to some strategy σ' instead of as. Then there exists $k < w$ such that $\sigma'(\pi(k)) \neq \sigma_i(\pi(k))$, and consequently from this point onwards $\langle(\sigma_j)_{\text{jETI}\setminus\{\text{i}\}}, \sigma'\rangle$ is consistent with τ-i, the optimal strategy ofthe coalition II\ {i} in $gi.$ Hence, τ_{-i} is not winning from $\pi(k)$. Ey determinacy, this implies that $\pi(k)$ E $W_i,$ a contradiction. Q.E.D.

As an immediate consequence, we get that the problem NE is in NP. However, in many cases, we can do better: For two payoff veetors $\overline{x}, \overline{y}$ E {0, *1}k,* let dist(x, $\overline{y}$) be the *Hamming distance* of $\overline{x}$ and $\overline{y}$, i.e., the numbel' $\sum_{i=1}^{k}$ IYi - *Xii* of nonmatching bits. Jurdziriski (1998) showed that the problem of deciding whether a vertex is in the winning region for player 0 in a two-player zero-sum parity game is in **UP** nco-UP. Recall that **UP** is the class of all problems decidabie by a nondeterministic Turing machine that runs in polynomial timp and has at most one accept.ing' run on every input. We show that the complexity of NE goes down to **UP** n co-**UP** if the Hamming distance of the thresholds is bounded. If additionally the number of priorities is bounded, the complexity reduces further to P.

Theorem 4.2 (Ummels, 2008). NE is in NP. If $\operatorname{dist}(\overline{x}, \overline{y})$ is bounded, NE is in **UP** nco-UP. If additionally the number of priorities is bounded for each player, the problem is in P.

Proo]. An NP algorithm for NE works as follows: On input $(\mathcal{G}, va),$ the algorithm starts by guessing a payoff $\overline{x} \leq \overline{z} \leq \overline{y}$ and the winning region W_i of each player. Then, for each vertex v and each player i, the guess whether v E W_i or $v \notin W_i$ is verified by running the **UP** algorithm for the respective problem. If one guess was incorrect, the algorithm rejects immediately. Otherwise, the algorithm checks whether there exists a winning play from *Va* in the one-player game arising from $\mathcal{G}$ by merging the two players, restricting the arena to G ⌈ *nZi =a(V* \ $W_{i)}$, and imposing the winning condition $\bigwedge_{z_i=1} \Omega_i \wedge \bigwedge_{z_i=0} \neg\Omega_i$, a Streett condition. If so, the algorithm accepts. Otherwise, the algorithm rejects.

The correctness of the algorithm follows from Lemma 4.1. For the complexity, note that deciding whether there exists a winning play in a one-player Streett game can be done in polynomial time (Emerson and Lei, 1985).

If dist(x, $\overline{y}$) is bounded, there is no need to guess the payoff $\overline{z}$. Instead, one can enumerate all of the constantly many payoffs $\overline{x} \leq \overline{z} \leq \overline{y}$ and check for each of them whether there exists a winning play in the respective one-player Streett game. If this is the case for some $\overline{z}$, the algorithm may accept. Otherwise it has to reject. This gives a **UP** algorithm for NE in the case that dist(x, $\overline{y}$) is bounded. Analogously, a **UP** algorithm for the complementary problem would accept if for each $\overline{z}$ there exists *na* winning play in the respective one-player Streett game.

For parity games with a bounded number of priorities, winning regions can actually be computed in polynomial time (see, e.g., Zielonka, 1998). Thus, if additionally the number of priorities for each player is bounded, the guessing of the winning regions can be avoided as weil, so we end up with a deterministic polynomial-time algorithm. Q.E.D.

It is a major open problem whether winning regions of parity games can be computed in polynomial time, in general. This would allow us to decide the problem NE in polynomial time for bounded dist(x, $\overline{y}$) even if the number of priorities is unbounded. Recently, Jurdzinski et al. (2006) gave a deterministic subexponential algorithm for the problem. It follows that there is a deterministic subexponential algorithm for NE if dist(x, $\overline{y}$) is bounded.

Another line of research is to identify structural properties of graphs that allow for a polynomial-time algorithm for the parity game problem. It was shown that winning regions can be computed in polynomial time for parity games played on graphs of bounded DAG-width (Berwanger et al., 2006; Obdrzálek, 2006) (and thus also for graphs of bounded tree width (Obdrzálek, 2003) or bounded entanglement (Berwanger and Grädel, 2005)), and also for graphs of bounded clique width (Obdrzálek, 2007) or bounded Kelly width (Hunter and Kreutzer, 2007). It follows that NE can be decided in polynomial time for games on these graphs if also dist(x, $\overline{y}$) is bounded.

Having shown that NE is in NP, the natural question that arises is whether NE is NP-complete. We answer this question affirmatively. Note that it is an open question whether the parity game problem is NP-complete. In fact, this is rather unlikely, since it would imply that NP = **UP** = co-**UP** = co-NP, and hence the polynomial hierarchy would collapse to its first level. As a matter of fact, we show NP-completeness even for the case of games with co-Büchi winning conditions, a class of games known to be solvable in polynomial time in the classical two-player zero-sum case.

AIso, it suffices to require that only one distinguished player, say the first one, should win in the equilibrium. **In** essence, this shows that NE is a substantially harder problem than the problem of deciding the existence of a winning strategy for a certain player.

Theorem 4.3 (Ummels, 2008). NE is NP-complete for co-Büchi games, even with the thresholds $\overline{x} = (1,0,\ldots,0)$ and $\overline{y} = (1,\ldots,1)$.

Proof. By Theorem 4.2, the problem is in NP. To show that the problem is NP-hard, we give a polynomial-time reduction from SAT. Given a Boolean formula $\varphi = C1 \wedge \ldots \wedge Cm$ in CNF over variables $X1, \ldots, X_n$, we build a game $\mathcal{G}_\varphi$ played by players $0,1,\ldots,n$ as follows. $\mathcal{G}_\varphi$ has vertices $C1,' \ldots, Cm$ controlled by player 0, and for each clause C and each literal $X;$ or $,Xi,$ a vertex (C, Xi) or $(C, ,Xi),$ respectively, controlled by player i. Additionally, there is a sink vertex $\bot$. There are edges from a clause C_j to each vertex (C_j, L) such that L occurs as a literal in C_j and from there to $C(j \bmod m)+1$. Additionally, there is an edge from each vertex $(C, ,Xi)$ to the sink vertex $\bot$. As $\bot$ is a sink vertex, the only edge leaving $\bot$ leads to $\bot$ itself. For example, Figure 2 shows the essential part of the arena of $\mathcal{G}_\varphi$ for the formula $\varphi = (X1 \vee X_3 \vee, X_2) \wedge (X_3 \vee \neg X_1) \wedge, X_3$. The co-Büchi winning conditions are as follows:

- Player 0 wins if the sink vertex is visited only finitely often (or, equivalently, if it is not visited at all).
- Player i E {I ... ,n} wins if each vertex (C, Xi) is visited only finitely often.

Clearly, $\mathcal{G}_\varphi$ can be constructed from φ in polynomial time. We claim that φ is satisfiable if and only if $(\mathcal{G}_\varphi, C_1)$ has a Nash equilibrium where at least player 0 wins.

$(\Rightarrow)$ Assume that φ is satisfiable. We show that the positional strategy profile where at any time player 0 plays from a clause C to a (fixed) literal that satisfies this clause and each player $j \neq 0$ plays from $\neg X_j$ to the sink if and only if the satisfying interpretation maps X_j to true is a Nash equilibrium where player 0 wins. First note that the induced play never reaches the sink and is therefore won by player 0. Now consider any player i that loses the induced play, which can only happen if a vertex (C, Xi) is visited infinitely often. But, as player 0 plays according to the satisfying assignment, this means that no vertex $(C1, ,Xi)$ is ever visited, hence player i has no chance to improve her payoff by playing to the sink vertex.

$(\Leftarrow)$ Assume that $(\mathcal{G}_\varphi, C_1)$ has a Nash equilibrium where player 0 wins, hence the sink vertex is not reached in the induced play. Consider the variabie assignment that maps $X;$ to true if some vertex (C, Xi) is visited infinitely often. We claim that this assignment satisfies the formula. To see

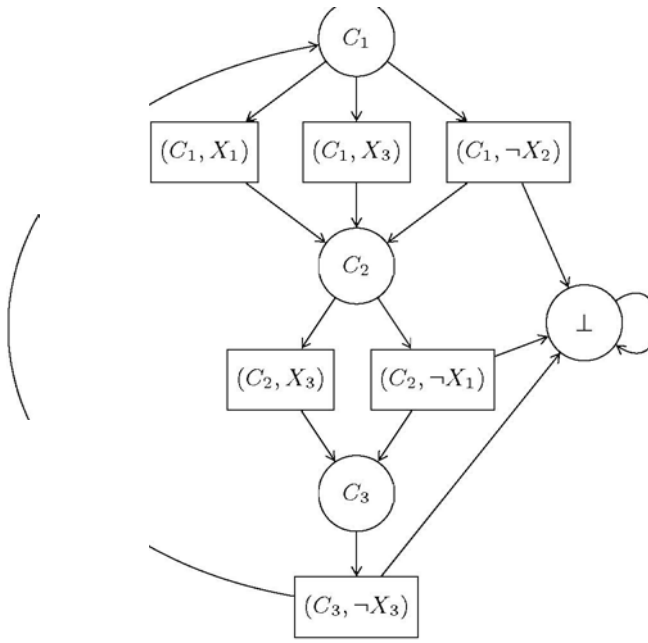


FIGURE 2. The game $\mathcal{G}_\varphi$ for $\varphi = (X_1 \vee X_3 \vee \neg X_2) \wedge (X_3 \vee \neg X_1) \wedge \neg X_3$.

this, consider any clause C_j. By the construction of $\mathcal{G}_\varphi$, there exists a literal X_i or $\neg X_i$ in C_j such that the vertex (C_j, X_i) or $(C_j, \neg X_i)$, respectively, is visited infinitely often. Now assume that both a vertex (C, X_i) and a vertex $(C, \neg X_i)$ are visited infinitely often. Then player i would lose, but could improve her payoff by playing from $(C, \neg X_i)$ to the sink vertex. Hence, in any case the defined interpretation maps the literal to true thus satisfying the clause. Q.E.D.

4.2 Subgame perfect equilibria

For subgame perfect equilibria, we are not aware of a characterisation like the one in Lemma 4.1 for Nash equilibria. Therefore, our approach to solve SPE is entirely different from our approach to solve NE. Namely, we reduce SPE to the nonemptiness problem for tree automata (on infinite trees). However, this only gives an ExpTllVIE upper bound for the problem as opposed to NP for the case of Nash equilibria. For the ful! proof of the fol!owing theorem, see (Ummels, 2006).

Theorem 4.4. The problem SPE is in ExpTIlVIE. **If** the number of players and priorities is bounded, the problem is in P.

Proo] sketch. Without loss of generality, let us assume that the input game $\mathcal{G}$ is *binaTY,* i.e., every vertex of $\mathcal{G}$ has at most two successors. Then we can arrange all plays of $(\mathcal{G}, va)$ in an infinite binary tree with labels from the vertex set *V.* Given a strategy profile $(\sigma_i)_{i\in\Pi}$ of $(\mathcal{G}, v_0)$, we enrich this tree with a second label component that takes the value 0 or 1 if the strategy profile prescribes going to the left or right successor, respectively.

The algorithm works as follows: We construct two *aliernaiinq paTity tree automata.* The first one checks whether some arbitrary tree with labels from the alphabet V x {0, I} is indeed a tree originating from a strategy profile of $(\mathcal{G}, va)$, and the second one checks for a tree originating from a strategy profile $(\sigma_i)_{i\in\Pi}$ of $(\mathcal{G}, va)$ whether $(\sigma_i)_{i\in\Pi}$ is a subgame perfect equilibrium with a payoff in between the given thresholds. The first automaton is actually a nondeterministic tree automaton with trivial acceptance (every run of the automaton is accepting) and has $O(|V|)$ states. The second automaton has $O(kd)$ states and $0(1)$ priorities where k is the number of players and d is the maximum number of priorities in a player's parity condition. An equivalent nondeterministic parity tree automaton has $2^{O(kd \log kd)}$ states and $O(kd)$ priorities (Muller and Schupp, 1995). Finally, we construct the product automaton of the first nondeterministic parity tree automaton with the one constructed from the alternating one. As the former automaton works with trivial acceptance, the construction is straightforward and leads to a nondeterministic parity tree automaton with $O(|V|) \cdot 2^{O(kd \log kd)}$ states and $O(kd)$ priorities. Obviously, the tree language defined by this automaton is nonempty if and only if $(\mathcal{G}, va)$ has a subgame perfect equilibrium with a payoff in between the given thresholds. By (Emerson et al., 1993) nonemptiness for nondeterministic parity tree automata can be decided in time polynomial in the number of states and exponential in the number of priorities. Q.E.D.

The exact complexity of SPE remains an open problem. However, NP-hardness can be transferred from NE to SPE. Hence, it is unlikely that there exists a polynomial-time algorithm for SPE, in genera!.

Theorem 4.5. SPE is NP-hard for co-Büchi games, even with the thresholds $\overline{x}$ = (1,0, ... ,0) and $\overline{y}$ = (1, ... ,1).

Proo]. The proof is analogous to the proof of Theorem 4.3. Just note that the Nash equilibrium of $(\mathcal{G}_\varphi, C_1)$ constructed in the case that φ is satisfiable is also a subgame perfect equilibrium. Q.E.D.

4.3 Secure **equilibria**

For secure equilibria we concentrate on two-player games as it is done in (Chatterjee et al., 2004a), where secure equilibria were introduced. If there are only two players, then there are only four possible payoffs for a secure equilibrium: (0,0), (1,0), (0,1), and (1,1). For each of these payoffs, we aim to characterise the existence of a secure equilibrium that has this payoff and analyse the complexity of deciding whether there exists such an equilibrium.

Lemma 4.6. Let $(\mathcal{G}, va)$ be a two-player game with determined winning conditions. Then $(\mathcal{G}, va)$ has a secure equilibrium with payoff (0,0) if and only if no player has a winning strategy.

Proo]. Clearly, if (σ, τ) is a secure equilibrium with payoff (0,0), then no player can have a winning strategy, since otherwise (σ, τ) would not even be a Nash equilibrium. On the other hand, assume that no player has a winning strategy. By determinacy, there exist a strategy σ of player 1 that guarantees payoff 0 for player 2 and a strategy τ of player 2 that guarantees payoff 1 for player 1. Hence, (σ, τ) is a Nash equilibrium. But it is also secure since every player receives the lowest possible payoff. Q.E.D.

Theorem 4.7. The problem of deciding whether in a two-player parity game there exists a secure equilibriurn with payoff (0,0) is in UP nco-UP. If the number of priorities is bounded, the problem is decidabie in polynomial time.

Proo]. By Lemma 4.6, to decide whether there exists a secure equilibrium with payoff (0,0), one has to decide whether neither player 1 nor player 2 has a winning strategy. For each of the two players, existence (and hence also non-existence) of a winning strategy can be decided in UP n co-UP (Jurdziriski, 1998). By first checking whether player 1 does not have a winning strategy and then checking whether player 2 does not have one, we get a UP algorithm for the problem. Analogously, one can deduce that the problem is in co-UP.

If the number of priorities is bounded, deciding the existence of a winning strategy can be done in polynomial time, so we get a polynomial-time algorithm for the problem. Q.E.D.

Lemma 4.8. Let $(\mathcal{G}, va)$ be a two-player game. Then $(\mathcal{G}, va)$ has a secure equilibrium with payoff (1,0) or payoff (0,1) if and only if player 1 or player 2, respectively, has a strongly winning strategy.

Proo]. We only show the claim for payoff (1,0); the proof for payoff (0,1) is completely analogous. Clearly, if σ is a strongly winning strategy for player 1, then (σ, τ) is a secure equilibrium for any strategy τ of player 2. On the other hand, if (σ, τ) is a secure equilibrium with payoff (1,0), then

for any strategy τ' of player 2 the strategy profile (σ, τ') has payoff (1, 0), hence σ is strongly winning. Q.E.D.

Theorem 4.9 (Chatterjee et al., 2004a). The problem of deciding whether in a two-player parity game there exists a secure equilibrium with payoff (1,0), or payoff (0,1), is co-NP-complete. If the number of priorities is bounded, the problem is in P.

Proof. By Lemma 4.8, deciding whether a two-player parity game has a secure equilibrium with payoff (1,0) or (0,1) amounts to deciding whether player 1 respectively player 2 has a strongly winning strategy. Assume that the game has parity winning conditions Ω_1 and Ω_2. Then player 1 or player 2 has a strongly winning strategy if and only if she has a winning strategy for the condition $\Omega_1 \wedge \neg\Omega_2$ respectively $\Omega_2 \wedge \neg\Omega_1$, a Streett condition. The existence of such a strategy can be decided in co-NP (Emerson and Jutla, 1988). Hence, the problem of deciding whether the game has a secure equilibrium with payoff (1,0), or (0,1), is also in co-NP.

In (Chatterjee et al., 2007) the authors showed that deciding the existence of a winning strategy in a two-player zero-sum game with the conjunction of two parity conditions as its winning condition is already co-NP-hard. It follows that the problem of deciding whether a player has a strongly winning strategy in a two-player parity game is co-NP-hard.

If the number of priorities is bounded, we arrive at a Streett condition with a bounded number of pairs, for which one can decide the existence of a winning strategy in polynomial time (Emerson and Jutla, 1988), so we get a polynomial-time algorithm. Q.E.D.

Lemma 4.10. Let $(\mathcal{G}, v_0)$ be a two-player game with a determined pair (Win., Win2) of prefix-independent winning conditions. Then $(\mathcal{G}, v_0)$ has a secure equilibrium with payoff (1,1) if and only if there exists a play π with payoff (1, 1) such that for all $k < w$ no player has a strongly winning strategy in $(\mathcal{G}, \pi(k))$.

Proof. Clearly, if (σ, τ) is a secure equilibrium with payoff (1, 1), then $\pi := \langle \sigma, \tau \rangle$ is a play with payoff (1,1) such that for all $k < w$ no player has a strongly winning strategy in $(\mathcal{G}, \pi(k))$, since otherwise one player could decrease the other players payoff while keeping her payoff at 1 by switching to her strongly winning strategy at vertex $\pi(k)$.

Assume that there is a play π with payoff (1,1) such that for all $k < w$ no player has a strongly winning strategy in $(\mathcal{G}, \pi(k))$. By determinacy, there exists a strategy σ_1 of player 1 and a strategy τ_1 of player 2 such that σ_1 and τ_1 are retaliating strategies in $(\mathcal{G}, \pi(k))$ for each $k < w$. Similarly to the proof of Lemma 4.1, we define a new strategy σ of player 1 for $(\mathcal{G}, v_0)$

by

$$\sigma(hv) = \begin{cases} \pi(k+1) & \text{if } hv = \pi(0)\dots\pi(k) \prec tt, \\ \sigma_1(h_2 v) & \text{otherwise.} \end{cases}$$

where in the latter case $h = hl\,.\,ba$, and hl is the longest prefix of h still being a prefix of tt. Analogously, one can define a corresponding strategy τ of player 2 for $(\mathcal{G}, va)$. It follows that the strategy profile (σ, T) has payoff (1, 1), and for each strategy σ' of player 1 and each strategy τ' of player 2 the strategy profiles (σ', τ) and (σ, τ') still give payoff 1 to player 2 respectively player 1. Hence, (σ, τ) is a secure equilibrium. Q.E.D.

Theorem 4.11 (Chatterjee et al., 2004a). The problem of deciding whether in a two-player parity game there exists a secure equilibrium with payoff (1,1) is in NP. If the number of priorities is bounded, the problem is in P.

Proof. By Lemma 4.10, to decide whether there exists a secure equilibrium with payoff (1,1), one has to decide whether there exists a play that has payoff (1, 1) and remains inside the set U of vertices where no player has a strongly winning strategy. By determinacy, the set U equals the set of vertices where both players have retaliating strategies. Assume that the game has parity winning conditions Ω_1 and Ω_2. Then a retaliating strategy of player 1 or player 2 corresponds to a winning strategy for the condition $\Omega_1 \vee \neg\Omega_2$ respectively $\Omega_2 \vee \neg\Omega_1$, a Rabin condition. Since positional strategies suffice to win a two-player zero-sum game with a Rabin winning condition (Klarlund, 1992), this implies that the set U also equals the set of vertices where both players have *positional* retaliating strategies.

An NP algorithm for deciding whether there exists a secure equilibrium with payoff (1,1) works as follows: First, the algorithm guesses a set X together with a positional strategy σ of player 1 and a positional strategy τ of player 2. Then, the algorithm checks whether σ and τ are retaliating strategies from each vertex $v \in X$. If this is the case, the algorithm checks whether there exists a play with payoff (1, 1) remaining inside X. If so, the algorithm accepts, otherwise it rejects.

The correctness of the algorithm is immediate. For the complexity, note that checking whether a positional strategy of player 1 or 2 is a retaliating strategy amounts to deciding whether the other player has a winning strategy for the condition $\Omega_2 \wedge \neg\Omega_1$ respectively $\Omega_1 \wedge \neg\Omega_2$, again a Streett condition, in the one-player game where the transitions of player 1 respectively player 2 have been fixed according to her positional strategy. Also, checking whether there exists a play with payoff (1, 1) remaining inside X amounts to deciding whether there exists a winning play in a one-player Streett game, namely the one derived from $\mathcal{G}$ by removing all vertices in X, merging the two players into one, and imposing the winning condition

$\Omega_1 \wedge \Omega_2$. As the problem of deciding the existence of a winning play in a one-player Streett game is decidabie in polynomial time, our algorithm runs in (nondeterministic) polynomial time.

If the number of priorities is bounded, we can actually compute the set U of vertices from where both players have a retaliating strategy in polynomial time, so the algorithm can be made deterministic while retaining a polynomial running time. Q.E.D.

References

Berwanger, D., Dawar, A., Hunter, P. & Kreutzer, S. (2006). DAG-width and parity games. **In** Durand, B. & Thomas, W., eds., *STACS 2006, 23rd Annual Symposium on Theoreiical Aspects of Computer Science, Marseille, France, Februaru* 23-25, *2006, Proceeduiqs,* Vol. 3884 of *Leciure Notes in Computer Science,* pp. 524-436. Springer.

Berwanger, D. & Grädel, E. (2005). Entanglement - a measure for the complexity of directed graphs with applications to logic and games. **In** Baader, F. & Voronkov, A., eds., *Logic fOT Proqramrninq, Artificial Intel!igence, and Reasoning, l1th International Conference, LPAR 2004, Montevideo, Uruguay, Mareli* 14-18, *2005, Proceedinqs,* Vol. 3452 of *Leciure Notes in Computer Science,* pp. 209-223. Springer.

Büchi, .J.R. & Landweber, L.H. (1969). Solving sequential conditions by finite-state strategies. *Tmnsactions of the AmeTican JvIathematical Society,* 138:295-311.

Chatterjee, K, Henzinger, T.A. & Jurdziriski, M. (2004a). Games with secure equilibria. **In** Ganzinger, H., ed., *Proceedinqs of the 19th Annual IEEE Symposium on Logic in Computer Science (LICS 2004),* 14-17 *July 2004, Tu rku , Finland,* pp. 160-169. **IEEE** Computer Society.

Chatterjee, K, Henzinger, T.A. & Piterman, N. (2007). Generalized parity games. **In** Seidl, H., ed., *Foundations of Softmore Science and Computational Siruciures, 10th International Conference, FOSSACS 2007, Held as Pari of the Joint Europeon Conferences on Theoru and Praciice of Softnoare, ETAPS 2007, Bmga, Poriuqal, MaTch 24-ApTil* 1, *2007, Proceedinqs,* Vol. 4423 of *Leciure Notes in Computer Science,* pp. 153-167. Springer.

Chatterjee, K, Jurdziriski, M. & Majumdar, R. (2004b). On Nash equilibria in stochastic games. **In** Marcinkowski, .1. & Tarlecki, A., eds., *Computer Science Logic, 18th International Workshop, CSL 2004, 13th Annual Confer-*

ence of the EACSL, Karpacz, Polarul, September 20-24, 2004, Proceedinqs, Vol. 3210 of *Leciure Notes in Computer Science,* pp. 26-40. Springer.

Emerson, A. & Jutla, C. (1991). Tree Automata, Mu-Calculus and Determinacy. **In** Sipser, M., ed., *Proceedinqs of 32nd Annual IEEE Symposium on Foundations of Computer Science (FOCS)91),* pp. 368-377.

Emerson, E.A. & Jutla, C.S. (1988). The complexity of tree automata and logies of programs (extended abstract). **In** *Proceedinqs of the 29th Annual Symposium on Foundations of Computer Science, FoCS*)88, pp. 328-337. **IEEE** Computer Society Press.

Emerson, E.A., Jutla, C.S. & Sistla, A.P. (1993). On model-checking for fragments of μ-calculus. **In** Courcoubetis, **C.,** ed., *Computer Aided Verification, 5th International Conference, CA V*)93, *Elounda, Creece, June* 28 - *July* 1, 1993, *Proceedinqs,* Vol. 697 of *Leciure Notes in Computer Science,* pp. 385-396. Springer.

Emerson, E.A. & Lei, C.-L. (1985). Modalities for model checking: Branching time strikes back. **In** Van Deusen, M.S., Galil, Z. & Reid, E.K., eds., *Conference Record of the Twelfth ACM Symposium on Pruiciples of Progmmming Languages. Extended Abstracts of Papers Presenied at the Symposium. Monieleone Hotel, New Orleans, Louisiana,* 14-16 *January 1985,* pp. 84-96. ACM Press.

Grädel, E. (2007). Finite model theory and descriptive complexity. **In** *Finite Model Theory and lts Applications,* pp. 125-230. Springer.

Hunter, P. & Kreutzer, S. (2007). Digraph measures: Kelly decompositions, games, and orderings. **In** Bansal, N., Pruhs, K. & Stein, **C.,** eds., *Proceedinqs of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January* 7-9, *2007,* pp. 637-644. SIAM.

Jurdziriski, M. (1998). Deciding the winner in parity games is in **UP** ∩ co-UP. *Information Processine Letters, 68(3):119-124.*

Jurdzinski, M., Paterson, M. & Zwick, U. (2006). A deterministic subexponential algorithm for solving parity games. **In** Stein (2006), pp. 117-123.

Klarlund, N. (1992). Progress measures, immediate determinacy, and a subset construction for tree automata. **In** Scedrov, A., ed., *Proceedinqs of the Seventh Annual IEEE Symposium on Logic in Computer Science, Santa Cruz, California, June* 22-25, 1992, pp. 382-393. **IEEE** Computer Society Press.

Kuhn, H.W. (1953). Extensive games and the problem of information. **In** Kuhn, H.W. & Tucker, A.W., eds., *Contribuiione to the Theoru of Games, vol. IJ,* Vol. 28 of *Annals of Mathematical Studies,* pp. 193-216. Princeton University Press.

Martin, D.A. (1975). Borel determinacy. *Annals of Mathematics,* 102(2):363-371.

Martin, D.A. (1990). An extension of Borel determinacy. *Annals of Pure and Applied Logic, 49(3):279-293.*

Mostowski, A.W. (1991). Games with fotbidden positions. Technical Report 78, Instytut Matematyki, Uniwersytet Cdariski, Poland.

Muller, D.E. & Schupp, P.E. (1995). Simulating alternating tree automata by nondeterministic automata: New results and new proofs of the theorems of Rabin, McNaughton and Safra. *Theoreiical Computer- Science, 141(1-*2):69-107.

Nash, Jr., J.F. (1950). Equilibrium points in N-person games. *Proceedings of the National Academy of Sciences of the United States of Amer-ica,* 36(1):48-49.

Obdrzálek, .1. (2003). Fast mu-calculus model checking when tree-width is bounded. **In** Hunt, Jr., W.A. & Somenzi, F., eds., *Computer- Aided Ver-ification, 15th International Confer-ence, CA V 2003, Boulder-, CO, USA, July 8-12, 2003, Proceedirujs,* Vol. 2725 of *Leciure Notes in Computer- Science,* pp. 80-92. Springer.

Obdrzálek, .1. (2006). DAG-width - connectivity measure for directed graphs. **In** Stein (2006), pp. 814-821.

Obdrzálek, .1. (2007). Clique-width and parity games. **In** Duparc, .1. & Henzinger, T.A., eds., *Computer- Science Logic, 21st International Workshop, CSL 2007, 16th Annual Confer-ence of the EACSL, Lausanne, Switzerland, September-* 11-15, *2007, Proceedinqs,* Vol. 4646 of *Leciure Notes in Computer- Science,* pp. 54-68. Springer.

Selten, R. (1965). Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit. *Zeitschrift fiir die gesamte Staatswissenschajt, 121:301-*324 and 667-689.

Stein, **C.,** ed. (2006). *Proceedinqs of the Seventeenth Annual ACM-SIAM Symposium on Discr-ete Alqorithme, SODA 2006, Miami, Ftorida, USA, Januari]* 22-26, *2006.* ACM Press.

Thomas, W. (1995). On the synthesis of strategies in infinite games. **In** Mayr, KW. & Puech, C., eds., *STACS* 95, *12th Annual Symposium on Theoreiical Aspects of Computer Science, Murucli, Germany, Mareli 2-4, 1995, Proceeduiqs,* Vol. 900 of *Leciure Notes in Computer Science,* pp. 1-13. Springer.

Ummels, M. (2006). Rational behaviour and strategy construction in infinite multiplayer games. **In** Arun-Kumar, S. & Garg, N., eds., *FSTTCS 2006: Foundations of Software Technology and Theoreiical Computer Science, 26th International Conference, Kolkata, India, December* 13-15, *2006, Proceedinqs,* Vol. 4337 of *Leciure Notes in Computer Science,* pp. 212-223. Springer.

Ummels, M. (2008). The complexity of Nash equilibria in infinite multiplayer games. **In** Amadio, KM., ed., *Foundations of Software Science and Computational Struciures, l1th International Conference, FOSSACS 2008, Held as Part of the Joint Europeon Conferences on Theory and Praciice of Software, ETAPS 2008, Budapest, Hungary, Mareli* 29 - *April* 6, *2008. Proceedinqs,* Vol. 4962 of *Leciure Notes in Computer Science,* pp. 20-34. Springer.

Zielonka, W. (1998). Infinite games on finitely coloured graphs with applications to automata on infinite trees. *Theoreiical Computer Science,* 200(1-2):135-183.

Games in Language

Gabriel Sandu

institut d'histoire et de philosophie des sciences et des techniques (IHPST)
Université Paris 1 Panthéon Sorbonne
13, rue du Four
75006 Paris, France
sandu@rnappi.helsinki.fi

Abstract

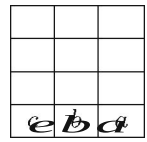
The paper deals with one but widespread natural language phenomenon: discourse anaphora. I make a bridge between three game-theoretical approaches to this phenomenon: (a) Hintikka's game-theoretical semantics (GTS); (b) Dekker and van Rooij's application of strategie games to the underspecification of anaphora; and (c) Abramsky's Dynamic Game Semantics. I see (b) as leading to a 'gamification' of a phenomenon which GTS (and other approaches such as Government and Binding Theory) sawasbelonging to semantics or syntax. I see (c) as solving some problems left open by the 'subgame interpretation' of GTS. The present paper draws some comparisons and methodological refiections prompted by the remarks of my commentator, Paul Dekker.

1 Rules **and** language games

One traditional view in philosophy and linguistics is that without rules of usage common to the speaker and the listener, communication would be impossible. According to it, every linguistic expression has a meaning which is determined by the rules for its correct use. This obviously brings language and games together, for it is in the latter that rules are explicitly given. Here are two examples of language games which iliustrate in a very simple and ideal way how a communication language could emerge out of *language games.* They are due to the Finnish logician Erik Stenius who thought that they are typical examples of the view of language advocated by Ludwig Wittgenstein in his later period.

The *Garden Game* is played by a gardener *A* and his assistant *B.* There are pieces in the game, the letters *'a', 'b', 'c'* , *'P'* and *'Q',* and a ftower bed divided into squares as in the figure below. In every square there is a plant.

Krzysztof R. Apt, Robert van Rooij (eds.). *New Perspectives on Games and Interaction.* Texts in Logic and Games 4, Amsterdam University Press 2008, pp. 179-196.



The game amounts to this: Every day B writes on a piece of paper the letters *'a'*, *'b'*, *'c'*, and to the left of any of these letters he writes either the letter *'P'* or the letter *'Q'*, according to whether the plant in the square corresponding for that day to the lower-case letter is in ftower or not. For instance, if in the rectangle for that day, the plant next to the path is in ftower, whereas the two ot hers are not, B will write:

$$Pa\ Qb\ Qc$$

The teaching of the game is done by simple gestures of approval and disapproval depending on whether B writes the correct tokens on the piece of paper or not.

Once the assistant masters the Garden Game, A and B move to play the *Report Game*. A does not need to accompany B any longer to the ftower-bed. A now partakes in the game only by receiving the written tokens from B. If B really follows the rules of the game, A can read off certain facts from what B has written.

It is obvious that by means of the report game, A and B have created a small language for communication: *'a'*, *'b'*, and *'c'* are used to denote certain squares, *'P'* and *'Q'* express certain properties, etc. These symbols have acquired a meaning.

Stenius' language games had more of a philosophical purpose, namely to give concrete examples of Wittgensteinian language-games. They inspired David Lewis who formulated them in a more precise way, using notions in classical game theory. **In** doing so, Lewis thought to respond to a challenge launched by Quine. The latter regarded with distrust conventional views of language and doubted that one can give a coherent account of how communication takes place without presupposing already some degree of linguistic competence. **In** response to Quine's challenge, Lewis formulated signalling games, that is, communication games played by two players, the Sender and the Receiver, the former sending messages or signals about the situation he or she is in, and the latter undertaking a certain action aftel' receiving it. The point to be emphasized is that the messages do not have a prior meaning: whatever meaning they are going to acquire, it will be the result of the interactive situation in the signalling game, or in other terms, they will be optimal solutions in the game. Let us have a closer look at the game-theoretical setting.

2 Strategie games and Nash equilibria

Let us shortly reeall some basic notions in classical game theory. We shall use as an example *Prisoner's dilemma:* Two criminais, 1 and 2, are interrogated in separate cells. If they both confess, each of them will be punished to stay 3 years prison. If only one of them confesses, he will be free while the other will be punished with 4 years in prison. If none of them confesses, each will stay 1 year in prison. The picture below depiets the choices and payoffs of the players:

	D	C
D	$(-1,-1)$	$(-4,0)$
C	$(0,-4)$	$(-3,-3)$

Prisoner's dilemma

'D' stands for "don't confess" and *'C'* stands for "confess."

As we see, a complete description of a strategic game with two players, requires a list of the players' action repertoires Al and A_2, and a specification of their utility functions u_1 and *ua-* The function u_i specifies, for each possible sequence of choices *(a, b)* (action profile) i's payoff. **In** our example we have: $u_1(D, D) = -1$, $u_1(D, \mathrm{C}) = -4$, $u_2(D, \mathrm{C}) = 0$, etc.

Given a strategic game, we are interested in optimal plays of the game, that is, in every player's action being the best response to the actions of his opponents. Consider a simple arbitrary action profile *(a, b).* **It** is a Nash equilibrium in the strategic game if none of the players would have been better off by making a different choice:

$$u_1(a, b) \geq u_1(c, b), \text{ for any } \mathrm{c} \text{ in } Al$$
$$u_2(a, b) \geq u_2(a, d), \text{ for any } d \text{ in } A_2.$$

Thus in the Prisonner's dilemma game, *(C,* C) is a Nash equilibrium, but there is no Nash equilibrium in the Matching pennies game. There are games which have two Nash equilibria, like the one below where both *(L, L)* and *(R, R)* are Nash equilibria.

	L	R
L	$(1,1)$	$(-1,-1)$
R	$(-1,-1)$	$(1,1)$

3 Signalling games

3.1 Lewis' signalling games

Stenius' games are games of complete information in the nontechnical sense that when the gardener teaches his assistant the Garden game, both the

Gardener and the assistant know the situation each of the letters *'a', 'b'* or *'c'* is supposed to be associated with. They both see whether, for a particular day, a plant is in the corresponding square or not. Lewis's signalling games abandon this assumption. Their aim is to model the kind of conventions that associate meanings to linguistic forms. This is achieved by Lewis's signalling systems, that is, ideal communicative situations in which the communicator or Sender *(S)* sends a message or a signal in a particular situation, in order to get a Receiver *(R)* to undertake a course of action.

One of Lewis's examples is that of a truck driver *A* trying to reverse. Ris assistant *B,* who is behind the truck, helps him to do that by making two kinds of gestures: If there is room to reverse, she is making the beckoning gesture; otherwise she is showing hands with palms outwards. The driver is taking a course of action conditional on his observations: if *B* makes the beckoning gestures, then he reverses; otherwise he stops.

It is straightforward to put this game in the above format of strategic games. The Sender's *(B's)* choices are functions $S : T \to F$, where T is a set of situations ("there is place" and "there is no place") and F a set of messages ("beckoning" message and "palms outwards" message), and the Receiver's (A's) choices are functions $R : F \to Act$, where *Act* is a set of actions ("reverse" and "stop"). There are 4 strategies for the Sender and 4 strategies for the Receiver.

Each simultaneous choice by the two players gives rise to an action profile *(S, R).* Utility functions u_S and u_R calculate the payoffs $u_S(S, R)$ and $u_R(S, R)$ for each action profile *(S, R).* Each of them will sum up, in a certain way, the payoffs of each state *t*

$$u_S(t, S, R), \quad u_R(t, S, R).$$

Given that the signalling games are cooperative games, in that both players try to achieve a common goal, communication, we take $u_S(t, S, R)$ and $u_R(t, S, R)$ to be equal in such a way that their value is 1 when communication is achieved, and 0, otherwise:

$$u_S(t, S, R) = u_R(t, S, R) = \begin{cases} 1 & \text{if } R(S(t)) = t \\ 0 & \text{otherwise.} \end{cases}$$

Finally, the expected utilities $u_S(S, R)$ and $u_R(S, R)$ are certain sums of *uit; S, R)* for all situations *t.* For the Sender, $u_S(S, R)$ is simply $\sum_{t \in T} P(t)$ x $u_S(t, S, R)$, where P is the prior probability distribution over the states in T (which is common knowiedge). But for the Receiver things are a bit more complicated. Re knows only the message but not the situation the Sender is in when sending it. For this reason, his expected utilities will be conditional on equivalence classes of states. We let S_t be the information

state the Receiver is in aftel' the Sender chooses strategy *W:*

$$W_t = \{t' : W(t') = W(t)\}.$$

Finally we define

$$u_R(S, R) = \sum_{t' \in W_t} P(t'/W_t) \times u_R(t, S, R).$$

A strategy profile *(S,R)* forms a (Bayesian) Nash equilibrium ifnone ofthe two players would be (strictly) better off by making another decision.

It is games of this kind which are called by Lewis signalling systems and which can be associated, as pointed out earlier, with linguistic meanings. As these games have several appropriate Nash equilibria, the choice between them is conventional. The interesting question is of course which one of them is to be chosen as the conventional one. Lewis's weil known answer is that it is the more salient one which is to be selected, but the question is of course what makes one separating Nash equilibrium more salient than another. **It** seems that Quine's challenge to give an explanatory account of conventional meanings without presupposing some kind of linguistic competence has not yet been completely met.

Now when such a signalling system occurs recurrently in a population, a convention for such a problem leads to a simple language. The signals used acquire meaning in virtue of the fact of being associated with particular states in the world and actions, and this association being common knowiedge. **If** in the truck game, the selected equilibrium is the pair *(S, R),* where

- *S* is the strategy: make the beckoning gesture, if place to reverse, and show the palms outwards, otherwise,
- *R* is the strategy: reverse, if shown the beckoning gesture, and stop otherwise

then Lewis would say that the beckoning gesture means 'Place to reverse, and the showing palms outwards gestures means 'No place to reverse, and this meaning has been settled by convention. **In** other words, *S(t)* means *t.*

(Notice however, that for this to work, the game has to have separated equilibria: *S* sends different messages in different states.)

3.2 Parikh's interpretation games

Parikh's games share the same format as their Lewisian relatives, with some variations. They are interpretation games: A Sender is sending messages that the Receiver tries to interpret. **It** is thus more appropriate to see the Sender's choices as functions $S : T \to F$ from a set of states to a set of

linguistic forms, and the Receiver's choices as functions $S : F \rightarrow T$. There are two other features which distinguish Parikh's games more essentially from their Lewisian relatives. For one thing, messages have costs which enter into the determination of payoffs. For another thing, some of the messages possess already a meaning. The typical situation is one in which the Sender has three messages, i, f' and I'', with f' and f'' possessing already a meaning: f' can be used only in *ti* ("f' means *ti"*) and f'' means *til.* i on the other side can be used in two distinct situations: the Sender may send it when she is in *t* to communicate that she is in *t;* and she may send it when she finds herself in *ti,* to communicate that she is in *ti.* Otherwise the setting is very much like in Lewis's signalling games. The Sender knows also the situation she is in, unlike the Receiver who associates with the state ta probability of 0.8 and with the state *t'* a probability of 0.2.

To give an idea of what is going on, here are two tables with the possible strategies of the two players

Sender:

	t	*ti*
Sl	i	f'
S2	i	i
S3	f''	i
S4	f''	f'

Receiver:

	i	f'	f''
Hl	*t*	*ti*	*t*
H_2	*ti*	*ti*	*t*

The fact that messages have costs leads to a different calculation of the payoffs of the two players than in the previous case. The expected utilities $u_S(S, R)$ and $u_R(S, R)$ are calculated as above, but for $u_S(t, S, R) = u_R(t, S, R)$, the complexity of the messages matters. Let us assume that f' and t'' are more complex than i. Let Compl(J) = 1, and Complrj") = Complrj'") = 2. Under the assumption that the Sender prefers to send cheaper messages to expensive ones, $u_S(t, S, R)$ and $u_R(t, S, R)$ are now redefined as:

$$u_S(t, S, R) = u_R(t, S, R) = \begin{cases} \frac{1}{\mathrm{Compl}(S(t))} & \text{if } R(S(t)) = t \\ 0 & \text{otherwise.} \end{cases}$$

The strategie game is now depicted in the table below:

	Hl	H_2
Sl	(0.9,0.9)	(0.1,0.1)
S2	(0.8,0.8)	(0.2,0.2)
S3	(0.4,0.4)	(0.6,0.6)
S4	(0.5,0.5)	(0.5,0.5)

It can be checked that the game has only two Nash equilibria, (Sl, H_1) and $(S3, H_2)$, but unlike Lewis, Parikh would not say that the choice be-

tween them is conventional, but uses the notion of Pareto optimality to choose between the two:

- An action profile *(S, R)* is Pareto more optimal than *(Sl, Rl)* if $S > Si$ and $R > Rl$.

The optimal solution ofthe game is thus *(Sl, Hl)'* (Cf. also van Rooij, 2002.)

4 Signalling games and Gricean pragmatics

In the Parikhian game we considered above, *t* is more likely than *ti* and in each state there are at least two forms that could express it. But because *f* is the less "complex" of the two forms, then Parikh's theory prediets that *t* will be expressed by *f* and *ti* by the more complex expression. The introduetion of costs leads to an ordering of messages. The game theoretical setting forces the Sender to consider alternative expressions he or she could have used together with their costs.

Van Rooij (2002) observes that a similar conclusion is reached by Blutner's bidimensional optimality theory (OT). We do not have the space here to enter into details. Suffice it to say that in this theory, for the hearer to determine the optimal interpretation of a given form, he must also consider alternative expressions the speaker could have used to express that meaning (interpretation). And the Speaker is foteed to consider the optimal form to express a particular meaning.

Blutner's bidemensional OT has been given a game-theoretical interpretation in (Dekker and van Rooij, 2000). According to it, communication (information exchange) is represented as a strategic game between speaker and hearer in the marmer described above. For a detailed comparison between Parikhian, Lewisian and the games introduced in (Dekker and van Rooij, 2000), the reader is referred to (van Rooij, 2002).

One of the interesting things in the game-theoretical setting of (Dekker and van Rooij, 2000) is its conneetion to Gricean Pragmatics, where the focus is not so much on the question of how expressions acquire their meanings, but rather on the distinction between what is said and what is conveyed or implied. The former is more or less conventional, semantic meaning, while the latter, although not explicitly stated, is something the speaker wants the Hearer to understand from what is said. **In** a seminal paper, Paul Grice tried to account for such pragmatic inferences in terms of maxims of conversations: He thought that the Hearer is able to associate the right interpretation with a particular assertion on the basis of an algorithm which computes it out of the relevant maxims. **In** optimality theory, the maxims lead to a ranked set of constraints which allow one to select the optimal syntactic form-interpretation pair. Dekker and van Rooij (2000) observe that the ranking of constraints in optimality theory has the same effect as the

ranking of action profiles in strategic games of the sort considered above: The Speaker wants to communicate a certain meaning and she has to choose a suitable formulation for it, while the Hearer wants to associate a correct interpretation with that form by considering all the other alternatives the speaker might have used.

Here are a couple of examples which have a bearing on the interpretation of anaphora.

Example 4.1 (Dekker and van Rooij, 2000).

John is happy. He smiles. (1)

A girl came in. She smiles. (2)

Bill tickled John. He squirmed. (3)

There are two principles at work here:

(a) The principle of salience: pronouns refer back to the (denotation of the) salient expression (subject expression) of the previous sentence, i.e., 'John', 'a girl' and 'BilI'.

(b) The naturalness principle: because of semantical facts associated with 'tickled', it is natural that in (3) the pronoun refers back to John.

In (3), principle (a) is overruled by principle (b) explaining thus why the correct interpretation to be associated with (3) is the one in which the head of the anaphorical pronoun is 'John' and not 'BilI'.

Example 4.2 (Hendriks & de Hoop, Dekker & Van Rooij).

Often when I talk to a doctor., the doctor{i,j} disagrees with him{i,j}' (4)

There are two general principles at work here:

(i) If two arguments of the same semantic relation are not marked as being identical, interpret them as being distinct.

(ii) Don't Overlook Anaphoric Possibilities.

Here (ii) is overruled by (i), explaining why the head of the anaphoric pronoun cannot be 'the doctor'.

I will take up in the next section an analysis of these examples in Cametheoretical semantics (CTS). The comparison is instructive, for CTS tries to account for the same phenomena in semantical terms. I will then sketch yet another approach to these examples in *Govemment and Binding* Theory.

5 Game-theoretical semantics (GTS)

5.1 Semantica1 games for quantifiers

Unlike communication games whose task is to model how expressions of the language acquire an interpretation, in semantical games associated with natural or formallanguages, it is presupposed that certain basic expressions and sentences of a language already have an interpretation. What is wanted, instead, is a way to give the truth-conditions of more complex sentences of the language by reducing them to the truth-conditions of the known ones. Here is a typical example from the mathematical vernacular.

A function *y = f(x)* is continuous at *Xa* if *given a* number α however smali, *we can find* ε such that *If(x) - f(xa)1* $<$ α, *given any x* such that $Ix\text{-}xal<c$.

The game-theoretical analysis is supposed to throw light on the interpretation of the expressions "we can find" and "given any" assuming that the interpretation of the other expressions $(11, \dot{l}, <)$ is already fixed (by a background model). This is done by a semantical game played by two players, the existential $\exists$ and respectively the universal player $\forall$, both choosing individuals from the relevant universe of discourse. The choices of the first correspond to "we can find" and those of the second to "given any". Unlike strategic games, which are one shot games, semantical games have a sequential element with later choices depending on earlier ones. Thus a play of the present game consists of a sequence of three choices of individuals in the universe: first $\forall$ chooses an in α, then $\exists$ chooses ε, and finaliy $\forall$ chooses *x.*

Given the sequential nature of semantical games, it is more appropriate to exhibit them, not in strategic, but in extensive form

$$G = (N, H, P, (u_i)_{i \in N})$$

where N is a coliection of players, H is a set of histories, P is a function attaching to each non-maximal history the player whose turn is to move, and u_i is the utility function for player i, i.e., a function which associates with each maximal history in H a payoff for player i. Each maximal history represents a play of the game, at the end of which each of the players is given a payoff. The games are strictly competitive û-sum games: for each maximal play, one of the player is winning and the other is loosing. The play is a win for the existential player if the terminal formula is true (in the background model). Otherwise it is a win for the universal player.

The crucial notion is that of a strategy for a player, a method which gives him or her the appropriate choice depending on the elements chosen earlier in the game. Such a strategy is codified by a mathematical function *9* which takes as arguments the partial histories $(aa, \ldots, a_{n-1})$ in H where the player is to move, and gives her an appropriate choice *g(aa, ... ,an-l).*

9 is a winning strategy if it guarantees him a win in every maximal play in which she uses it.

In our particular example, a maximal play of the game is any sequence (α, ε, x) chosen be the players in the order specified above. **If** the chosen elements stand in the appropriate relations, i.e.,

$$\text{if } (|x - xo| < \varepsilon) \text{ then } (|f(x) - f(xo)1 < \alpha), \tag{5}$$

then we deciare the play to be a win for $\exists$ and a loss for $\forall$. Otherwise, it is a win for $\forall$ and a loss for $\exists$. A strategy for $\exists$ is any function *9* whose arguments are all the individuals α chosen by $\forall$ earlier in the game. *9* is a winning strategy if,

$$\text{if } (|x - xo| < g(\alpha)) \text{ then } (|f(x) - f(xo)1 < \alpha). \tag{6}$$

Thus the continuity of a function has been characterized by the truth of the second-order sentence:

$$\exists g \forall \alpha \forall x[(|x - xo| < g(\alpha)) \rightarrow (|f(x) - f(xo)1 < \alpha)] \tag{7}$$

5.2 Anaphora and the subgame interpretation

GTS has been extensively applied to the analysis of anaphoric descriptions and anaphoric pronouns. For a simple illustration, consider our earlier sentence (2), reproduced here as (8):

A girl came in. She is happy. (8)

The semantical game associated with (8) is completely analoguous to the quantifier games (in fact the game involves quantifiers), except that

- Games are divided into subgames, one for each subsentence of (8).
- The rules of the game are extended to cover also anaphoric pronouns.

In our example, 'She' prompts a move by the existential player who must now choose the unique individual available from the earlier subgames.

The only such individual is the one introduced for "A girl", and thus the game-theoretical analysis correctly prediets that (8) may receive an interpretation in which "A girl" is the head of "She".

There are two main problems here.

The first one has to do with semantical games being undetermined by the game rules. Here are a couple of examples.

Λ problem of underdetermination ariscs when there is more than one individual available from earlier subgames. This is the case with our earlier example (3) reproduced here as (9):

Bill tickled John. He squirmed. (9)

In order to obtain the right interpretation of this sentence, Hintikka and Sandu (1991) make use oflexical rules which eneode the semantic properties of lexical words like "tickled" and "squirmed". Disregarding some of the details, such rules have the effect that only the individual denoted by "John" remains available for later choices. The mechanism by which Bil! is excluded has to do with the fact that in virtue of the semantic properties of "tickled", Bil! is assigned an agentive and John a patient role. The semantic properties of 'squirmed', on the other side, require an argument which has the patient role, that is, John.

The second example concerns the correct interpretation of (4) reproduced here as (10):

Often when I talk to a doctor., the doctor{i,j} disagrees with him{i,j}. (10)

Here we need additional principles which limit the sets which choices corresponding to anaphoric pronouns can be made from: the individual chosen as the value of an anaphoric pronoun cannot be the same individual which has been chosen earlier as the value of an expression in the same clause as the anaphoric pronoun. **In** the case of him{i,j} the rule has the effect that the chosen individual must be disjoint from the individual chosen earlier for "the doctor."

We postpone the discussion of the second problem for later on. For the moment let us take stock.

6 Three kinds of explanations

There is a phenornenon of undeterrnination in sentences (3)-(4). **In** GTS it is manifest at the level of the application of the game rules to anaphoric pronouns: the rules do not determine completely which individual is to be chosen in the play of the games. For the choice to be settled, we need to supplement them with additional principles. But one must be clear of what is going on here: the additional principles do not, properly speaking, have any game-theoretical content. **In** the play of the relevant semantical game, it is enough for the players to lay back and wait for syntactical principles or thematical roles associated with lexical items to do their job.

In Dekker and van Rooij's analysis, there is an undetermination of the truth-conditions of sentences (3)-(4): their semantic content ("w hat is said") must be supplemented by additional principles motivated by Grice's principles of rational communication (Grice's maxims of conversation). These principles lead to an appropriate ranking of the form-interpretation pairs which lend themselves to a game-theoretical interpretation in terms of strategie games of communication. The correct intepretation is eventual!y obtained as the solution of such a game.

If one looks at the two additional sets of principles in GTS and in Dekker and van Rooij's analysis, one notices that they are very much variations of each others. For instance, the principle which says that

- the choice of (the value of) an anaphoric pronoun cannot come from the same set as that from which an individual chosen as the value of an expression being in the same clause comes from,

is clearly the counterpart in GTS of the maxim (i) relativized to anaphoric pronouns (as distinguished from refiexives).

The main difference between these two approaches lies, as I see it, in what is taken to belong to semantics as opposed to pragmatics. This is an ongoing debate. The motivation behind the GTS treatment of (3)-(4) has to do with a familiar conception according to which, if asemantic category (indexicals, refiexives, etc) have a syntactical counterpart, then, it should be treated semantically, even if underspecified. By relegating underspecified anaphoric pronouns to the realm of pragmatic phenomena pertaining to strategic communication, Dekker and van Rooij brought this kind of undetermination under the incidence of strategic communication games, à la Parikh. The gain in game-theoretical content seems to be obvious. The only qualms I have is about this kind of undetermination belonging to the pragmatics of *communication*. I won't settie this matter here. I will shortly describe, instead, a third approach to the very same phenomena: *Government and Binding* (GB) Theory. Initially GB tries to explain the mechanism underlying the behaviour of pronouns in natural language in purely syntactical terms, using notions like C-command, governing category, etc. Again, I will be very sketchy.

GB theory contains a class of binding principles like:

- A refiexive pronoun must be bound in its local domain
- A pronominal must be free in its local domain

(Chomsky, 1986, p. 66) where the notion of "free", "bound", "Iocal domain" are syntactical notions matching familiar notions in logic: free variabie, binding, scope, etc.

Applied to our earlier examples, these principles predict that in (4) 'him' and 'the doctor' cannot be coindexed. Interesting enough, these principles are not sufheient to yield the correct interpretation for (3). They must be supplemented by "thematic relations" of the same sort we used in the game-theoretical analysis above which, in the end, had the effect of allowing "he" in (3) to be coindexed with "John". (A detailed explanation is contained in Hintikka and Sandu, 1991.)

7 Semantics for dynamic game languages

One can interpret the two preceding sections in the following way: discourse anaphora brought along a phenomena of underdetermination in CTS. The solution proposed inside the CTS community was to enrich that framework with more syntactical or lexical principles, which lacked, however, a game-theoretical motivation. I regard Dekker and van Rooij's analysis as an alternative to that move. **It** certainly has the merit of making a bridge between phenomena which traditionally were regarded as syntactical or semantical at most, and issues in strategie games of communication.

In this section lintend to show how certain contemporary developments in dynamic game semantics can be seen as solving another problem (the second problem I mentioned above) in the CTS treatment of anaphoric pronouns.

The problem I have in mind appears in any of our earlier examples, say (1). **In** the (Carlson and Hintikka, 1979) and (Hintikka and Kulas, 1985) subgame interpretation of this sentence, a subgame is played first with the first subsentenee of (1). **If** the existential player has a winning strategy in it, then the players move to play the second subgame, remembering the winning substrategy in the first subgame. For (1) this substrategy reduces to the choice of an individual, who is then available for being picked up as the value of the anaphoric pronoun in the second subgame (cf. above).

This interpretation is problematic in at least one respect: the anaphoric resolution is dependent on truth, while things should go the other way around. The anaphoric link between the pronoun and 'John' or 'A girl' is established only aftel' the truth of the first subsentenee has been established. This is manifest in the fact that what is "transmitted" from one subgame to another are *winning* strategies, and not strategies *simpliciter*. There is a general agreement, however, that the dependenee between the anaphor and its head in all these sentences is prior and independent of truth. This brings me to the purpose of this section: to present aDynamie Came Language and a compositional intepretation of it, due to Abramsky (2006) in which the strategies of a given game are determined compositionally from the strategies of subgames.

I restriet my presentation only to a subfagment of Abramsky's language which is given by the clauses:

$$\varphi := 1 \mid \mathrm{At} \mid Q_\alpha x \mid \varphi.\psi \mid \varphi \| \psi,$$

where 1 is a propositional constant, At stands for atomie first-order formulas, and $\varphi.\psi$ and $\varphi \| \psi$ represent the operations of sequential and parallel composition, respectively.

Foliowing insights from Dynamic Game Logic (van Benthem, 2003), Abramsky represents quantifiers by

$$Q_\alpha x := Q_\alpha x.1$$

This syntax is very powerful, it turns out that it can represent paraliel sequences of quantifiers, like Henkin's branching quantifier

$$\begin{pmatrix} \forall x & \exists y \\ \forall z & \exists w \end{pmatrix} A(x, y, z, w) \Longleftrightarrow \forall x.\exists y \| \forall z.\exists w.A(x, y, z, w).$$

Two interpretations are given for this language:

(i) Statie Semantics assigns a game to each formula in such a way that every Iogical operator is interpreted by an operation on games.

(ii) Dynamic semantics assigns to each game and each player associated with a formula, a strategy in such a way that the strategies for complex formulas are obtained compositionaliy from the strategies of subformulas.

7.1 Statie semanties (game-semanties)

The games in the statie semantics are actualiy comparable with games in extensive form. They have a sequential nature, but their structure is more fine grained so that games associated with complex formulas are formed compositionaliy from simpler ones. A game has the form

$$G = (M, \lambda_M)$$

where M is a *Concrete Data Struciure* (CDS), that is, a quadruple

$$M = (CM, V_M, D_M, \vdash_M)$$

such that:

- CM is a set of celis
- V_M is a set of values
- DM is a set of decisions, $DM \subseteq CM \times VM$
- $\vdash_M$ is the enabling relation which determines the possible flow of events (decisions)

and

$$\lambda_M : CM \rightarrow \{\exists, \forall\}$$

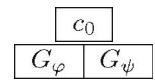
As pointed out, complex games are build up compositionaliy from simpier ones according to the foliowing rules (again, we restriet ourselves to few cases relevant for later examples).

Rules for atomie games:

- $G_{\mathrm{I}} = ((0,0,0,0),0)$
- $G_{\mathrm{At}} = ((0,0,0,0),0)$

The formulation of these two games indicate the fact that there are no moves associated with them.

- The rule for $G_{\varphi\vee\psi}$. First let $G_\varphi = ((C_M, V_M, D_M, \vdash_M), \lambda_M)$ and $G_\psi = ((C_N, V_N, D_N, \vdash_N), \lambda_N)$. Then $G_{\varphi\vee\psi} = ((C_M \cup C_N \cup \{\mathrm{co}\}, V_M \cup V_N \cup \{\mathrm{I}, 2\}, D_M \cup D_N \cup \{(\mathrm{co}, 1), (\mathrm{co}, 2)\}, \vdash_{\varphi\vee\psi}), \lambda_{\varphi\vee\psi})$



In other words, the combined game has a new position (cell), Co which can be filled in by the existential player with either 1 or 2.

From now on I restriet myself to the graphical representation of the games.

- $G_{\exists x\varphi}$:

The complex game has a new position Co which can be filled in by Eloise with any element in the domain.

- $G_{\varphi.\psi}$:

That is, G_φ is played first followed by a play of G_ψ. This fact is encoded in the enabling relation which is now defined by

$$r \vdash_{\varphi.\psi} c \;\Leftrightarrow\; r \vdash_{G_\varphi} c \vee (r = s \cup \Delta \wedge s \in \mathrm{Max}(G_\varphi) \wedge \Delta \vdash_{G_\psi} c).$$

which says that when a maximal play of G_φ is reached, then the play is continued as in G_ψ.

Once the games have been defined, strategies may be built up compositionally from substrategies:

- The strategy set in $G_{\exists x\varphi}$:

  We fix σ, a strategy of $\exists$ in G_{φ}, and a an individual in the domain. We may think of σ as a set which is a (partial) function. We define (roughly) the notion of a strategy for the existential player:

$$\mathrm{up}_a(\sigma) = \{(c_0, a)\} \cup \sigma.$$

  That is, uPa(σ) is formed by adding to the strategy σ in the game φ the new position Co filled in with the individual *a.*

  Finally a strategy set for the existential player is the collection of all strategies

$$\mathrm{Str}_{\exists}(G_{\exists x\varphi}) = \{\mathrm{up}_a(\sigma) : a \in D \wedge \sigma \in \mathrm{Str}_{\exists}(G_{\varphi})\}$$

- The strategy set in $G_{\varphi;\psi}$:

  We fix σ, a strategy of $\exists$ in G_{φ}, and a family $(\mathsf{Ts})_{s\in \mathrm{Max}(G_{\varphi})}$ of strategies of $\exists$ in G_{φ}, indexed by maximal histories of G_{φ}. (We reeall that $G_{\varphi.\psi}$ consists of plays of G_{φ} followed by plays of G_{ψ}.) A strategy for the existential player in the game is

$$(\sigma \cdot (\tau_s)_s)(t) = \begin{cases} \sigma(t) & t \text{ is a nonmaximal state in } G_{\varphi}, \\ \sigma(t) \cup \tau_{\sigma(t)}(\varnothing) & \sigma(t) \text{ is a maximal state in } G_{\varphi}, \\ s \cup \tau_s(u) & t = s \cup u,\ s \in \mathrm{Max}(G_{\varphi}),\ u. \text{ is a state in } G_{\psi}. \end{cases}$$

  In other words, $\exists$ plays according to σ in G_{φ}, and when a maximal state t has been reached, she plays according to $\tau_{\sigma(t)}$.

  Finally the strategy set for the existential player is the collection of all strategies:

$$\mathrm{Str}, (G_{\varphi;\psi}) = \{(\sigma \cdot (Ts)s) : \sigma \in \mathrm{Str}_{\exists}(G_{\varphi}) \wedge \forall s \in \mathrm{Max(Grp)}(\mathrm{T_s} \in \mathrm{Str}_{\exists}(G_{\psi}))\}$$

7.2 **Dynamic** semantics: **Solution** concepts

Once the games are fixed together with the strategies of the players in a compositional way, we can now go one step further than in GTS (extensive form), and bring the semantics to life, that is, define solution concepts in terms of the strategie interaction of the players. For this purpose, let us fix a concrete CDS with a strategy set S_{α} for each agent $\alpha \in A$. A strategy profile is defined in the usual way as a mernber of the product of all strategy sets:

$$(\sigma_{\alpha})_{\alpha\in\mathcal{A}} \in \prod_{\alpha\in\mathcal{A}} S_{\alpha}$$

The idea is that each state ansmg from $(\sigma_\alpha)_{\alpha\in\mathcal{A}}$ is a maximal one reached by starting in the initial state 0 and repeatedly playing the strategies in the profile until no further moves can be made. We still need a way to evaluate maximal states (payoffs, utilities, truth-valuation). **In** the present case, one uses Boolean valuations functions, from the formula of the state into the set {0,1}. For instance, if the formula of the maximal state is $Qa\,.\,Pa$, then we let its value (in the background model) be 1 if both Qa and Pa hold.

Finally, let us give an example which illustrates the treatment of anaphora.

Example 7.1. We consider the game associated with the formula $\exists x Qx \cdot Px$ which can be thought of as the logical form of our earlier example (2).

The corresponding CDS has only one cell (Recall that the games for Qx and Px have the form (0,0,0,0)):

$$\exists x$$

In a maximal play, the cell is filled with an element of the domain:

$$\begin{matrix}\exists x\\ a\end{matrix}$$

The maximal play gets payoff 1 for Eloise if

$$oa.>«$$

holds in the relevant model:

$$Qa \text{ and } Pa$$

Then we can define notions like Nash equilibria, etc. The important thing to notice is that whatever Nash equilibria gets selected as the solution of the game, the situation is going to be such that the syntactically free variabie *"x"* is going to be semantically bound by the existential quantifier. The dynamic game semantics has thus led to an extension of CTS in the direction of dynamic logic.

References

Abramsky, S. (2006). Socially Responsive, Environmentally Friendly Logic. **In** Aho, T. & Pietarinen, A.-V., eds., *Truth and Games: Essays in horcour of Cabriel Sandu,* Vol. 78 of *Acta Philosophica Fennica,* pp. 17-46. Societas Philosophica Fennica. Reprinted as Abramsky (2007).

Abramsky, S. (2007). A compositional game semantics for multi-agent logies ofpartial information. **In** van Benthem, .1., Gabbay, D. & Löwe, B., eds., *Intemctive Logic: Selected Papers [rom the 'lth Augustus de Morgan Workshop, London,* Vol. 1 of *Texts in Logic and Games,* pp. 11-47. Amsterdam University Press.

van Benthem, J. (2003). Logic Games are Complete for Game Logies. *Studia Logica, 75(2):183-203.*

Blutner, R. (2000). Some Aspects of Optimality in Natural Language Interpretation. *Journalof Semantics, 17:189-216.*

Carlson, L. & Hintikka, .1. (1979). Conditionals, generic quantifiers, and other applications of subgames. **In** Guenthner, T. & Smith, S.J., eds., *Formol Semantics and Fmgmatics of Notural Languages.* D. Reidel Co., Dordrecht, Holland.

Chomsky, N. (1986). *Knowledge of Language.* Convergence.

Dekker, P. & van Rooij, R. (2000). Bi-Directional Optimality Theory: an Application of Game-Theory. *Journalof Semantics, 17(3):217-242.*

Grice, H.P. (1989). *Studies in the Way of Words.* Harvard University Press, Cambridge, Mass.

Hintikka, .1. & Kulas, .1. (1985). *Anaphom and Definite Descriptions: Two Applications of Game-Theoreiical Semantics.* D. Reidel, Dordrecht.

Hintikka, J. & Sandu, G. (1991). *On the Methodology of Linguistics.* Basil Blackwell.

Lewis, D. (1969). *Convention.* Harvard University Press.

Parikh, P. (1991). Communication and Strategie Inference. *Linguistics and Fhilosophy, 14(5):473-513.*

van Rooij, R. (2002). Signalling Games Select Hom Strategies. **In** Katz, *G.,* Reinhard, S. & Reuter, P., eds., *Sinn und Bedeutung VI, Proceedinqs of the Sixth Annual Meeting of the Gesellschaft fiir Semantik,* pp. 289-310. University of Osnabrück.

van Rooij, R. (2006). Optimality-theoretic and game-theoretic approaches to implicature. **In** Zalta, E.N., ed., *The Stanford Encyclopedia of Fhilosophy.* Winter 2006 edition. http://plato.stanford.edu/archives/win2006/entries/implicature-optimality-games/.

Stenius, E. (1967). Mood and Language-Game. *Synthese, 17(1):254-274.*

.Games That Make Sense': Logic, Language, and Multi-Agent Interaction

Johan van Benthem

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam, The Netherlands

Department of Philosophy
Stanford University
Stanford, CA 94305, United States of Ameriea
johan@seienee.uva.nl

Abstract

Gabriel Sandu tells an appealing story of naturallanguage viewed in tenns of games and game theory, bringing together several strands from the philosophical, logical, and even computationalliterature. In this short invited note, I will take the cruising altitude a few levels up from his, and show you a panoramic picture where the clamour of the raw facts on the ground has become just soothing, but wholly negligible background noise.

1 Meaning is a many-mind notion

What do games have to do with natural language? On the traditional view of linguists and logicians, *syntax* is about grammatical code, *semantics* is about mathematical relationships between syntactic code and structures in reality, while the rest of language use is the bustling but unsystematic world of *pmgmatics*. **In** particular, on this view, meaning does not involve agency of any kind: it is a 'O-agent notion'. But starting from the 1970s, another view emerged placing actions of language users at centre stage, making meaning the 'information change', or more general 'context change potential' of linguistic expressions. Speakers or writers change the information states of their heaters or readers, and semantics should describe these changes. This action and update-oriented 'I-agent view' of meaning is the basis of the well-known Amsterdam paradigm of 'dynamic semantics' developed by Groenendijk, Stokhof and Veltman and their students, and it also underlies the well-known 'discourse representation theory' of Hans Kamp and Irene Heim.[1] Of course, this move also involves shifting the agenda. **In**

1 See the *Handbook of Logic and Language* (van Benthem and ter Meulen, 1997) for a survey of paradigms and sourees in dynamie semanties broadly eoneeived sinee the

Krzysztof R. Apt, Robert van Rooij (eds.). *New Perspectives on Games and Interaction.* Texts in Logic and Games 4, Amsterdam University Press 2008, pp. 197-209.

particular, it relocates the traditional boundary line between semantics and pragmatics in the study of language, and entire philosophical conferences have been devoted to that teetonic movement.é

But once on this raad, it seems strange to stop here. Psychologists of language like Herb Clark (1996) have shown convincingly that much oflanguage use is directed toward the hearer's rather than the speaker's perspective, it is the hearer's uptake which determines the success of communication. And once you start thinking all this through, you wind up in a 'hermeneutical circle' of speakers taking into account how their hearers will interpret what they are saying, and heaters taking into account how speakers will phrase what they are saying, and level aftel' level of stacked mutual information unfolds, leading to the iterative 'theory of mind' and mutual expectations that keep human behaviour stabie according to philosophers and psychologists. It also leads naturally to *game iheoru,* since that is where these circles find their resting place in refiective and action equilibria.

2 Games have a history with natural language

Indeed, the idea that natural language has an intimate relationship with games has recurred through the 20th century. **In** the 1950s, the later Wittgenstein famously moved away from the crystalline logical structure of the *Tmctatus* to a paradigm of rule-generating 'Ianguage games', and as Gabriel Sandu shows, authors like Stenius tried to put more substance into the game metaphor. Also in the 1950s, maybe under the infiuence of the then nascent game theory.f various proposals were made for analyzing logic in terms of 'Iogic games', casting basic logical activities like argumentation (Lorenzen, 1955) or model comparison (Ehrenfeucht-Fraïssé; cf. Ehrenfeucht, 1957) as two-player games, with winning strategies encoding proofs, modeis, or invariance relations, as the case might be." **In** particular, Gabriel Sandu discusses one of these, Hintikka's evaluation games for first-order logic (Hintikka, 1973), which later made its way into the study of naturallanguage under the name of 'Game-Theoretical Semantics' (GTS). We will return to these games later, which mainly analyze the 'Iogical skeleton' of *sentence construciion:* connectives, quantifiers, and anaphoric referential relationships. Thus, logic is the driver of the analysis here--and the expression 'game-theoretic' does not suggest any deep contacts with game theory.f

1970s, which also run over into computer scicnce.

2 Viewed in this way, natural language is no longer a descriptive medium, but rather a *program.m.ing* language for bringing about cognitive changes.

3 Much of the modern history of logic and its interfaces remains to be written, since authors usually stay with the aftermath of the foundational era in the 1930s.

4 (Van Benthem, 2007a) is an extensive survey and discussion of logic games today.

5 But see below for some mathematical contacts between logic games and game theory.

Also in the same 1960s, another, logic-free, style of game-theoretic analysis for natural language came up in Lewis' work (cf. Lewis, 1969), going back to (Schelling, 1960) on *signaling games.* **In** this way of looking at language, Nash equilibria establish stabie meanings for *lexical items,* the smallest atoms of sentence construction. While this new view long remained largely a small undercurrent," it has now become a major contender, with the authors discussed by Gabriel Sandu: (Parikh, 1991), (Dekker and van Rooij, 2000), (van Rooij, 2002), (.Jäger and van Rooij, 2007). While logic games are largely about winning and losing only, these modern signaling games involve real preferences that communicating linguistic agents have about matching up intended and perceived meaning, grammatical structure,[7] as weil as computational costs in doing so. Thus, they involve more serious connections with game theory, and at the same time, with the topological and metric structure of human perceptual and conceptual spaces (cf. Gärdenfors and Warglien (2006)). This may weil be the most serious encounter between linguistics and game theory today,8 and there are many interesting questions about its conneetion to the earlier logic-game based approaches like GTS. Sandu is quite right in putting this link on the map in his piece, though much still remains to be clarified.

3 Evaluation games, language, and interactive logic

The basic idea of Hintikka-style evaluation games is that two players, **Verifier** and **Falsifier,** disagree about whether a given first-order formula φ is true in a given model $\mathcal{M}$, under some assignment sof objects to variables.? The rules of the game refiect this scenario-and they may be seen as describing dynamic mechanisms of evaluation or investigation of facts about the world. With disjunctions $\varphi \vee \psi$, **Verifier** must *choose* a disjunct to defend **(Falsifier** is opposed to both), with conjunctions $\varphi \wedge \psi$, the choice is Falsifier's. A negation $\neg\varphi$ triggers a *mie switch,* where players change roles in the game for φ. Moreover, quantifiers let players choose an object from the domain: $\exists x\varphi$ lets **Verifier** choose a 'witness', $\forall x\varphi$ lets **Falsifier** choose a 'challenge', aftel' which play continues with the game for the formula φ. These moves change assignments of objects to variables, because the new

6 Lewis himself did add interesting thoughts on 'Score-Keeping in a Language Game'. Also, the stream of work on common knowledge in epistemic logic relates to Lewis' study of conventions, though there are even some earlier sourees in the social sciences.

7 This scenario comes part.ly from linguistic Optimality Theory and its 'rule-free' paradigm which casts language users as optimizing syntactic and semantic analysis of assertions along a set of constraint-based preferences.

8 Economics and cognitive science are other natural partners in this mix, as in the newly established interdisciplinary Bielefeld Heisenberg Center in 'Games and Cognition'.

9 lt has often been fruitful---e.g., in situation theory and in dynamic semantics-to use first-order logic, not as a literal translation medium for nat.ural language, but as a methodological 'test lab' for investigating basic features of actual usage.

value of x now becomes the chosen object d. When the game reaches an atomie formula, it is checked against the cutrent assignment, and **Verifier** wins if it is true, and loses otherwise. **In** all, this produces a two-agent scenario of changing assignments which has the following basic property. A formula φ is true at $(\mathcal{M}, s)$ iff **Verifier** has a *winning stmtegy* in the evaluation game $Game(\varphi, \mathcal{M}, s)$.

Much can be said about this simple game. For instanee, the dynamic view of logical constants as moves in a game is intriguing, and so is the muiti-agent 'pulling apart' of basic logical notions into different roles for different players. **In** this setting, players' strategies become logicalobjects in their own right now, expressing 'dependencies' in interactive behaviour. This powerful and appealing viewpoint also underlies other logic games, and its many repercussions are still not fully developed today, where we seem to be witnessing the birth pangs of an 'interactive logic'.[10] Van Benthem (2007b) also points out surprising connections with the early foundations of game theory. **In** particular, the law of Excluded Middle for first-order logic says that **Verifier** can always win games ofthe form $\varphi \vee \neg\varphi$. Unpacking this by the above rules, the law says that either **Verifier or Falsifier** has a winning strategy in the evaluation game for any formula φ. This 'determinacy' can be proven via Zermelo's Theorem about zero-sum two-player games of finite depth, which in its turn also follows from Excluded Middle plus some logically valid game transformations.[11] Thus, semantical evaluation, and hence also linguistic meaning in a procedural sense, meets with classical game theory-a conneetion elaborated in (van Benthem, 2007b).

In particular, viewed in this way, major issues in natural language semantics meet in inLeresLing ways wiLh basic quest.ions about games. Here is one. As we said, applying logical operations in formal languages serves as a model for sentence construction in natural language. And the most famous semantic issue arising then is Frege's Principle of *compositionality:* which says that the meaning of any linguistic expression can be determined stepwise, in tandem with its construction out of grammatical parts. Here, too, games offer a fresh perspective. As we saw, logical operations correspond to moves in an evaluation game--but we can also state the above scenario differently, since it has nothing to do with the specific games involved. Disjunction and conjunction are really quite general *game operations,* taking two games C, H to a *choice game* $C \cup H$ or $C \cap H$ starting with a choice by one of the players. Likewise, negation forms the obvious *dual game* to any given game. Thus, issues of linguistic compositionality become questions

10 The recent strategie EuroCoRes Project 'LogiCCC: Modeling Intelligent Interaction in the humanities, computational and social sciences' is an effort to put this development on the map in a much more general setting.

11 Evaluation games for other logical languages can be much more complex, involving infinite histories---e.g., with the modal μ-calculus: cf. (Bradfield and Stirling, 2006).

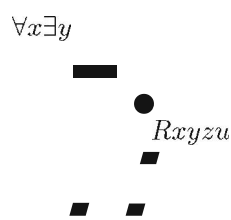
about *game algebm,* and the laws satisfied by natural game operations. For instance, van Benthem (2003) shows how the complete game algebra underlying first-order logic is a decidabie mixture of principles from Boolean Algebra plus laws for a left-, though not right-distributive operation G; *H* of *sequential composition* of games. Thus, if we take the evaluation game perspecton naturallanguage seriously as a view of multi-agent processes, we must understand the algebraic structure of the natural operations creating complex games for compound linguistic expressions out of simple ones.

4 Imperfect information and dependence

But logical evaluation games in CTS have further interesting features from realistie game theory, viz. *imperfect information.* Standard logic games, with the above evaluation games as a prime example, assume perfect information: players can observe each move that is played, and their only uncertainties are about future moves yet to be played. Cabriel Sandu has been one of the prime movers in a generalization, however, where the perfect information is abandoned in the process of semantic evaluation. Quantifier sequences in natural language sometimes show patterns of dependence and independence where it seems very natural to assume that access is blocked to objects chosen earlier. **In** the 'slash notation' of 'independence-friendly logic' ('IF logic'), a sequence like

$$\forall x \exists y \forall z \exists u / x \; Rxyzu$$

represents a 'branching quantifier' that can be written two-dimensionally as



This is true iff **Verifier** has a winning strategy consisting of responses to objects chosen by **Falsifier,** where the choice for *u.* only depends on the object chosen for *z.* **In** this scenario, evaluation games are no longer determined, and they may even have only mixed equilibria in random strategies, letting probability into the inner sanctum of logic. There is a large technical literature on this generalization of classical evaluation games, but its game content is under debate, and Hintikka has been downplaying the original game motivation. Indeed, **IF** logic has inspired a mathematical analysis as

generalized predicate logic by Hodges (1997), while Väänänen (2007) extracts the abstract logic of *dependenee* at stake here without game modeis.

But the jury is still out. For instance, van Benthem (2003) analyzes branching quantifiers in terms of a new game operation of the *parallel product* G x *H* of two games being played simultaneously without intermediate communication.P One reason why this is of interest to natural language is as follows. It has been claimed that IF logic is deeply non-compositional, a difficulty related to the absence of natural 'sub-games' in games with imperfect information (Osborne and Rubinstein, 1994). But introducing parallel product operations makes the underlying game algebra compositional again. Sandu's artiele actually discusses another recent game-theoretic take on IF, stemming more from the game semantics of programming languages. Abramsky (2006) makes connections between IF logic and fragments of linear logic, whose parallel products do allow for intermediate communication, copying moves from one sub-game to another. In all then, the question of the complete multi-agent game algebra behind evaluation processes for natural language seems open, although by this stage, we have a much deeper mathematical take on 'language games' than that of the 1950s.

5 Which games 'make sense' for natural language?

Our story so far does not exhaust the varieties of games that have been, or can be, brought to bear on natural language. There is a throng of further candidates, reftecting the many levels at which language can be studied.

5.1 Logic games

For a start, there are many logic games, and some fit natural language just as weil as evaluation games for sentences φ against models $\mathcal{M}$. In much 01'-dinary communication, there is no model at all of the described situation to evaluate against. What seems much more realistic then is 'consistency management'. We take in what the speaker says, and try to integrate this into consistent 'discourse representation structures' or more abstract semantic information states, unless the pressures on the successive updates become too high, and a conversational collapse takes place. But for this consistency management, a much more appropriate scenario might be *logic games of model construciiori,* which build models for sets of formulas (Hodges, 1997; van Benthem, 2003). In the semantics of natural language, the relevant distinction is 'dynamics of evaluation' (as in systems like DPL) versus *'dynamics of interpretation',* viewed as constructing a model or 'discourse representation' that makes sense of the cutrent linguistic utterances.l'

12 Van Benthem, Ghosh & Liu (2007) provide its complete game logic and algebra.

13 Indeed, van Benthem & van Eijck (1982) already proposed that the proper take on Hintikka's view of natural language would be model building games associated with the methad of *semantic tableaux* rather than with semantic model checking.

Interestingly, from a logical point of view, model building games are closely related to *dialogue games fOT pTOOf.* As we said earlier, these were already introduced by Lorenzen (1955), who wanted to explain logical validity of inferences P ⇒ C as the existence of a winning strategy in argumentation or debate for the **Proponent** of the conclusion C against any **Opponent** granting the premises P. This raises the issue of *infereniial* views of language and communication, which we will not pursue here. Historically, through the intermediate stage of (Blass, 1992), Lorenzen dialogue games eventually led to game semantics for linear logic and programming languages in Abramsky's style. Thus, the games that Sandu tries to conneet with **IF** logic seem quite different in spirit-but a link may be made through 'proof-theoretic' or category-theoretic semantics (Abramsky, 2üü7b).14

5.2 **Signaling** games

Now add the signaling games from the recent work by Parikh, van Rooij, and others, mentioned above. Sandu makes a simple and prima facie seamless connection, but I wonder about the consistency of scenarios. Signaling games really represent a very different scenario of language use, prior to the level of logic games. A logical evaluation game can only work when two things have already been settled: (a) the meaning of the logical operations, and (b) the denotations of the basic lexical items such as predicates and object names. But signaling games are about establishing the latter, and maybe even the former, connections in the first place!

Now in standard communication scenarios, we may assume that this initial phase has been achieved already, so that a global or at least alocal, 'linguistic convention' has arisen. **In** that case, we can focus on the higher tasks of making claims, and convincing others. But there can be cases where the two tasks meet, as in the creation of the right anaphoric links, which do not have fixed conventional meanings. It is here where Sandu focuses his discussion, and I have nothing to add to that.[15] Even so, it seems fair to say that we have no integrated technical theory of logic games and signaling games, and I wonder what would be a good way of combining them. Do we need a game algebra for natural language which allows for composition of heterogeneous games of quite different sorts?

Finally, from the viewpoint ofnaturallanguage, we have not even reached the complete picture of what goes on in ordinary conversation. There may be games that fix meanings for lexical items and for truth or falsity of expressions whose meaning is understood. But having achieved all that, the 'game of conversation' only starts, since we must now convey information,

14 This take on natural language interpretation seems closer to Categorial Grammar and its semantics in the lambda calculus, cf. (van Benthem, 1991; Moortgat, 1997).

15 Other natural examples arise in the semantic secnarlos of 'bi-directional Optimality Theory', many of which go beyond anaphora.

try to persuade others, and generally, further our goals-and maybe those of the ot hers as weil. In this area, Dutch-style logicians have developed a broad family of 'dynamic-epistemic logies' for analyzing information update and belief revision (cf. Baltag et al. (1998); Gerbrandy (1999); van Ditmarsch et al. (2007); van Benthem et al. (2006)). These systems have already been given game-theoretic interpretations (van Benthem, 2001, 2007c), and recent twists toward rational agency include dynamic logies for preferenee change (cf. the dissertations of Liu (2008), Girard (2008) and Roy (2008)).

But conversation and communication is also an arena where game theorists have entered independently, witness the earlier references in (van Rooij, 2002), and the recent signaling games for conversation proposed in (Feinberg, 2008). Again, there is an interface between logic and game theory to be developed here, and it has not happened yet.

5.3 The long term: language communities

Finally, there is one more level where games meet with natural language. We have talked about lexical meaning assignment, compositional semantics for single expressions, about checking for truth, argumentation, or information flow. But these are all short-term processes that run against the backdrop of a much larger, and potentially infinite process, viz. natural language use in communities with its *conventions over- time*. In terms of computer science, the former are terminating special-purpose processes for concrete tasks, while the latter are about the never-ending 'operating system' of natural language. Here, again, signaling games are relevant, and they have been applied to such diverse issues as the emergence of Gricean norms in pragmatics (van Rooij, 2006) or of warning signais, or argumentative strategies (Rubinstein, 2000).

In these scenarios, a significant move takes place, from single games to iterated games with *infinite runs*. Scenarios often involve thought experiments in terms of biological fitness and evolutionary stability against 'invaders' deviating from equilibrium. This is still about games and natural language, but with a very different agenda of explaining global, rather than local features of linguistic behaviour. And it is a far cry from logic games, involving rather dynamical systems theory for computing equilibria. Even so, it makes sense to ask for contacts aftel' all. Infinite games like repeated Prisoner's Dilemma are iterated game constructions out of simple base games, so a discrete algebra of game constructions still makes sense in this extended setting. Moreover, logic games are often infinite, most clearly in the game semantics for linear logic and associated programming languages. And even from a narrowly logical point of view, questions about stability of long-term natural reasoning practices make just as much sense as they do for linguistic conventions in natural language.

Thus, despite some conceptual and technical differences of emphasis and style in the literature right now, the encounter between logic and game theory in the arena of natural language seems far from concluded.

5.4 Naturallanguage as a circus: a carroussel of games

I started by saying that naturallanguage has three main aspects of syntax, semantics and pragmatics. By now it will be clear that 'linguistics' can ask questions about many levels of language use, asking for explanations of simple word meanings to successful discourse, and eventually the existence of broad norms and conventions that hold linguistic communities together. It also seems clear that games, whether from inside logic or directly from game theory, have an attractive role to play here, as an explicit way of bringing out the interactive multi-agent character of language use.

But what is the total picture? I have described natural language as a carroussel of games, where you can walk from one activity to another, and line up for the associated game. Is there a unifying principle, perhaps, one 'super-game'? Should we find clues in mathematics, at some level of 'deep game algebra', or rather in the communicative character of homo sapiens? I do not know, but I think that these questions are worth asking, if 'games and language' is to be more than a bunch of separate clever techniques.

6 Coda: but what about 'logic of games'?

Many people have heard of fruitful, and even Nobel-prize winning connections between logic and game theory-but the above story would probably leave them bewildered. What we have discussed in this note are game-theoretic models for basic linguistic and logical activities. Dut there is a quite different interface, too, where logic and language play their traditional role, viz. the description and analysis of game forms, strategies, information and reasoning of agents. This involves epistemic, doxastic and dynamic logics, providing analyses of notions such as rationality and its associated game solution procedures. In this descriptive guise, logic plays the same role toward game theory as it does toward multi-agent systems or process theories in computer science. Indeed, this more traditional use of logical techniques constitutes the main thrust of work in my own ILLC environment in Amsterdam, where games serve as one rich and intuitively appealing model of intelligent interaction that we want to capture by logical means.[16] This is also the sense in which computer scientists have embraced game theory as a richel' model for computation (Grädel, 2004), and philosophical logicians as a concrete model for rationality (Stalnaker, 1997). All these contacts can take place while logic keeps its standard semantic and proof-theoretic

16 Cf. (van Benthem, 1991,2006); as weil as the bundIe of European projeets eonstituting the recent LoglCCC team 'Logies for Interaction'.

face. Of course, game-theoretic ideas can reach logic in this way, and they do-but there is no need for logic to 'let' game theory 'under its skin', and recast itself as a family of games, as we have suggested in the above.

Nevertheless, the latter more radical view, too, has its basis in the history of logic, and it constitutes what Abramsky (2007b) calls *logic as embodying process* rather than logic as external process description.!" Indeed, the two directions are related. We can use standard logical languages to describe games, and then go on to use games to reinterpret what these logical languages are. The result is a wonderful circle--carroussel?-where the two fields spin happily together on each other's backs. I find that interactive view weil in line with the spirit of the present book.

References

Abrarnsky, S. (2006). Socially Responsive, Environmentally Friendly Logic. **In** Aho and Pietarinen (2006), pp. 17-46. Reprinted as Abramsky (2007a).

Abramsky, S. (2007a). A compositional game semantics for multi-agent logics of partial information. **In** van Benthem, .1., Gabbay, D. & Löwe, R, eds., *Intemctive Logic: Selected Papers [rom the 'lih Augustus de Morgan Workshop, London,* Vol. 1 of *Texts in Logic and Games,* pp. 11-47. Amsterdam University Press.

Abrarnsky, S. (2007b). Information, Processes, and Games. **In** Adriaans, P. & van Benthem, .1., eds., *Handbook of the Fhilosophy of Injormaiion:* Elsevier, Amsterdam.

Aho, T. & Pietarinen, A.-V., eds. (2006). *Truth and Games: Essays in horcour of Cabriel Sandu,* Vol. 78 of *Acta Fhilosophica Fennica.* Societas Philosophica Fennica.

Baltag, A., Moss, L.S. & Solecki, S. (1998). The Logic of Common KnowIedge, Public Announcements, and Private Suspicions. **In** Gilboa, 1., ed., *Proceedinqs of the 'lih Conference on Theoretical Aspects of Rationality and Knowledge, (TARK'98),* pp. 43-56.

van Benthem, .1. (1991). *Language tri Action: Caieqories, Lambdas, and Dynamic Logic.* Elsevier, Amsterdam.

van Benthem, .1. (2001). Games in Dynamic Epistemic Logic. *Bulletin of Economic Research. 53(4):219-248.*

17 The lecture notes *Logic in Games* (van Benthem, 2007b) eall this fundamental distinetion one of 'logie games' versus 'game logies'.

van Benthem, .1. (2003). Logic Games are Complete for Game Logics. *Studia Logica,* 75(2): 183-203.

van Benthem, .1. (2006). Logical Construction Games. **In** Aho and Pietarinen (2006), pp. 123-138.

van Benthem, .1. (2007a). Logic Games: From Tools to Models of Interaction. **In** Gupta, A., Parikh, R. & van Benthem, .1., eds., *Logic at the Crossroads: An Inierdisciplinaru View,* Vol. 1, pp. 283-317. Allied Publishers, Mumbai.

van Benthem, .1. (2007b). Logic in Games. Lecture notes. Institute for Logic, Language and Computation.

van Benthem, J. (2007c). Logic, Rational Agency, and Intelligent Interaction, illc research report. **In** Westerstahl, D., ed., *Proceedinqs of Beijing Conqress on Logic, Methodology, and Fhilosophy of Science,* London. College Publications. To appeal'.

van Benthem, .1., Gosh, S. & Liu, F. (2007). Modeling Simultaneous Games with Concurrent Dynamic Logic. **In** van Benthem, .1., Ju, S. & Veltman, F., eds., *A Meeting of the Minds, Fmceedings LORI Beijing,* pp. 243-258. College Publications.

van Benthem, .1. & ter Meulen, A., eds. (1997). *Handbook of Logic and Language.* Elsevier, Amsterdam.

van Benthem, .1. & van Eijck, .1. (1982). The Dynamics of Interpretation. *Joumal of Semantics, 1(1):3-20.*

van Benthem, .1., van Eijck, J. & Kooi, B. (2006). Logics of Communication and Change. *Ïnformaiioti and Computation, 204(99):1620-1666.*

Biass, A. (1992). A Game Semantics for Linear Logic. *Annals of FUTe and Applied Logic, 56(1-3):183-220.*

Bradfield, J. & Stirling, C. (2006). **PDL** and Modal μ-Calculi. **In** *Handbook of Modal Logic.* Elsevier, Amsterdam.

Clark, H. (1996). *Using Language.* Cambridge University Press, Cambridge.

Dekker, P. & van Rooij, R. (2000). Bi-Directional Optimality Theory: an Application of Game-Theory. *Journol of Semantics, 17(3):217-242.*

van Ditmarsch, H., van der Hoek, W. & Kooi, B. (2007). *Dynamic Epistemic Logic,* Vol. 337 of *Synthese Libraru.* Springer.

Ehrenfeucht, A. (1957). Application of games to some problems of mathematical logic. *Bulletin de l'Académie Polonaise des Sciences, Cl. JIJ, 5:35-37.*

Feinberg, Y. (2008). Meaningful Talk. This volume.

Gärdenfors, P. & Warglien, M. (2006). Cooperation, Conceptual Spaces, and the Evolution of Semantics. **In** Vogt, P., Sugita, Y., Tuci, E. & Nehaniv, C., eds., *Symbol Crounduiç and Beyond: Third International Workshop on the Emergence and Evolution of Linguistic Communications, EELC 2006, Rome, Jtaly, September 30-0ctober* 1, *2006, Proceedinqs.* Springer.

Gerbrandy, .1. (1999). *Bisimulations on Plasiet Kripke.* **PhD** thesis, University of Amsterdam. *ILLC Publications* DS-1999-01.

Girard, P. (2008). *Modal Logic [or Preferenee and Belief Change.* **PhD** thesis, Universiteit van Amsterdam and Stanford University. *ILLC Publication* DS-2008-04.

Grädel, E. (2004). Games and Automata for Synthesis and Validation. *ERCJM News, 57:36-37.*

Hintikka, .1. (1973). *Logic, Language Games and Injormation.* Clarendon, Oxford.

Hodges, W. (1997). Compositional Semantics for a Language of Imperfect Information. *Logic Joumal of the JGPL, 5(4):539-563.*

Jäger, G. & van Rooij, R. (2007). Language Structure: Psychological and Social Constraints. *Synthese, 159(1):99-130.*

Lewis, D. (1969). *Convention.* Harvard University Press.

Liu, F. (2008). *Changing [or the Better. Preferenee Dynamics and Agent Diversity.* **PhD** thesis, University of Amsterdam. *ILLC Publications* DS-2008-02.

Lorenzen, P. (1955). *Einführung in die Operative Logik und Mathematik.* Springer, Berlin.

Moortgat, M. (1997). Categorial Type Logies. **In** van Benthem, .1. & ter Meulen, A., eds., *Handbook of Logic and Language,* pp. 93-177. Elsevier, Amsterdam.

Osborne, M. & Rubinstein, A. (1994). *A Course in Game Theory.* **MIT** Press, Cambridge (Mass.).

Parikh, P. (1991). Communication and Strategie Inference. *Linguistics and Philosophy, 14(5):473-513.*

van Rooij, R. (2002). Signalling Games Select Hom Strategies. In Katz, *G.,* Reinhard, S. & Reuter, P., eds., *Sinn und Bedeutung VI, Proceedinqs of the Sixth Annua! Meeting of the Gesellschaft fiir Semantik,* pp. 289-310. University of Osnabrück.

van Rooij, R. (2006). Optimality-theoretic and game-theoretic approaches to implicature. In Zalta, E.N., ed., *The Stanford Encyclopedia of Philosophy.* Winter 2006 edition. http://plato.stanford.edu/archives/win2üü6/entries/implicature-optimality-games/.

Roy, O. (2008). *Thinking before Acting: Intentions, Logic, Rationol Choice.* PhD thesis, University of Amsterdam. *ILLC Pub!ications* DS-2008-03.

Rubinstein, A. (2000). *Economics and Language.* Cambridge University Press, Cambridge.

Schelling, T. (1960). *The Strategy of Conflict.* Harvard University Press.

Stalnaker, R. (1997). Reference and Necessity. In Hale, B. & Wright, C., eds., *A Companion to the Philosophy of Language.* Blackwell, Oxford.

Väänänen, .J. (2007). *Dependence Logic: A New Appmach to Independence Friendls] Logic,* Vol. 70 of *London Mathematica! Society Student Texts.* Cambridge University Press, Cambridge.

Solution of Church's Problem: A Tutorial

Wolfgang Thomas

Lehrstuhl lnformatik 7
RWTH Aachen University
52074 Aachen, Germany
thornas@inforrnatik.rwth-aachen.de

Abstract

Church's Problem (1957) asks for the construction of a finite-state procedure that transforms any input sequence α letter by letter into an output sequence β such that the pair (α, β) satisfies a given specification. Even after the solution by Büchi and Landweber in 1969 (for specifications in monadic second-order logic over the structure $(\mathbb{N}, +1)$), the problem has stimulated research in automata theory for decades, in recent years mainly in the algorithmic study of infinite games. We present a modern solution which proceeds in several stages (each of them of moderate difficulty) and provides additional insight into the structure of the synthesized finite-state transducers.

1 Introduction

Fifty years ago, during the "Summer Institute of Symbolic Logic" at Cornell University in 1957, Alonzo Church (1957) considered a problem which is both simply stated and fundamental.

Imagine a scenario in which an infinite bit stream α is to be transformed, bit by bit, into an infinite stream β, as indicated in the following figure.

The task is to construct a finite-state proeedure for this transformation when we are given a "specification" of the relation between α and β. This specification is usually presented as a formula of a logical system. **In** short words: We have to fill the box, given a description of the desired relation R between input α and output β. The problem is a question on automatic program synthesis which surprisingly can be answered positively when the specification language is not too expressive.

This setting for program synthesis is fundamentally different from the classical framework in which terminating programs for data transformations

Krzysztof R. Apt, Robert van Rooij (eds.). *New Perspectives on Games and Interaction.* Texts in Logic and Games 4, Amsterdam University Press 2008, pp. 211-236.

are considered. For correctness of a terminating program one relates the data given to the program before the start of the computation to those produced by the program at the end of the computation. Usually the data are from an infinite domain like the natural numbers. **In** Church's Problem, we deal with non-terminating computations in which inputs and outputs are interleaved, and the aspect of infinity enters in the dimension of time. On the other hand, the data processed in a single step are from a finite domain (in our example just {0, I}). **It** is this shift of infinity from data to time that allows to avoid undecidability results as known from the verification (or even synthesis) of terminating programs over infinite data spaces.

Let us look at an example. The relation *R* is defined by the conjunction of three conditions on the input-output stream (α, β). We use self-explanatory notation: $\alpha(t)$ is the *tth* bit of α *(t* = 0,1, ...), and $\exists^\omega$ is the quantifier "there exist infinitely many".

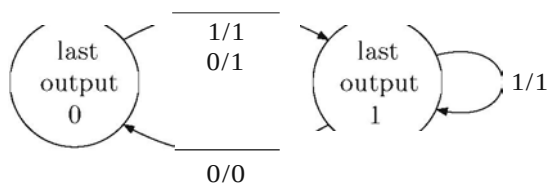
1. $\forall t(\alpha(t) = 1 \rightarrow$ *;3(t)* = 1)
2. $\neg\exists t$ *;3(t)* = $\beta(t + 1) = 0$
3. $\exists^\omega t\ \alpha(t) = 0 \rightarrow \exists^\omega t$ *;3(t)* = 0

The first two conditions are satisfied easily by producing output 1 at each moment. But the last condition, which has the form of a fairness constraint, excludes this simple solution; we cannot ignore the zero bits in α. A natural idea is to alternate between outputs 0 and 1 if the inputs are only 0. We arrive at the following procedure:

- Ior input 1 produce output 1
- for input 0 produce

  output 1 if last output was 0
  output 0 if last output was 1

This procedure is executable by the finite-state transducer displayed below. **It** is presented as an automaton in which each transition is labelled with an input bit and the cortesponding output bit. As initial state we take, for example, the left-hand state.



For a more precise statement of Church's Problem, it is necessary to fix the format of the specifications and that of the "solutions". Let us first address the solutions. Among the many concepts of transformations of sequences, only a very special form is admitted for Church's Problem. Two aspects are relevant, the requirement of a computation "bit by bit", and the restriction to "finite-state" solutions. The first requirement means that the output bit *j3(t)* has to be produced without delay aftel' receipt of the input bit $\alpha(t)$. Thus *j3(t)* can depend only on the input bits up to time *t,* i.e., on the input prefix $\alpha(0) \ldots \alpha(t)$. This is a sharper restriction than that of "continuity" (in the Cantor topology over *{0, 1}W),* which would mean that *j3(t)* depends on some finite prefix of α—possibly $\alpha(0) \ldots \alpha(s)$ with $s > t$. As an illustration, consider the transformation Tl with $T_1(\alpha) = \alpha(0)\alpha(2)\alpha(4)\ldots$. It is continuous but excluded as a solution for Church's Problem (since $T_1(\alpha)(t)$ depends on $\alpha(2t)$). A fortiori, non-continuous transformations are excluded, such as T_2 defined by $T_2(\alpha) = 111 \ldots$ if α has infinitely many letters 1, otherwise $T_2(\alpha) = 000 \ldots$ (note that no finite prefix of α determines even the first bit of $T_2(\alpha)$).

The restriction to "finite-state" solutions means, in Church's words, that the desired sequence transformation should be realizable by a "circuit". This is a much stronger assumption on the admissible transformations than the dependency of the *tth* output bit on the inputs bits up to time *t* only: One requires that the computation is realizable with a fixed finite memory (independent of *t),* as with the two states of memory in our example. It is remarkable that this restricted type of procedure actually suffices for solutions of Church's Problem. In this paper we work with finite-state transducers in Lhe Iorrnat of Mealy autornata. Forrnally, a *Jvlealy automaton* is a structure $\mathcal{M} = (S, \Sigma, \mathrm{T}, sa, \delta, T)$ where 8 is the finite set of states, Σ and f are the input alphabet and output alphabet, respectively, *sa* the initial state, $\delta : 8 \times \Sigma \to 8$ the transition function and $\tau : 8 \times \Sigma \to \mathrm{f}$ the output function. In a graphical presentation we label a transition from *P* to $\delta(p, a)$ by *a/T(p, a).* Later we shall also allow that certain transitions may not produce an output letter (but the empty word *c* instead). The function δ is extended to $\delta^* : 8 \times \Sigma^* \to 8$ by setting $\delta^*(s, \varepsilon) = s$ and $5^*(s,wa) = \delta(\delta^*(s, w), a)$ for $w \in \Sigma^*, a \in \Sigma$. For the input sequence $\alpha = \alpha(0)\alpha(1)\ldots$, the output sequence *j3* computed by $\mathcal{M}$ is given by *j3(t)* $= \tau(\delta^*(s_0, \alpha(0) \ldots \alpha(t - 1)), \alpha(t))$.

Let us now make precise the specification language. We consider the system of monadic second-order logic (MSO) over the successor structure $(\mathbb{N}, +1)$, also called SIS (for "second-order theory of one successor") or "sequential calculus". This case was emphasized by Church (1963) as an open problem, and today it is understood that "Church's Problem" refers to SIS. In the logical context, one identifies a bit sequence α with a set P_α of nat-

ural numbers that contains the numbers t with $\alpha(t) = 1$. We use $s, t, \ldots$ as first-order variables (ranging over natural numbers, or time instances) and $X, Y, \ldots$ as second-order variables (ranging over sets of natural numbers, or bit sequences). We view the latter as predicate variables and write $X(t)$ rather than $t \in X$. **In** SIS, one has quantifiers $\exists, \forall$ for both kinds of variables. One can define $s < t$ by saying that t belongs to each set which contains $s + 1$ and is closed under successor. Now our example specification takes the form of the fol!owing SIS-formula $\varphi_0(X, Y)$ (where we write $\exists^\omega t \ldots$ for $\forall s \exists t(s < t \wedge \ldots)$):

$$\forall t(X(t) \rightarrow Y(t)) \ \wedge \ \neg\exists t(\neg Y(t) \wedge \neg Y(t+1)) \ \wedge \ (\exists^\omega t \neg X(t) \rightarrow \exists^\omega t \neg Y(t))$$

In general, we have SIS-specifications that speak about sequences $\alpha \in (\{a, 1\}^{m1})^\omega$ and $\beta \in (\{a, 1\}^{m2})^\omega$. Then we consider bit veetors rather than single bits, and use ml-tuples $\underline{X}$ and m2-tuples $\underline{Y}$ of second-order variables in place of X, Y in the specifications. Similarly we write $\underline{P_\alpha}$ for the predicate tuple associated with α. Church's Problem now asks: *Given an SIS-specification $\varphi(\underline{X}, \underline{Y})$, construct a Mealy automaton $\mathcal{M}$ with input alphabet $\Sigma = \{0,1\}^{m1}$ and output alphabet $\Gamma = \{0,1\}^{m2}$ such that for each input sequence $\alpha \in (\{a, 1\}^{m1})^\omega$, an output sequence $\beta \in (\{a, 1\}^{m2})^\omega$ is produced by $\mathcal{M}$ with $(\mathbb{N}, +1) \models \varphi[P_\alpha, P_\beta]$, or prooide the arisuier that such an automaton does noi exist.*

An alternative view to study Church's Problem is to consider arelation $R \subseteq (\{a, 1\}^{m1})^\omega \times (\{0,1\}^{m2})^\omega$ as the definition of an infinite two-person game between players A and B who contribute the input-, respectively the output-bits in turn. A play of this game is the sequence of pairs $(a(t), \beta(t))$ of bits supplied for $t = 0, 1, \ldots$ by A and B in alternation, and the play $(a(0), \beta(0))\ (\alpha(1), \beta(1)) \ldots$ is won by player B iff the pair (α, β) belongs to R. A Mealy automaton as presented above defines a winning strategy for player B in this game; so we speak of a "finite-state winning strategy".

In 1969, Büchi and Landweber (1969) solved Church's Problem. The original proof involved a complicated construction. **It** took some time until more accessible proofs were available. The purpose of this tutorial is to present a construction which is made easy by a decomposition of the task into sirnpler modules (fol!owing Thomas, 1995; see also Thomas, 1997; Grädel et al., 2002). The construction also gives extra information on the structure of the finite-state machines that serve as solutions.

We wil! show the Büchi-Landweber Theorem in four stages: **In** a preliminary step, the SIS-specifications are converted into automata over infinite words ("w-automata"). Here we use, without going into details, classical results of Büchi and McNaughton that provide such a conversion (Büchi, 1962; McNaughton, 1966). We wil! il!ustrate this step by an example. Then we transform the obtained automaton into a game between the input player A

and the output player B (played essentiallyon the transition graph of the automaton). The task is then to decide whether B has a winning strategy and-if so-to construct a finite-state machine executing a winning strategy. The last two stages serve to obtain such a machine. First, we define its state space and transition function, and secondly we fix the output function. Only in this last step the decision about solvability of the specification will be obtained.

There is also an alternative approach to Church's Problem, developed by Rabin (1972) in the framework of tree automata theory. Let us briefiy sketch the idea. In the situation where both players A and B select bits, Rabin codes a strategy of player B by a labelling of the nodes of the infinite binary tree: The root has no label, the directions left and right represent the bits chosen by A, and the labels on the nodes different from the root are the bits chosen by B according to the considered strategy. When player A chooses the bits $b_a, \ldots, b_k$, he defines a path to a certain node; the label b of this node is then the next choice of player B. Note that a node labelling by bits corresponds to a subset X of the tree (containing the nodes with label 1). Now the paths through the (X-labelled) tree capture all plays that are compatible with B's strategy coded by X. One can write down a formula $X(X)$ in MSO-logic over the binary tree which states that the winning condition is satisfied byeach path; thus $X(X)$ says that "X is a winning strategy". By Rabin's Tree Theorem (Rabin, 1969) one can convert $X(X)$ into a Rabin tree automaton A_X (for definitions see e.g., Thomas, 1997), check whether this automaton accepts some tree, and-if this is the case--construct a "regular" tree accepted by A_X. This regular tree can then be inLerpreLed as a finite-state winning strategy Ior player B.

In the present notes we pursue the "linear" approach in which single plays are the main objects of study; so we avoid here the infinite tree structure that captures *all* plays for a given strategy.

2 From logic to automata and games

2.1 From logic to automata

Our first step for solving Church's Problem consists of a transformation of a specification $\varphi(\underline{X}, \underline{Y})$ into a semantically equivalent but "operational" form, namely into a deterministic automaton $\mathcal{A}_\varphi$ werking over w-sequences. This puts Church's Problem into the framework of automata theory. It is remarkable that we do not have any solution of Church's Problem that avoids this transformation at the start-e.g., by a compositional approach of synthesis that is guided by the structure of the formula φ.

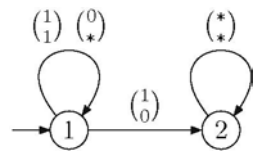
For an ml-tuple $\underline{X}$ and an m2-tuple $\underline{Y}$, the input alphabet of $\mathcal{A}_\varphi$ is $\{0,1\}^{m_1+m_2}$. The automaton is said to be equivalent to $\underline{cp(X,Y)}$ if it accepts precisely those w-words which define tuples $(\underline{P}, \underline{Q})$ of sets such that

$(\mathbb{N}, +1) \models$ *cp[P,Q].* **In** the game theoretic view explained above, one may consider the automaton as a referee who watches the play evolving between players A and B that consists of the two sequences α and *j3* (logically speaking: of the set tuple (P_α, Q_α) built up by A and B), and who decides at infinity (by the acceptance condition) whether B has won or not.

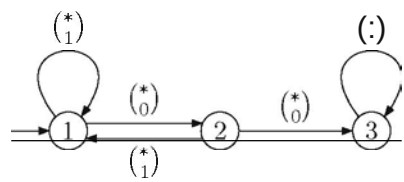
An appropriate acceptance condition is the so-called Muller condition. **It** is specified by a collection $\mathcal{F} = \{F_1, \ldots, F_k\}$ of state sets, and the automaton accepts an w-word γ if the set of the states visited infinitely often in the unique infinite run on γ is one of the F_i. (The sets F_i are called *accepting loops;* indeed, if the states in F_i are visited again and again they form a strongly connected set ("loop") in the transition graph of the automaton.)

We use here two core results of the theory of w-automata due to Büchi (1962) and McNaughton (1966) (see e.g., (Thomas, 1997) or (Grädel et al., 2002) for more recent expositions). They allow to translate an SIS-formula into an equivalent (non-deterministic) Büchi automaton, which is then transformed into a deterministic Muller automaton: *FOT each SlS-joTmula* $\varphi(X, Y)$ *one can construct an equivalent Muller automaton* $\mathcal{A}_\varphi$. As a drawback in this result we mention that the size of $\mathcal{A}_\varphi$ cannot be bounded by an elementary function in the length n of φ (see, e.g., Grädel et al., 2002); in other words, for no k, the k-fold iteration of the function $n \mapsto 2^n$ can serve as an upper bound for the size of $\mathcal{A}_\varphi$.

Let us illustrate the theorem for our example specification above. The formula $\forall t(\alpha(t) = 1 \rightarrow j3(t) = 1)$ is equivalent to the Muller automaton



with accepting loop {I} only (and where * stands for an arbitrary bit). The formula $\neg\exists t\beta(t) = j3(t + 1) = 0$ is expressed by the following Muller automaton with accepting loops {I} and {I, 2}:



The automaton for $\exists^\omega t\alpha(t) = 0 \rightarrow \exists^\omega t\beta(t) = 0$ is best explained as follows: There are four states, denoted by the four possible bit-pairs, with say (0,0)

as initial state. From each state we have, for each bit pair (b_1, b_2), a transition labelled (b_1, b_2) to the state (b_1, b_2). A set F is accepting if it satisfies the following condition: If the first component is 0 in some state of F, then the second component is 0 for some (possibly different) state of F.

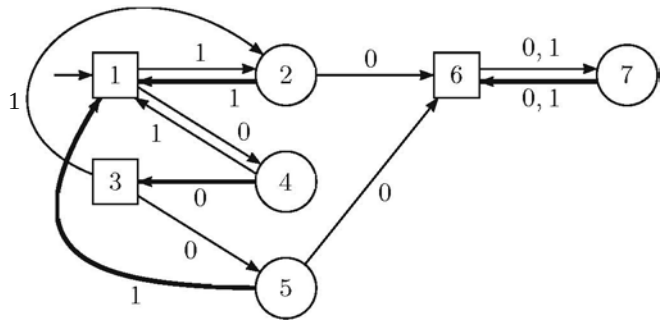
It is known how to combine Muller automata for several conditions to a single Muller automaton for their conjunction. We do not present it explicitly here for our example. Rather we turn to a variant, called "finite-state game with Muller winning condition". This approach, introduced by McNaughton (1993), is motivated by the view that the two components of an input letter of the Muller automaton are contributed by two players A and B who pursue antagonistic objectives: A aims at violating the condition φ and B at satisfying it.

2.2 From automata to games

We distinguish the contribution of bits (in the general case: bit vectors) by two players A and B by introducing two kinds of states, called A- and B-states. **In** an A-state, the next bit is to be picked by player A, in aB-state by player B. We indicate A-states by boxes and B-states by circles. The figure below indicates how we dissolve transitions from a state in the given Muller automaton by introducing intermediate states and corresponding transitions.

Note that we keep every state of the Muller automaton as an A-state. For each A-state q and bit b, we introduce a b-labelled transition to a new state called (q,b), and from (q,b) for each bit ca c-labelled transition to the state p which was reached from q by $\binom{b}{c}$ in the original automaton. For such a state p we call c the corresponding "output bit", denoted $\mathrm{out}(q,b,p)$. (If both c-transitions from (q,b) lead to the same state p we agree that $\mathrm{out}(q,b,p) = 0$.) If the input alphabet is $\{0,1\}^{m_1}$ and the output alphabet $\{0,1\}^{m_2}$, we introduce B-states $(q,\bar{b})$ with $\bar{b} \in \{0,1\}^{m_1}$, and define $\mathrm{out}(q,\bar{b},p)$ as a vector in $\{0,1\}^{m_2}$.

The result is a "game graph". For our example specification above, we can obtain the following game graph from a corresponding Muller automaton (the reader should ignore for the moment the boldface notation of some arrows).



The three conditions of our example formula are indeed captured by this graph. The first condition requires that a bit 1 chosen by A has to be answered by the bit 1 chosen by B. Ir this is violated (starting from the initial state 1), state 6 (and hence the loop consisting of states 6 and 7) is entered. The second condition says that player B should not piek two zeroes in succession. Ir this is violated, we would reach 6 and 7 again. We thus exclude states 6 and 7 from the accepting loops. The third condition (on fairness) means that if A chooses 0 infinitely often (which happens by going to 4 or 5), then B has to choose 0 infinitely often (which is only possible by going from 4 to 3). Altogether we deciare a loop F as accepting if it does not contain 6 or 7 and satisfies $(4 \in F \vee 5 \in F \rightarrow 3 \in F)$.

How should player B piek his bits to ensure that the play visits precisely the states of one of these loops F infinitely often? We have to fix how to move from states 2, 4, 5, 7. From 7 player B has to move to 6 since there is no ot her choice. The ot her choices can be fixed as follows: From 2 to 1, from 4 to 3, and from 5 to 1 (see boldface arrows). Then, depending on what Player A does, a play starting in 1 will visit infinitely often the states 1 and 2, or the states 1 to 4, or the states 1, 3, 4, 5, or the states 1 to 5. Each of these loops is accepting.

We see that the acceptance condition of a Muller automaton is thus turned into a winning condition in the associated game (Muller winning condition). Furthermore, we see that player B has a winning strategy by fixing his moves as stated above. This winning strategy can be converted into a Mealy automaton when we combine again each pair of two successive moves (by player A and then B) into a single transition. We get an automaton with the states 1 and 3 and the following transitions: From 1 via $\binom{1}{1}$ back to 1, from 1 via $\binom{0}{0}$ to 3, and from 3 via $\binom{0}{1}$ and via $\binom{1}{1}$ back to I. Up to names of states (and the irrelevant initial state), this is precisely the Mealy automaton mentioned in the Introduction.

In the remainder of the paper, we shall give a general construction that starts with a finite game graph equipped with a Muller winning condition,

provides the decision whether player B wins, and in this case yields a finite-state winning strategy.

We add some remarks on the step from automata to game graphs. First let us note that there is more behind this reshaping of the model than introducing the attractive idea of a game. The game theoretic view is most useful for introducing a symmetry between inputs and outputs in Church's Problem. The two players A and B represent the antagonistic aims of falsifying the specification (player A, supplying input) and satisfying it (player B, supplying output). **It** will turn out that either A or B has a winning strategy, an aspect which is hidden in the original formulation of Church's Problem.

Secondly, in studying plays over a given game graph, it is useful to ignore the special role occupied by the initial state. Rather we shall be interested in plays wherever they start, and we shall determine for each state which player has a winning strategy for plays starting from there.

On the other hand, we shall simplify the model in a different detail: We cancel the labels on the transitions. This is motivated by the fact that the winning condition is formulated in terms of visits of states only, regardless of the labels that are seen while traversing edges. When a winning strategy over the unlabelled game graph is constructed, it will be easy to re-introduce the labels and use them for a Mealy automaton as required in the original formulation of Church's Problem.

In our example, a Mealy automaton with two states was sufheient to solve Church's Problem for the specification φ_0. These two states were already present in the game graph G_{φ_0} corresponding to the Muller automaton $\mathcal{A}_{\varphi_0}$. (We took the states 1 and 3.) Given the game graph G_{φ_0}, we were able Lo fix Lhe moves of player B Irorn 1 and 3 independenL of Lhe "play history", i.e., independent of the path on which either of these states was reached. **In** general we shall need additional memory to define the right choice. We shall see that a finite memory suffices; so we can work with winning strategies that are implementable by finite-state machines. Such a finite-state machine S works on the game graph G. The states of S and of G should not be confused. For the solution of Church's Problem (given a logical formula φ) we have to combine the states of S with the states of G. We describe this in detail at the end of the next section.

3 Infinite games and the Büchi-Landweber Theorem

A *game graph {or arena)* has the form G = *(Q, QA, E)* where $QA \subseteq Q$ and $E \subseteq Q \times Q$ is the transition relation, satisfying $v« \in Q : qE \neq 0$ (i.e., $\forall q \exists q' : (q, q!) \in E$). This condition ensures that plays cannot end in a deadlock. (So a subset *Qo* of *Q* induces again a game graph if from each $q \in Qo$ there is an edge back to *Qo.)* We set $QB := Q \setminus QA$. **In** this paper edges will always lead from Q_A-states to QB-states or conversely; however

the results do not depend on this assumption. *We restriet to finite game gmphs ihrouqhoui the paper.*

A play over C from q is an infinite sequence $p = qOqlq2\ldots$ with $qo = q$ and $(qi, q_{i+1}) \in E$ for $i \geq 0$. We assume that player A chooses the next state from a state in Q_A, and player B from a state in Q_B. Note that the game graph is finite whereas the plays on it are infinite; thus one speaks of "finite-state infinite games".

Forrnally, a *game* is a pair (C, W) where $C = (Q, Q_A, E)$ is a game graph and $W \subseteq QW$ a "winning condition" for player B. Player B wins the play $p = qOqlq2\ldots$ if $p \in W$, otherwise player A wins p. The use of such "abstract" winning conditions W is pursued in descriptive set theory, see (Moschovakis, 1980). **In** our algorithmic context we have to work with finitely presentabie sets W. For our considerations below, we work with two finite presentations of winning conditions, either by a collection $\mathcal{F} \subseteq 2^Q$ of sets $R \subseteq Q$, or by a coloring $c : Q \to \{0, \ldots, k\}$ for some natural number k. **In** the special case $c : Q \to \{0,1\}$ we also consider the subset $F = \{q \in Q \mid c(q) = 1\}$ instead.

First we introduce two winning conditions connected with a collection $\mathcal{F} \subseteq 2^Q$. The first is the *Mulle» winning condition;* it refers to the set Inf(p) of states visited infinitely often in a play p:

$$\text{Inf(p)} := \{q \in Q \mid \exists^\omega i\ p(i) = q\}$$

Player B wins the play p if $\text{Inf(p)} \in \mathcal{F}$. With these conventions we speak of a *MulieT game* $(C, \mathcal{F})$.

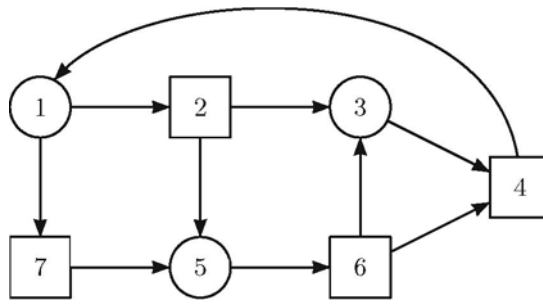
There is also a "weak" version of this winning condition, called *weak Muller condition* (or *Staiqer- WagneT condition),* which refers to the visited states in a play ("occurrence set"):

$$Occ(p) := \{q \in Q \mid \exists i\ p(i) = q\}$$

Player B wins a play p according to the weak Muller condition if $Occ(p) \in \mathcal{F}$. We speak of the *weak Muller game* $(C, \mathcal{F})$.

An important special case of weak Muller games is the *reachability game,* given a set $F \subseteq Q$ of states of the game graph (Q, Q_A, E). The winning condition for player B is satisfied for a play p if some state of p belongs to F. We obtain an equivalent weak Muller condition if we set $\mathcal{F} = \{R \subseteq Q \mid R \cap F \neq \varnothing\}$.

The next step is to introduce the concepts of strategy, winning strategy, and winning region. Let us look at examples first, using the following game graph Cl.



The reachability game $(\mathrm{Cl}, \{3\})$ with "goal set" $F = \{3\}$ is won by player B if the play starts in state 3. Otherwise player A can avoid this state by going from 2 to 5 and from 6 to 4. We shall say that the *winning reqion* of player A in this game is the set {I, 2,4,5,6, 7} and that of player B the set {3}. As a second example, consider the condition that states 2 and 7 both have to be visited again and again. Formally, this is the Muller game $(\mathrm{Cl}, \mathcal{F})$ where $\mathcal{F}$ consists of all sets $R \supseteq \{2,7\}$. Obviously, player B can win from any state: From 1 he proceeds to 2 and to 7 in alternation, from 5 he moves to 6, and from 3 to 4. So the winning region of A is empty in this case, and that of B the set of all states. Note that switching between the moves from 1 to 2 and from 1 to 7 means to use memory (here only one bit) when executing the strategy.

Formally, a *stmtegy fOT plosier B [rom q* is a function $f : Q+ \to Q$, specifying for any play prefix $qo \ldots qk$ with $qo = q$ and $qk \in QB$ some vertex $r \in Q$ with $(qk,T) \in E$ (otherwise the value of f is chosen arbitrarily). A play $p = qOql \ldots$ from $qo = q$ is played according to strategy f if for each $qi \in QB$ we have $q_{i+1} = f(qo \ldots qi)'$ A strategy f for player B from q is called *winning stmtegy fOT plosier B [rom q* if any play from q which is played according to f is won by player B. **In** the analogous way, one introduces strategies and winning strategies for player A. We say that A (resp. B) *wins from q* if A (resp. B) has a winning strategy from q.

For a game *(C, W)* with $C = (Q, Q_A, E)$, the *winning reqions of plasjers A and* Bare the sets $W_A := \{q \in Q \mid$ A wins from $q\}$ and $W_B := \{q \in Q \mid$ B wins from $q\}$. It is obvious that a state cannot belong to both W_A and W_B; so the winning regions W_A, WBare disjoint. But whether these sets exhaust the whole game graph is a more delicate question. One calls a game *deiermined* if $W_A \cup W_B = Q$, i.e., from each vertex one of the two players has a winning strategy. Determinacy of infinite games is a central topic in descriptive set theory; with the axiom of choice one can construct games that are not determined. For the games considered in this paper (i.e., games defined in terms of the operators Occ and Inf), determinacy

is well-known. Nevertheless we state (and prove) this claim in the results below, since determinacy is the natural way to show that envisaged winning strategies are complete: **In** order to show that the domain D of a strategy covers the entire winning region of one player, one verifies that from each state outside D the other player has a winning strategy.

By the solution of a game (C, W), with game graph C = (Q, QA, E) and a finitely presented winning condition W, we mean two tasks:

1. to decide for each $q \in Q$ whether $q \in W_B$ or $q \in W_A$
2. and depending on q to construct a suitable winning strategy from q (for player B, respectively A).

Item 2 asks for a winning strategy that has a finite presentation. Two kinds of strategies will be central in the sequel, the positional and the finite-state strategies. A strategy $i : Q^+ \to Q$ is *positional* if the value of $i(q1 \ldots qk)$ only depends on the "current state" qk. So a positional strategy for B can also be presented as a function $i : Q_B \to Q$, or-in graph theoretical terms-by a subset of the edge set where from QA-states all edges are kept but from each QB-state precisely one edge is chosen. For the definition of finite-state strategies, we first observe that over a finite state set Q, a strategy $i : Q^+ \to Q$ can be considered as a word function. We say that i is a *finite-state stmtegy* if it is computed by a Mealy automaton. **In** the present context a Mealy automaton is of the form $S = (S, Q, Q, sa, \delta, 7)$ with state set S, input alphabet Q, output alphabet Q, initial state sa, transition function $\delta : S \times Q \to S$, and output function $\tau : S \times QA \to Q$ for player A (respectively $7 : S \times QB \to Q$ for player B). The *stmtegy is computed by* S is now defined by $is(qo \ldots qk) = \tau(\delta^*(s_0, qo \ldots qk\text{-}l), qk)$ (where $\delta^*(s, w)$ is the state reached by S from s via input word w, as defined as above in the Introduction, and τ is chosen for the player under consideration).

Now we can state the main theorem on weak Muller games and on Muller games. We include part (a) for reasons of exposition; part (b) is the Büchi-Landweber Theorem.

Theorem 3.1. (a) Weak Muller games are determined, and for a weak Muller game $(C, \mathcal{F})$, where C has n states, one can effectively determine the winning regions of the two players and construct, for each state q of C, a finite-state winning strategy from q for the respective winning player, using 2^n memory states.
(b) Muller games are determined, and for a Muller game $(C, \mathcal{F})$, where C has n states, one can effectively determine the winning regions of the two players and construct, for each state q of C, a finite-state winning strategy from q for the respective winning player, using $n! \cdot n$ memory states.

Before entering the proof, we remark that part (b) gives the desired solution of Church's Problem. For this, we proceed as in the previous section,

i.e., we transform a given S1S-formula φ to a Muller automaton $\mathcal{M}$ which is then converted to a game graph G with Muller winning condition (see Section 2). Note that the game graph G inherits an initial state from $\mathcal{M}$. Using the Büchi-Landweber Theorem, one checks whether this initial state belongs to the winning region of player B, and in this case one obtains a Mealy automaton *S* that realizes a winning strategy from the initial state. The desired finite-state strategy for the original formula φ is now easily constructed as a product automaton from $\mathcal{M}$ and *S*.

We give the complete definitions for the reader who wants to see the details. For simplicity we consider the case $\varphi(X, \mathrm{Y})$ where each player picks single bits only. Let $\mathcal{M}$ be the Muller automaton obtained from $\varphi(X, Y)$, say with state set *Q*. The game graph G derived from $\mathcal{M}$ has *Q* as the set of A-states and *Q* x {0, 1} as the set ofB-states. Denote QU(Q x {0, 1}) by *Qo*. Let *S* = *(S, Qo, Qo, sa,* δ*, T)* be the Mealy automaton that realizes a finite-state winning strategy for player B in the Muller game over G from *qo* (the initial state of the Muller automaton). We construct the Mealy automaton *A* solving the considered instanee φ of Church's Problem as follows: *A* has the state set *Q* x *S* and the initial state *(qO' sa).* We have to specify a transition for each state *(q, s)* and input bit *b*, i.e., an output bit *bi* and a new state *(ql, Si).* For this we compute the state *q* = (q, b)* of the game graph and the associated S-state s" = $\delta(s, q^*)$. The output function of *S* yields the state *ql = T(S, q*)* of G and the new memory state *Si* = $\delta(s^*, ql)$. The output bit *bi* is the value out(*q, b, ql)* associated to the transition from *q* = (q, b)* to *ql* (cf. Section 2.2).

The memory of the automaton *A* combines the state space of the Muller autornaton **Mand** that of Lhe strategy autornaton *S*. IL is not yeL weil understood how these two aspects play together in genera!. Our example in Sections 1 and 2 illustrates the case that in addition to the states of $\mathcal{M}$ no additional memory is necessary.

4 Reachability games and weak Muller games

In this section we outline the proof of Theorem 3.1 (a). As a preparatory step we solve reachability games. The fundamental construction involved in this solution (computation of "attractor") later enters also in the solution of weak Muller games and Muller games. For this purpose, we introduce a second technique, called "game simulation". **It** allows to transform a given game into another one with an "easier" winning condition, namely such that the method as known from reachability games applies. We shall illustrate this approach first for weak Muller games (in this section) and then for Muller games (in the next section).

4.1 Reachability games

Recall that a reachability game (C, F) involves the winning condition (for player B) that the play should reach somewhere a state from the set F.

Theorem 4.1. A reachability game (C, F) with $C = (Q, Q_A, E)$ and $F \subseteq Q$ is determined, and the winning regions W_A, W_B of players A and B, respectively, are computable, as weil as corresponding positional winning strategies.

Proof. The proof follows a natural idea, namely to compute, for $i = 0,1, \ldots$, the vertices from which player B can force a visit in F within i moves. We call this set the ith "attractor" (for B). lts computation for increasing i is known from the theory of finite games (and corresponds to the well-known analysis of AND-OR-trees).

$$\mathrm{Attr}_B^i(F) :=$$
$$\{q \in Q \mid \text{from } q \text{ player B can force a visit of } F \text{ in } \leq i \text{ moves}\}$$

The inductive computation is obvious:

$$\mathrm{Attr}_B^0(F) = F,$$
$$\mathrm{Attr}_B^{i+1}(F) = \mathrm{Attr}_B^i(F) \cup \{q \in Q_B \mid \exists (q,r) \in E : r \in \mathrm{Attr}_B^i(F)\} \cup \{q \in Q_A \mid \forall (q,r) \in E: r \in \mathrm{Attr}_B^i(F)\}$$

So for step $i+1$ we include a state of Q_B if from it some edge can be chosen into $\mathrm{Attr}_B^i(F)$. We can fix such a choice for each QB-state in $\mathrm{Attr}_B^{i+1}(F)$ ($i = 0,1, \ldots$) in order to build up a positional strategy. We include a state in Q_A in $\mathrm{Attr}_B^{i+1}(F)$ if all edges from it lead to $\mathrm{Attr}_B^i(F)$. The sequence $\mathrm{Attr}_B^0(F) \subseteq \mathrm{Attr}_B^1(F) \subseteq \mathrm{Attr}_B^2(F) \subseteq \ldots$ becomes stationary for some index $\ll$ IQI. We define $\mathrm{AttrB(F)} := \bigcup_{i=0}^{|Q|} \mathrm{Attr}_B^i(F)$.

Later we shal! also use the set Attr, (F), defined in the analogous way for player A.

With the inductive construction it was explained that $\mathrm{Attrj}, (F) \subseteq W_B$; furthermore we have defined a uniform positional winning strategy which can be applied to any state in W_B regardless of the start of the play. (For states in $Q_B \cap F$ the choice of the next state is arbitrary.)

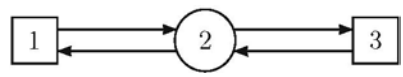
For the converse inclusion $W_B \subseteq \mathrm{AttrB(F)}$ we show that AttrB(F) exhausts the winning region W_B. For this, we show that from each state in the complement of AttrB(F), player A has a winning strategy (which is again positional). It suffices to verify that from any state q in $Q \setminus \mathrm{AttrB(F)}$ player A can force to stay outside $\mathrm{Attrj}, (F)$ also in the next step. This is checked by a case distinction: If $q \in Q_A$, there must be an edge back into $Q \setminus \mathrm{Attrj}, (F)$, otherwise all edges from q would go to $\mathrm{Attrj}, (F)$ whence q

would belong to Attr$_B$(F). If $q \in Q_B$, all edges from q must lead to $Q \setminus$ AttrB(F), because otherwise there would be an edge to Attrj, (F) and q would again belong to AttrB(F). Q.E.D.

4.2 Weak Muller games

In a weak Muller game $(C, \mathcal{F})$, player B wins the play p iff $Occ(p) \in \mathcal{F}$, i.e., the states visited during p form a set in $\mathcal{F}$. **It** is a useful exercise to verify that weak Muller games are precisely those where the winning condition can be expressed as a Boolean combination of reachability conditions.

Positional strategies do not suffice in general to win weak Muller games. As an example, consider the following game graph and the weak Muller condition given by $\mathcal{F} = \{\{1, 2, 3\}\}$ (requiring that player B should visit all states in order to win).



From vertex 2 there is no positional winning strategy: Neither the choice to move to I nor the choice to move to 3 will enable us to reach each vertex. On the other hand, a one-bit memory will do: When coming back to 2 we should know whether I or 3 was visited before, and then we should move to 3, respectively 1. A general principle derivable from this solution is to "rernember where we have been already". This principle corresponds to a simple experience of every-day life: When there is a task ahead consisting of several items, keep a list of what was done already (and thus of what still has to be done).

We shall see that this idea suffices completely for setting up the transition structure of a finite-state winning strategy. Given a weak Muller game $(C, \mathcal{F})$ with C = (Q, QA, E) and $\mathcal{F} = \{F_1, \ldots, F_k\}$, $F; \subseteq Q$, we define the transition structure of a Mealy automaton S with the power set of Q as its set of states and Q as its input alphabet. Having read the input word $ql \ldots qk$, its state will be $\{ql, \ldots, q_k\}$. So the initial state is 0 and the transition function $\delta : 2^Q \times Q \to 2^Q$ is defined by $\delta(R, p) = Ru\{p\}$. This memory of subsets of Q with the mentioned update rule is called *appeamnce record.* We shall see that this memory structure suffices for winning strategies in arbitrary weak Muller games over C. What remains is to fix the output function. For this purpose we study an expanded game into which the memory contents from 2^Q are incorporated. **It** will turn out that-based on this extra information-the winning condition can be reformulated for the expanded game. We call this transformation of the game a "game simulation". For the new game we shall provide *positional*

winning strategies, which will supply the desired output function for the strategy automaton S.

4.3 Game simulation

During a play p, the set of visited states increases weakly monotonically and finally reaches the value $Occ(p)$ on which it stays fixed. Similarly the cardinality of the set of visited states increases until it reaches the value IOcc(p)l. This observation enables us to reformulate the weak Muller winning condition "$Occ(p) \in F$". We associate a number $c(R)$ with each subset R of Q, also called its color, which conveys two pieces of information: the size of R, and whether R belongs to $\mathcal{F}$ or not. **In** the first case, we take the even color $2 \cdot$ IRI, otherwise the odd color $2 \cdot$ IRI - 1. Let

$$c(R) := \begin{cases} 2 \cdot \text{IRI} & \text{if } R \in \mathcal{F} \\ 2 \cdot \text{IRI} - 1 & \text{for } R \notin \mathcal{F} \end{cases}$$

for $R \neq 0$ and set $c(0) := 0$. Then the following claim is obvious:

Remark 4.2. Let p be a play and $Ra, Rl, R_2, \ldots$ be the value sequence of the associated appearance records. Then $Occ(p) \in \mathcal{F}$ iff the maximal color in the sequence $c(R_0)c(R_1)c(R_2)\ldots$ is even.

This remark motivates a new winning condition over game graphs G = (Q, QA, E) that are equipped with a coloring c: $Q \to \{0, \ldots, k\}$. The *weak paTity condition* with respect to coloring c says: Player B wins the play $p = TOT1T2\ldots$ iff the maximum color in the sequence $c(r_0)c(r_1)c(r_2)\ldots$ is even. Given a game graph G and a coloring c with the weak parity winning condition, we speak of the *weak paTity game* (G, c).

Using the idea above, one transforms a weak Muller game $(G, \mathcal{F})$ into a weak parity game (G', c): Given G = (Q, QA, E) let $G' = (2^Q \times Q, 2^Q \times QA, E')$ where $((P,p), (R,T)) \in E'$ iff $(p,T) \in E$ and $R = PU\{p\}$, and for nonempty R define $c(R, T) := 2 \cdot$ IRI if $R \in \mathcal{F}$, otherwise $2 \cdot$ IRI - 1 (and let $c(0, T) = 0$).

Each play $p = TOTl\ldots$ in Ginduces the play $\rho' = (0, TO)(\{Ta\}, r_1)\ldots$ in G', which is built up according to the definition of E', We have by construction that p satisfies the weak Muller condition w.r.t. $\mathcal{F}$ iff p' satisfies the weak parity condition w.r.t. c.

This transformation of $(G, \mathcal{F})$ into (G', c) (with a change of the winning condition) is a "game simulation". **In** general, we say that the game (G, W) with G = (Q, QA, E) is simulated by (G', Wl) with $G' = (Ql, Q'_A, E')$ if there is a finite automaton $S = (S, Q, sa, \delta)$ without final states such that

- $Ql = S \times Q, Q'_A = S \times QA$,

- *((r,p), (s, q))* ∈ *EI* iff *(p, q)* ∈ *E* and $\delta(r, p) = s$ (which means that a play $p = qOql \ldots$ in C induces the play $pi = (sa, qo)(5(so, qo), q_1) \ldots$ over *CI),*
- a play pover C belongs to *W* iff the cortesponding play *pi* over *CI* belongs to W'.

Ifthese conditions hold we write $(C, W) \leq_S (CI, WI)$.

This relation has an interesting consequence when the latter game allows positional winning strategies. Namely, positional strategies over *CI* are easily translated into finite-state strategies over C: The automaton *S* used for the simulation realizes such a strategy when equipped with an output function that is obtained from the positional strategy over *CI* = $(S \times Q, S \times QA, E')$.

Remark 4.3. Let $S = (S, Q, sa, \delta)$ and assume $(C, W) \leq_S (CI, W^{I)}$. If there is a positional winning strategy for player B in *(CI, WI)* from (sa, q), then player B has a finite-state winning strategy from *q* in *(C, W).* The analogous claim holds for player A.

Proo]. Consider the case of player B. We extend the automaton *S* by an output function that is extracted from the winning strategy $\sigma : Q'_B \to QI$. It suffices to define $\tau : S \times Q_B \to Q$ by $\tau(s, q) :=$ second component of $\sigma(s, q)$. Then any play *p* according to the strategy *S* belongs to *W* iff the corresponding play *pi* (obtained as defined via *S)* belongs to W_I. Since σ was assumed to be a winning strategy, so is the strategy executed by *S.* The case of player A is handled analogously. Q.E.D.

We apply this remark for the concrete simulation of weak Muller games by weak parity games mentioned above. We show "positional determinacy" for weak parity games and thus-by the preceding remark-finish the proof of part (a) of Theorem 3.1, concerning weak Muller games.

Theorem 4.4. A weak parity game *(C,* c) is determined, and one can compute the winning regions *WA, W*Band also construct corresponding positional winning strategies for the players A and B.

It may be noted that we suppressed the initial states *q* when speaking about positional winning strategies. **In** the proof we shall see that-as for reachability games-the strategies can be defined independently of the start state (as long as it belongs to the winning region of the respective player).

Proo]. Let C = (Q, Q_A, E) be a game graph (we do not refer to the special graph *CI* above), c: $Q \to \{0, \ldots, k\}$ a coloring (w.l.o.g. *k* even). Set ei = $\{q \in Q \mid c(q) = i\}$.

We first compute the attractor for B of the states with maximal color, which is even. When player B reaches such a state the play is won whatever happens later. So A_k := AttrB(C_k) is a part of the winning region of player B.

The remaining nodes form the set $Q \setminus A_k$; this is again a game graph. Note that from each state q in $Q \setminus A_k$ there is at least one edge back to $Q \setminus A_k$, otherwise (as seen by case distinction whether q E QA or q E QB) q would belong to A_k = AttrB(C_k).

In the subgame induced by $Q \setminus A_k$, we compute A_{k-1} := AttrA($C_{k-1} \setminus A_k$); from these vertices player A can reach the highest odd color k - 1 and guarantee to stay away from A_k, in the same way as explained above for reachability games (see Section 4.1).

In both sets we can single out positional winning strategies, over A_k for B, and over A_{k-1} for A. **In** this way we continue to adjoin "slices" of the game graph to the winning regions of Band A in alternation. The next set A_{k-2} is the set of all states q E $Q \setminus (A_{k-1} \cup A_k)$ from which player B can force the play to $C_{k-2} \setminus (A_{k-1} \cup A_k)$. We denote this set by $\mathrm{Attr}_B^{Q \setminus (A_{k-1} \cup A_k)}(C_{k-2} \setminus (A_{k-1} \cup A_k))$. The exponent indicates the set of states that induces the subgame in which the attractor computation takes place. **In** order to facilitate the notation for the general case, set Qi := $Q \setminus (A_{i+1} \cup \ldots \cup Ak)$.

So we compute the sets $A_k, A_{k-1}, \ldots, Aa$ inductively as follows:

$$A_k := \mathrm{AttrB}(C_k)$$
$$A_{k-1} := \mathrm{Attr}_A^{Q_{k-1}}(C_{k-1} \setminus A_k)$$

and for i = k - 2, ... , 0:

$$A_i := \begin{cases} \mathrm{Attr}_B^{Q_i}(C_i \setminus (A_{i+1} \cup \ldots \cup A_k)) & \text{if i even} \\ \mathrm{Attr}_A^{Q_i}(C_i \setminus (A_{i+1} \cup \ldots \cup A_k)) & \text{if i odd} \end{cases}$$

The positional strategies for A and Bare chosen as explained for the initial cases A_k, A_{k-1}. Now we have

$$W_B = \bigcup_{\text{i even}} A_i \quad \text{and} \quad W_A = \bigcup_{\text{i add}} A_i$$

For the correctness, one verifies by induction on j = 0, ... ,k:

$$\bigcup_{\substack{i=k-\mathrm{j} \\ \text{i even}}}^{k} A_i \subseteq W_B \qquad \bigcup_{\substack{i=k-j \\ \text{i add}}}^{k} A_i \subseteq W_A$$

We do not give this proof in detail; it is done in analogy to the case of reachability games (Section 4.1). Q.E.D.

Returning to the solution of weak Muller games, we first note that the claim of Theorem 3.1(a) on the memory size of a finite-state winning strategy (2^n memory states over a game graph with n states) is clear from the game simulation using the structure of appearance record. It is remarkable that this method yields a finite-state winning strategy (for either player) where the transition structure depends solely on the underlying game graph; the winning condition given by the family $\mathcal{F}$ enters only later in the definition of the output function.

5 Muller games and parity games

5.1 An example

As a preparation of the proof of the Büchi-Landweber Theorem 3.1(b), we consider a game that was introduced by Dziembowski et al. (1997). The game is parametrized by a natural number n; we consider here the case $n=4$.

The game graph consists of A-states 1,2,3,4 (called number-states) and B-states A, B, C, D (called letter-states). There is an edge from each A-state to each B-state and conversely.

The winning condition for B is the following, for a play p: *The number of letter-statee occurruiq infinitely aften in p has to coincide with the highest number that* OCCUTS *infinitely aften among the nurnber-siaies in p.* More formally we can write IInf(p)n{A, B, C, D}I = max(Inf(p)n{1, 2, 3,4}). Note that this defines a Muller game; the family $\mathcal{F}$ of accepting loops contains each set R such that IR n {A, B, C, D}I = max(R n {1, 2, 3,4}).

It is the job of player A to choose letter-states. H, for instance, player A decides aftel' some time to stick just to the letters A and D (in some order) and not to visit Band C anymore, then player B should infinitely often piek state 2 and only finitely often the larger states 3 and 4.

From a naive point ofview it is hard to imagine how player B can win this game. Aftel' a finite play prefix, nothing about the set Inf(p) n {A, B, C, D} is decided (in fact, player A has complete freedom to go for any nonempty subset of {A, B, C, D}). However, a strategy has to select one number vertex on the basis of the cutrent finite play prefix alone.

Nevertheless, player B wins this game from each of the states, and the winning strategy illustrates again that for appropriate decisions on the future it may be sufheient to remember relevant facts from the past. We shall use a refined version of the appearance record, in which not only the visited states, but also their *order of last visits* is taken into account. **In** the present example, it suffices to record the list of previously visited letter-states in

the order of their last visits-most recently visited states noted first. If the current (letter-) state was already visited before, then it is shifted from its previous position, say at place h in the list, to the front. The position h from which it was taken is underlined; we call it the "hit". This structure was introduced by McNaughton (1965) under the name "order-vector". Later Gurevich and Harrington suggested in their fundamental paper (Gurevich and Harrington, 1982) the name "latest appearance record" (LAR) under which the structure is known today.

Let us study an example. Suppose player A picks successively the letter-states A, C, C, D, B, D, C, D, D, ... We note this sequence on the left, and the associated sequence of latest appearance records on the right:

Visited letter	Reached **LAR**
A	(A)
C	(CA)
C	$\underline{(CA)}$
D	(DCA)
B	$(BDCA)$
D	$\underline{(DBCA)}$
C	$\underline{(CDBA)}$
D	$\underline{(DCBA)}$
D	$\underline{(DCBA)}$

Now assume that player A indeed sticks to the states C and D and repeats these two infinitely often. Then the states A and B will finally stay on the last two LAR-positions and not be touched anymore. Thus the hit value will be only 1 or 2 from some point onwards, and the maximal hit value visited infinitely often will be 2. In fact, if only position 1 is underlined from some point onwards, then only the same letter would be chosen from that point onwards (and not two states C and D as assumed).

We conclude that player B should always move to the number state named by the cutrent hit value. In the scenario mentioned, this would mean to move finally only to states 1 or 2, and to 2 infinitely often. If at some point player A would decide to go to one state only, this state would be repeated at the head of the **LAR** and underlined; so the maximal hit value visited infinitely often would be 1 (and correct again).

We leave it to the reader to show that "to move to the number given by the cutrent hit value" is a winning strategy of player B in the game (see also Remark 5.1 below). Since the required memory is finite and the update rule is defined in terms of the previous **LAR** and the current state, this is a finite-state winning strategy.

The example suggests a solution of Muller games in very close analogy to the case of weak Muller games, using the latest appearance record in place

of the appearance record. We shall introduce the LAR-structure in general (i.e., we take it to cover *all* states of the game graph under consideration and not only a subset, such as the letter-states in our example). From each LAR we extract an "index". Whereas in an appearance record we referred to its cardinality, we use the hit value for a LAR. Then we introduce the "parity condition" (a variant of the weak parity condition) as new winning condition and apply it in a game simulation. The solution of the parity game that arises in this way gives a solution of the original Muller game.

5.2 Parity games

Given a Muller game (G,F) with G = (Q,QA,E) and Q = {I, ...,n}, we define the transition structure of a finite-state machine S = (S, Q, sa, δ). Its state set is the set ofLAR's over Q, and its purpose is to realize, given a play prefix $i_1 \ldots ik \in Q^*$, the computation of the corresponding LAR. Formally, an LAR is a pair $((j_1 \ldots i\text{-}), h)$ where the $i_\cdot$ are pairwise distinct states from Q and $0 \leq h \leq r$. The initial state is sa = (( ),0) (empty list and hit 0). The transition function $\delta : S \times Q \to S$ realizes the update of the LAR as indicated in the example above: We set $\delta(((i_1 \ldots i_r), h), i) = ((i i_1 \ldots i_r), 0)$ if i does not occur in $(i_1 \ldots i_r)$. Otherwise, if i = ik we cancel i from $(i_1 \ldots i_r)$ to obtain $(j_1 \ldots jr\text{-}1)$ and set $\delta(((i_1 \ldots i_r), h), i) = ((ij_1 \ldots j_{r-1}), k)$.

An essential ingredient of a LAR $((i_1 \ldots i_r), h)$ is the *hit set* $\{i_1, \ldots, i_h\}$ of states listed up to and including the hit position h. Consider a play pover Q and the associated sequence of LAR's, denoted p', If h is the maximal hit assumed infinitely often in ρ', we may piek a position in ρ' where no unlisted state enters any more later in the play and where only hit values $\leq h$ occur afterwards. From that point onwards the states listed aftel' position h stay fixed, and thus also the hit set for the hit value h stays fixed. We call this set *the hit set fOT the maximal hit occurruiç infinitely often in pl.*

Remark 5.1. Let p be a sequence over Q and pi be the associated sequence of LAR's. The set Inf(p) coincides with the hit set H for the maximal hit h occurring infinitely often in *pl.*

Proo]. Consider the point in p from where no new states wil! occur and where al! visits of states that are visited only finitely often are completed. Aftel' a further visit of each state in Inf(p), these states wil! stay at the head of the LAR's (in various orders), and the hit values will be $\leq k$:= IInf(p)l. It remains to show that the hit value in *pi* reaches k again and again (so that k is the maximal hit occurring infinitely often in *pi).* If the hit was $< k$ from some point onwards, the state q listed on position k would not be visited later and thus not be in Inf(p). Q.E.D.

Using the remark, we can reformulate the Mul!er winning condition for the play *p: The hit set fOT the highest hit occurruiq infinitely often in pi*

belangs to $\mathcal{F}$. This allows us to extract two data from the LAR's which are sufheient to decide whether the play p satisfies the Muller condition: the hit value and the information whether the corresponding hit set belongs to $\mathcal{F}$. We combine these two data in the definition of a coloring of the LAR's. Define, for $h > 0$,

$$c(((i_1 \ldots i_r), h)) := \begin{cases} 2h & \text{if } \{i_1, \ldots, i_h\} \in \mathcal{F} \\ 2h-1 & \text{if } \{i_1, \ldots, i_h\} \notin \mathcal{F} \end{cases}$$

and let $c(((i_1 \ldots i_r), 0)) = 0$.

Then the Muller condition $\mathrm{Inf}(p) \in \mathcal{F}$ is satisfied iff the maximal color occurring infinitely often in $C(p|(0))C(p|(1)) \ldots$ is even. This is a "parity condition" (as introduced by Mostowski (1984) and Emerson and Jutla (1991)). The only difference to the weak parity condition is the reference to colors occurring infinitely often rather than those which occur at all.

In general, the parity condition refers to a coloring $c : Q \to \{0, \ldots, k\}$ of a game graph C; it is the following requirement on a play p:

$$\bigvee_{j \text{ even}} (\exists^\omega i : c(p(i)) = j \ \wedge \neg \exists^\omega i : c(p(i)) > j)$$

The pair (C, c) with this convention for the winning condition for player B is called a *paTity game*.

Similar to the case of weak Muller games, one can set up a game simulation of a Muller game $(C, \mathcal{F})$ bya parity game $(C|, c)$: We use the finite-state machine S introduced before that transforms a given play pover C into the corresponding sequence $p|$ of LAR's (realized in the states visited by S), and we use the coloring c defined above. The game graph $C|$ is fixed as in Section 4.3 above, using the new machine S. We obtain the game simulation $(C, \mathcal{F}) \leq_S (C|, c)$ where $(C|, c)$ is a parity game.

Remark 5.2. There is a variant of Sin which some ofthe states are spared. We cancel the initial LAR's (corresponding to hit value 0), starting (over states $1, \ldots, n$) with the LAR $((1 \ldots n), 1)$ rather than $(( ), 0)$, and keeping the update rule as before. With this change, one cannot distinguish between first and repeated visits of states, but clearly this loss of information is inessential for the satisfaction of the winning condition. The number of states of the reduced machine is then $n! \cdot n$ over a graph with n states.

One can use S as the transition structure of automata realizing winning strategies in the Muller game $(C, \mathcal{F})$. In order to provide also the output function, we have to solve parity games, again by positional winning strategies.

Theorem 5.3. A parity game (C, c) is determined, and one can compute the winning regions W_A, W_B and also construct corresponding positional winning strategies for the players A and B.

Proof. Given $C = (Q, Q_A, E)$ with coloring $c: Q \to \{0, \ldots, k\}$ we proceed by induction on $|Q|$, the number of states of C.

The induction start (Q is a singleton) is trivial!. **In** the induction step assume that the maximal color k is even (otherwise switch the roles of players A and B). Let q be a state of the highest (even) color k and define $A_a = \text{Attr}_j(\{q\})$. As the complement of an attractor, the set $Q \setminus A_a$ induces a subgame. The induction hypothesis ensures a partition of $Q \setminus A_a$ into the winning regions U_A, U_B of the two players (with corresponding positional winning strategies) in this subgame.

We now distinguish two cases:

1. From q, player B can ensure to be in $U_B \cup A_a$ in the next step,
2. From q, player A can ensure to be in U_A in the next step.

Let us first verify that one of the two cases applies (which gives a kind of local determinacy). Assume Case 1 fails. **If** $q \in Q_B$, then all transitions from q have to go to U_A, otherwise we would be in Case 1. By the same reason, if $q \in Q_A$, then some transition from q goes to U_A; so Case 2 applies.

In Case 1, one shows $W_B = U_B \cup \text{Attr}_B(\{q\})$ and $W_A = U_A$, applying the positional strategies of the induction hypothesis over U_A, U_B, the attractor strategy over $\text{Attr}_j(\{q\})$, and (if $q \in Q_B$) the choice of the next state from q according to Case 1. For the first claim, note that a play in $U_B \cup \text{Attr}_j(\{q\})$ either remains in U_B from some point onwards, whence Player B wins by induction hypothesis, or it visits (by choice of player A) the attractor A_a and hence q again and again, so that player B wins by seeing the highest color (even!) repeatedly. The second claim $W_A = U_A$ is now clear by induction hypothesis.

We turn to Case 2. **In** this case we know that $q \in \text{Attr}_A(U_A)$ and consider the set $\text{Al} = \text{Attr}_A(U_A \cup \{q\})$, clearly of cardinality > 1. So we can apply the induction hypothesis to the subgame induced by $Q \setminus \text{Al}$. We obtain a partition of this domain into winning regions V_A, V_B for A and B, with corresponding positional winning strategies. Now it is easy to verify $W_B = V_B$ and $W_A = V_A \cup \text{Al}$, with positional winning strategies again provided by the induction hypothesis and the attractor strategy over Al.

Finally we note that the inductive construction can be turned to a recursive procedure which produces, given C and the coloring c, the desired winning regions and positional strategies. Q.E.D.

The recursive procedure appearing in this proof involves a nested call of the inductive hypothesis, which means that for each induction step the

computational effort doubles, resulting in an overall exponential runtime. It is known that the problem "Given a parity game (C, c) and a state q, does q belong to the winning region of B?" is in the complexity class NP nco-NP. Whether this problem is decidabie in polynomial time is one of the major open problems in the algorithmic theory of infinite games.

As mentioned above, Theorem 5.3 on positional determinacy of parity games completes the solution of Church's Problem. The claim on the numbel' of states of a finite-state winning strategy $(n! \cdot n$ memory states over a graph with n states) is clear from Remark 5.2. As shown in (Dziembowski et al., 1997), the factorial function also supplies a Iower bound on the memory size of winning strategies in Muller games.

It is worth noting that the claim on positional determinacy of parity games also holds for infinite game graphs (however, without a statement on computability of winning strategies). This "infinite version" of the theorem can be applied for the complementation of automata over infinite trees (see Thomas, 1997).

6 Conclusion

Let us reeall the three major steps for a solution of Church's Problem: First we relied on a translation from the logic SIS to Muller automata, which were then changed into game graphs with Muller winning condition. From Muller games we constructed parity games via the **LAR** structure; and finally we presented a solution of parity games. All three steps are nontrivial. As mentioned, the first step involves a non-elementary blow-up (from length of formula to size of automaton). For each of the other two steps, an exponential time procedure was presented; a direct construction is possible, however, resulting in a single exponential altogether (see Zielonka, 1998). On the other hand, our two-step approach showed that finite-state winning strategies for a Muller game over a graph C can be constructed with a transition structure that depends on C alone, and that only for the output function the winning condition has to be invoked.

Church's Problem and its solution were the starting point for a highly active area of research in computer science, first restricted to pure automata theory, but in the last 20 years with a great inftuence in algorithmic verification and program synthesis. A problem in cutrent research is to find classes of infinite game graphs over which games with MSO-definable winning conditions can still be solved algorithmically. Some results (on so-called pushdown graphs) are mentioned in (Grädel et al., 2002). Another direction is to modify or to generalize the specification language in Church's Problem (see, e.g., Rabinovich and Thomas, 2007). **In** a wider context, more general models of games are studied, for instanee "concurrent games" (where the two players move simultaneously), "timed games" (generalizing the model

of timed automata), stochastic games (in which random moves enter), and multiplayer games.

Acknowledgments

Thanks are due to Erich Grädel, Wong Karianto, Detlef Kähler, Christof Löding, and Michaela Slaats for their helpful comments on a previous version of this paper.

References

Büchi, .LR. (1962). On a decision method in restricted second order arithmetic. **In** Nagel, K, Suppes, P. & Tarski, A., eds., *Logic, methodology and philosophy of science. Proceedinqs of the 1960 International Conqress,* pp. 1-11. Stanford Uni versity Press.

Büchi, .LR. & Landweber, L.H. (1969). Solving sequential conditions by finite-state strategies. *Tmnsactions of the AmeTican Mathematical Society,* 138:295-311.

Church, A. (1957). Applications of recursive arithmetic to the problem of circuit synthesis. pp. 3-50. Mimeographed note, distributed as *Summaries of talks presenied at the Surnmer Institute fOT symbolic logic: Cornell Unioersits] 1957.*

Church, A. (1963). Logic, arithmetic, and automata. **In** Stenström, V., ed., *Proceedinqs of the International Conqress of Mathematicians,* 15-22 *August* 1962, pp. 23-35, Djursholm, Sweden. Institut Mittag-Leffier.

Dziembowski, S., Jurdziriski, M. & Walukiewicz, I. (1997). How much memory is needed to win infinite games? **In** Winskei, *G.,* ed., *Proceedinqs. 12th Annual IEEE Symposium on Logic in Computer Science (LICS*)97). *WaTsaw, Foland, June 29-July* 2, 1997, pp. 99-110. **IEEE** Computer Society Press.

Emerson, A. & Jutla, C. (1991). Tree Automata, Mu-Calculus and Determinacy. **In** Sipser, M., ed., *Proceedinçs of 32nd Annual IEEE Symposium on Foundations of Computer Science (FOCS)91),* pp. 368-377.

Grädel, K, Thomas, W. & Wilke, T. (2002). *Automata, Logics and Infinite Games,* Vol. 2500 of *Leciure Notes in Computer Science.* Springer.

Gurevich, Y. & Harrington, L. (1982). Trees, automata, and games. **In** Lewis, H.R., Simons, B.B., Burkhard, W.A. & Landweber, L., eds., *Preeeedings of the Fourieenili Annual ACM Symposium on Theoru of Computing. San Francisco, Califomia. May* 5-7, 1982, pp. 60-65. ACM Press.

McNaughton, R. (1965). Finite-state infinite games. Project MAC Report, MIT.

McNaughton, R. (1966). Testing and generating infinite sequences by a finite automaton. *Injormaiion and* Co*nirol,* 9:521-530.

McNaughton, R. (1993). Infinite games played on finite graphs. *Anna!s of Pure and Applied Logic, 65(2):149-184.*

Moschovakis, Y. (1980). *Descripiioe Set Theoru,* Vol. 100 of *Studies in Logic and the Foundations of Mathematics.* North-Holland, Amsterdam.

Mostowski, A.W. (1984). Regular expressions for infinite trees and a standard form of automata. *Computation Theoru,* pp. 157-168.

Rabin, M.O. (1969). Decidability of second-order theories and automata on infinite trees. *Transaciions ofthe Amer-ican Mathematica! Society, 141:1-35.*

Rabin, M.O. (1972). *Automata on infinite objects and Churcli's Problein.* American Mathematical Society, Providenee **RI.**

Rabinovich, A. & Thomas, W. (2007). Logical refinements of Church's Problem. **In** Duparc, .l. & Henzinger, T.A., eds., *Proceedinçs of Computer-Science Logic, 21st International Workshop, CSL 2007, 16th Annua! Con-fer-ence of the EACSL, Lausanne,* Vol. 4646 of *Leciure Notes in Computer-Science,* pp. 69-83. Springer.

Thomas, W. (1995). On the synthesis of strategies in infinite games. **In** Mayr, KW. & Puech, **C.,** eds., *STACS* 95, *12th Annua! Symposium on Theor-etica! Aspects of Computer- Science, Murucli, Germany, Mareli 2-4, 1995, Proceeduiqs,* Vol. 900 of *Leciure Notes in Computer- Science,* pp. 1-13. Springer.

Thomas, W. (1997). Languages, automata, and logic. **In** Rozenberg, G. & Salomaa, A., eds., *Handbook of Formal Languages, Vol.* 3, pp. 389-455. Springer.

Zielonka, W. (1998). Infinite games on finitely coloured graphs with applications to automata on infinite trees. *Theor-etica! Computer- Science,* 200(1-2):135-183.

Modal Dependence Logic

Jouko Väänänen

institute for Logic, Language and Computation
Universiteit van Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam, The Netherlands
J.A.Vaananen@uva.nl

Abstract

We introduce a modal language which involves the concept of *dependence.* We give two game-theoretic definitions for the semantics of the language, and one inductive, and prove the equivalence of all three.

1 Introduction

Is it possible that in the future currency exchange rates depend only on government decisions? It is perhaps possible, but it is certainly not necessary. **In** (Väänänen, 2007) we outlined the basics of the logic of dependence. **In** this paper we take it upon ourselves to start a study of the logic of "possible dependence" .

Ey *dependenee* we mean dependence as it occurs in the foliowing contexts: Dependence of

- a move of a player in a game on previous moves
- an attribute of a database on other attributes
- an event in history on other events
- a variabie of an expression on other variables
- a choice of an agent on choices by other agents.

We claim that there is a coherent theory of such dependence with applications to games, logic, computer science, linguistics, economics, etc.

There is an earlier study of the closely related concept of independence in the form of the independence friendly logic, by Jaakko Hintikka (1996). **In** that approach independence is tied up with quantifiers. We find dependence a more basic and a more tractable concept than independence. AIso, we find that dependence (or independence) is not realiy a concept limited to quantifiers but a more fundamental property of individuals. Likewise, we do

Krzysztof R. Apt, Robert van Rooij (eds.). *New Perspectives on Games and Interaction.* Texts in Logic and Games 4, Amsterdam University Press 2008, pp. 237-254.

not study here dependence or independence of modal operators from each other.

The basic concept of our approach to the logic of dependence is the dependence atom:

$$=(p_1, \ldots, p_n, q). \tag{1.1}$$

with the intuitive meaning that q depends only on $p_1, \ldots, p_n$. The quantities $p_1, \ldots, p_n$ and q can be propositions or individuals, in this paper they are propositions.

Definition 1.1. The *modal language of dependence* has formulas of the form:

1. $P, q, \ldots$ proposition symbols
2. $=(p_1, \ldots, p_n, q)$ meaning "q depends only on $p_1 \cdots p_n$"
3. $A \vee B$
4. $\neg A$
5. $\Diamond A$

The logical operations DA (i.e., $\neg\Diamond\neg A$) and $A \wedge B$ (i.e., $\neg A \vee \neg B$), $A \to B$ (i.e., $\neg A \vee B$), $A \leftrightarrow B$ (i.e., $(A \to B) \wedge (B \to A)$), are treated as abbreviations.

The intuition is that a set of nodes of a Kripke structure satisfies the formula $=(p_1, \ldots, p_n, q)$ if in these nodes the truth value of q depends only on the truth values of $p_1, \ldots, p_n$. Note that this criterion really assumes, as emphasized in a similar context in (Hodges, 1997), a *set* of nodes, for one cannot meaningfully claim that the propositional symbols true or false in one single node manifest any kind of dependence. Figures 1 and 2 give examples of dependence and lack of it.

We think of the sentence

$$\mathrm{DO}(=(p, q) \wedge A)$$

as being true in a Kripke structure if every node accessible from the root has access to a node with A in such a way that in these nodes q depends only on p. A practical example of such a statement could be:

> *Whatever decisions the govemments make in the next 10 yoors, it* is *possible that by the yoor 2050 the sea levels rise and whether the rise* is *over 50* cm *depends only on how many countries have reduced their greenhouse gas emissions.*

We define now the game-theoretical semantics of our modal dependence language:

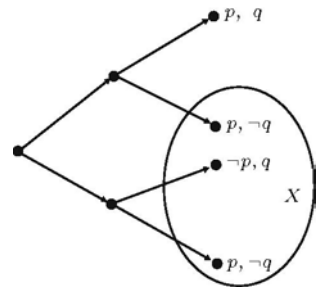


FIGURE 1. q depends only on P in X.

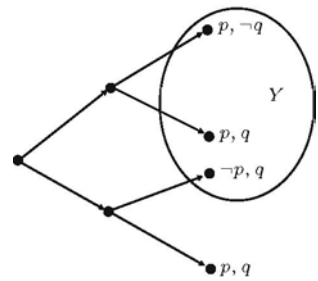


FIGURE 2. q does not depend only on pin Y.

Definition 1.2. The semantic game Cl*(A)* is defined as follows: Positions are of the form *(s, B, d),* where s is a node of the Kripke structure, *B* is a modal formula and *d* is a player (lor 11). **In** the beginning of Csem(A), played at *Sa,* the position is *(sa, A,* 11). The rules of the game are:

1. Position is *(s,p, d):* Player *d* wins if *p* is true in s, otherwise the opponent wins.
2. Position is *(s, =*(Pl, ... ,*Pn, q), d):* Player *d* wins.
3. Position is *(s,* $\neg A$*, d):* The next position is *(s, A, d*),* where *d** is the opponent of *d.*
4. Position is *(s, A* V *B, d):* Player *d* chooses C from *{A, B}.* The next position is *(s,* C, *d).*
5. Position is *(s,* $\Diamond A$*, d):* Player *d* chooses *a* node *si,* accessible from s. The next posi tion is *(si, A, d).*

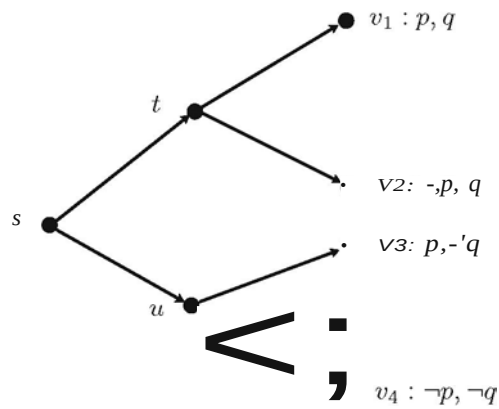


FIGURE 3. A Kripke model M.

A strategy σ of d is *uniform* if in any two plays where d uses σ and the game reaches a position $(s,=(Pi, \ldots ,Pn,q),d)$ in the first play and $(Si, =(Pi, \ldots ,Pn, q), d)$ in the second play, with the same subformula $=(Pi" " ,Pn, q)$ of A and the same truth values of Pi, ... ,Pn, the truth value of q is also the same. (By the "same subformula" we mean the same formula occurring in the same position in A.) **In** the extreme case of $=(p)$, the truth value of P has to be the same every time the game ends in position $(s, =(p), d)$ with the same $=(p)$.

Note that the game $G_{sem}(A)$ is determined and a perfect information game. Thus one of the players has always a winning strategy. However, there is no guarantee that this winning strategy is uniform (see Section 2). Thus the requirement of uniformity changes the nature of the game from determined to non-determined. **In** a sense the game loses the perfect information characteristic as the player who counts on a dependence atom $=(Pi, \ldots ,Pn, q)$ being true has to choose the possible worlds without looking at other parameters than Pi, ... ,Pn, as far as the truth of q is concerned. Rather than putting explicit information related restrictions on the moves of the players, we simply follow how they play and check whether the moves *seem* to depend on parameters not allowed by the winning positions $(s, =(Pi, \ldots ,Pn, q), d)$. **In** a sense, a player is allowed to know everything all the time, but is not allowed to use the knowiedge.

Definition 1.3. A is true at a node s if player 11 has a uniform winning strategy in the game $G_{sem}(A)$ at s.

The sentences

$$\Diamond\Box q$$
$$\Diamond\Box{=}(p, q)$$
$$\mathrm{OO}(\mathrm{p}\ \mathrm{V}\ \text{-},p)$$

are all true at the root of the Kripke model of Figure 3. By the definition of the meaning of the negation, -,A is true in a node s if and only if player I has a uniform winning strategy in position *(s, A,* 11). Ey a *logical consequence* $A \Rightarrow B$ in this context we mean that the formula B is true in every Kripke model at every node where A is true. Respectively, $A \Leftrightarrow B$ means that both $A \Rightarrow \mathrm{Band}\ B \Rightarrow A$ hold. Finally, A is called valid if it is true in every Kripke structure at every node.

Example 1.4.

1. $A \wedge (A \rightarrow B) \Rightarrow B$
2. $A \Rightarrow (B \rightarrow A)$
3. $(A \rightarrow (B \rightarrow C)) \wedge (A \rightarrow B) \Rightarrow A \rightarrow \mathrm{C}$
4. -,$B \rightarrow$ -,$A \Rightarrow A \rightarrow B$
5. $A \vee B \Leftrightarrow B \vee A$
6. $A \wedge B \Leftrightarrow B \wedge A$
7. $A \wedge A \Leftrightarrow A$
8. $A \wedge (B \wedge \mathrm{C}) \Leftrightarrow (A \wedge B) \wedge \mathrm{C}$
9. $A \vee (B \vee \mathrm{C}) \Leftrightarrow (A \vee B) \vee \mathrm{C}$
10. $=(p, q,T) \Leftrightarrow\ =(q,p,T)$
11. $(=(p, q) \wedge\ =(q, T)) \Rightarrow\ =(p, T)$
12. $=(p, T) \Rightarrow\ =(p, q,T)$
13. If A is valid, then so is $\mathrm{o}A$
14. $\mathrm{O}(A \rightarrow B) \wedge \mathrm{o}A \Rightarrow \mathrm{O}B$
15. $\mathrm{o}A \wedge \mathrm{O}B \Leftrightarrow \mathrm{O}(A \wedge B)$
16. $\Diamond A \vee \Diamond B \Leftrightarrow \Diamond(A \vee B)$

2 An example of non-determinacy

Consider the Kripke model M of Figure 3. The sentence

$$\Box\Diamond(p \leftrightarrow q)$$

is clearly true at the root 8 of the model, as both extensions of 8 have an extension in which p and q have the same truth value. On the other hand, the sentence

$$A: \quad \mathrm{DO}(=(\mathrm{p}) \wedge (p \leftrightarrow q))$$

is not true at the root for the following reason. Aftel' the move of player I, the node is t or u. Suppose it is t. Now player 11, in order not to lose right away, has to commit herself to $=(p)$ and the node with $p \wedge q$. Suppose the game is played again but Player I decides to move to node u. Now player 11 has to commit herself to $=(p)$ and the node with $\text{-}'p \wedge \text{-}'q$. At this point we see that the strategy that Player 11 is using is not uniform, for two plays have reached the same dependenee atom $=(p)$ with a different truth value for p. This contradiets the very definition of uniformity. However, the sentence

$$\text{-},A: \quad \text{-},\mathrm{DO}(=(\mathrm{p}) \wedge (p \leftrightarrow q))$$

is not true either, that is, neither does Player I have a uniform winning strategy in position $(8, A, 11)$. To see why this is so, let us assume I has a winning strategy (uniform or non-uniform) in position $(8, A, 11)$ and derive a contradiction. The position

$$(8, \mathrm{DO}(=(\mathrm{p}) \wedge (p \leftrightarrow q)), 11)$$

is actually the position

$$(8, \neg\Diamond\neg\Diamond(=(p) \wedge (p \leftrightarrow q)), 11),$$

from which the game moves automatically to position

$$(8, \Diamond\neg\Diamond(=(p) \wedge (p \leftrightarrow q)), I).$$

So in this position, according to the rules, Player I makes a move and chooses according to his strategy, say, t. We are in position

$$(t, \text{-},O(=(p) \wedge (p \leftrightarrow q)), I)$$

from which the game moves automatically to position

$$(t, O(=(p) \wedge (p \leftrightarrow q)), 11).$$

Now it is Player lI's turn to make a choice. We let her choose the node with $p \wedge q$. So we are in position

$$(\mathrm{Vl}, =(p) \wedge (p \leftrightarrow q), 11)$$

which leads to the position

$$(\mathrm{Vl}, \neg=(p) \vee \neg(p \leftrightarrow q), I).$$

Player I is to move. He does not want to play $\neg(p \leftrightarrow q)$ for that would lead to position

$$(v_1, \neg(p \leftrightarrow q), \mathrm{I}),$$

that is,

$$(Vl, p \leftrightarrow q, 11),$$

which is a winning position for Player 11. So Player I is foteed to play $\neg=(p)$, leading to position

$$(\mathrm{Vl}, \neg=(p), I),$$

that is,

$$(\mathrm{Vl}, =(p), 11).$$

But this is a winning position for Player 11, too. So again I has lost. If Player I moved *u.* instead of *t,* the argument would be essentially the same. So we may conclude that I simply does not have a winning strategy in position *(s,A,II)*. The game Gsem(A) is in this case *non-determined.*

We may conclude that the sentence $A \vee \neg A$ is not true at the root of *M*. Thus the *Law of Excluded Middle* is not valid in this logic. Also, the implication $A \rightarrow A$ is not valid. How can this be understood? The explanation lies in our game-theoretic concept of truth. For Player 11 to have a uniform winning strategy in position $(s, A \rightarrow \mathrm{A}, \mathrm{II})$, she has to count on herself or Player I having a uniform winning strategy in position $(s, A, 11)$. As we have seen, the game $G_{sem}(A)$ has no Gale-Stewart Theorem to guarantee it being determined. We have to give up-in the context of dependence logic-the idea that the meaning of $A \rightarrow B$ is that if A is true then B is true. Rather, we should think of $A \rightarrow B$ meaning that if Player I does not have a uniform winning strategy in Gsem(A), then Player 11 has a uniform winning strategy in Gsem(B).

3 A non-idempotency phenomenon

Consider the Kripke model *N* of Figure 4 and the sentence

$$B: \mathrm{D}=(\mathrm{p}).$$

It is clear that although Player 11 trivially wins every round of the game $G_{sem}(B)$ at *s,* she does not have a uniform winning strategy at *s,* because

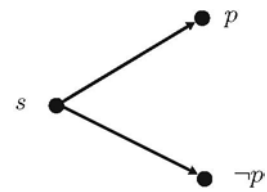


FIGURE 4. A Kripke model *N*.

depending on which extension of *s* Player I chooses, the value of *P* is true or false. On the other hand, Player I does have a uniform winning strategy, namely he simply plays the node with *P* during every round of the game.

Let us then look at

$$C\colon D(=(p) \vee =(p)).$$

Now Player 11 has a uniform winning strategy: ***Ij 1*** *plays the node with p, she plays the left disjunct, and otherunse the riqht disjunct.* So we have shown that

$$D(D \vee D) \not\Rightarrow DD.$$

4 Inductive truth definition

There is an alternative but equivalent truth definition, similar to the inductive truth definition of Hodges (1997) for Hintikka's **IF** logic. The basic concept here is a set *X* of nodes satisfying a formula, rather than a single node. We define:

- *p* is true in *X* if *p* is true in every node in *X*.
- *p* is true in *X* if *p* is false in every node in *X*.
- $=(P1,'''\ ,Pn, q)$ is true in *X* if any two nodes in *X* that agree about $pi, \ldots .P«$ also agree about *q*.
- $\neg=(p_1, . .\ ,Pn, q)$ is true in *X* if $X = 0$.
- $A \vee B$ is true in *X* if *X* is the union of a set where *A* is true and a set where *B* is true (see Figure 5).
- $A \wedge B$ is true in *X* if both *A* and Bare.
- $\Diamond A$ is true in *X* if *A* is true in some set Y such that every node in *X* has an extension in Y (see Figure 5).
- DA is true in *X* if *A* is true in the set consisting of all extensions of all nodes in *X* (see Figure 5).

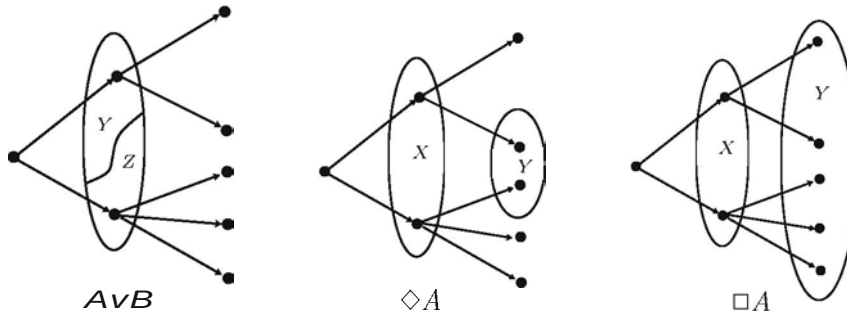


FIGURE 5. Truth definition.

More formally:

Definition 4.1. A is true in X if and only if $(X, A, 11)$ ET, where the set T is defined as follows:

(Tl) (X,p,lI) E T iff pis true in every node in X.

(T2) $(X,p,$ I) E T iff $,p$ is true in every node in X.

(T3) $(X,$ =(Pl, ... ,$Pn,q),$ 11) ET iff any two nodes in X that agree about Pl, ... ,Pn also agree about q.

(T4) $(X,$ =(Pl, ... ,»«, $q),$ I) ET iff $X = 0$.

(T5) $(X, ,A, d)$ ET iff (X, A, d^*) ET

(T6) $(X, A$ V$B,$ 11) E T iff X is contained in the union of a set Y and a set Z such that $(Y, A,$ 11) ET and $(Z, B,$ 11) ET.

(T7) $(X, A$ V$B,$ I) E T iff X is contained in the intersection of a set Y and a set Z such that $(Y, A,$ I) ET and $(Z, B,$ I) ET.

(T8) $(X, \Diamond A,$ 11) ET iff $(Y, A,$ 11) ET for some set Y such that every node in X has an extension in Y.

(T9) $(X, \Diamond A,$ I) E T iff $(Y, A,$ I) E T for the set Y consisting of all extensions of all nodes in X.

An easy induction shows that, as shown in (Hodges, 1997):

Lemma 4.2.

1. (X, A, d) E Timplies (Y, A, d) E T for all Y $<$ X. *The Downward Closure Propertsj.*

2. $(X, A \wedge \neg A, 11) \in \mathrm{Timplies}\ X = 0$. *The Consistency Propertq.*

From the Downward Closure Property it follows that (T6) can be replaced by

(T6)' $(X, A \vee B, 11)$ ET iff X is the union of a set Y and a set Z such that $(Y, A, 11)$ ET and $(Z, B, 11) \in T$.

and (T7) can be replaced by

(T7)' $(X, A \vee B, \mathrm{I}) \in T$ iff (X, A, I) ET and (X, B, I) ET.

The way we defined the game $G_{sem}(A)$ there was always an initial node from which the game started. We can generalize the setup up a little by allowing a set X of initial nodes. A strategy of a player d in $G_{sem}(A)$ is a *winning stmtegy in X* if the player wins every game started from a position $(s, A, 11)$, where $s \in X$. The strategy is *uniform in X* if in any two plays Pl and $P2$, started from positions $(x1, A, 11)$ and $(x2, A, 11)$, with $x1, x2 \in X$, where d uses the strategy and the game reaches a position $(s,=(Pl, \ldots ,Pn,q),d)$, with the same $=(Pl, \ldots ,Pn,q)$ and the same truth values of Pl, ... ,*Pn*, the truth value of q is also the same. Thus a player has a uniform winning strategy (in the original sense) at s iff he or she has a uniform winning strategy in $\{s\}$.

Theorem 4.3. If in the game Gsem(A) Player 11 has a uniform winning strategy in the set X, then $(X, A, 11) \in T$, i.e., A is true in the set X.

Proof. Suppose 11 has a uniform winning strategy σ in $Gsem(A_a)$ in the set X_a. We prove by induction on subformulas A of Aa that if $\Gamma(A, d)$ denotes the set of nodes s such that position (s,A,d) is reached while $Gsem(A_a)$ is being played, 11 following σ, then $(\mathrm{r}(A, d), A, d) \in T$. This will suffice, for the initial position $(s, Aa, 11)$ can be reached for any $s \in X_a$ and so it will follow that Aa is true in X_a. When dealing with $\mathrm{r}(A, d)$ we have consider different occurrences of the same subformula of Aa as separate. So, e.g., $=(p)$ may occur in Aa in two different places and $\Gamma'(=(p), d)$ is computed separately for each of them.

Case **i:** $X = \mathrm{r}(p, 11)$. Since σ is a winning strategy, P is true at every $s \in X$. Thus $(X,p, 11) \in T$ by (Tl).

Case ii: $X = \mathrm{r}(p, \mathrm{I})$. Since σ is a winning strategy, $\neg p$ is true at every $s \in X$. Thus (X,p, I) ET by (T2).

Case **iii:** $X = r(=(Pl, \ldots ,Pn,q),II)$. Let us consider $s,t \in X$ that agree about Pl, ... ,*Pn*. Since σ is a uniform strategy, s and tagree about q. Ey (T3), $(X, =(\mathrm{Pl}, '\ldots ,»«, q), 11) \in T$.

Case **iv:** $X = \mathrm{r}(=(\mathrm{Pl}, \ldots ,Pn, q), \mathrm{I})$. Since σ is a winning strategy of 11, $X = 0$. Ey (T4), $(X, =(\mathrm{Pl}, \ldots ,t\text{-}: q), \mathrm{I})$ ET.

Case v: X = f($\neg A, d$). Note that X = f(A, d^*). Ey induetion hypothesis, (X, A, d^*) ET, and henee (X, $\neg A, d$) ET.

Case vi: X = f(A VB, 11). Note that $X \subseteq Y \cup Z$, where Y = f(A, 11) and Z = f(B,II). Ey induetion hypothesis, (Y, A, 11) E Tand (Z, B, 11) E T. Thus (X, A VB, 11) ET by (T6).

Case vii: X = f(A VB, I). Note that $X \subseteq Y \cap Z$, where Y = f(A, I) and Z = f(B, I). Ey induetion hypothesis, (Y, A, I) ET and (Z, B, I) ET. Thus (X, A VB, I) E T by (T7).

Case viii: X = f(OA, 11). For eaeh sEX there is some s_i reaehable from s that 11 ehooses aeeording to her winning strategy σ in position ($s, \Diamond A$, 11). Let Y be the set of all sueh s_i. Note that then Y $\subseteq$ f(A, 11). By induetion hypothesis, (f(A, 11),A, 11) E T. Ey (T8), (X, $\Diamond A$, 11) E T.

Case ix: X = I'($\Diamond A$, I). For eaeh sEX there may be some s_i reaehable from s that I eould ehoose in position ($s, \Diamond A$, I). Let Y be the set of all sueh possible s_i (i.e., Y is the set of all possible extensions of all SEX). Ey induetion hypothesis (Y, A, I) E T. Ey (Tg), (X, $\Diamond A$, I) ET. Q.E.D.

Corollary 4.4. If 11 has a uniform winning strategy in $G_{sem}(A)$ at s, then A is true in $\{s\}$.

5 **Truth** strategy

We define a new game $G_2(A)$, whieh we eall the *set game* as follows: Positions are of the form (X, B, d), where X is a set of nodes, B is a modal dependenee formula, and d is either I or 11. The rules of the game are as follows:

(SI) (X,p,II): Player 11 wins *ifp* is true at every node in X, otherwise I wins.

(S2) (X,p,I): Player 11 wins *ifp* is false at every node in X, otherwise I wins.

(S3) (X, =($po, \ldots, Pn, q$), 11): Player 11 wins if any two nodes in X that agree about Pl, ...P« also agree about q. Otherwise I wins.

(S4) (X, =($po, \ldots, Pn, q$), I): Player 11 wins if X = o, otherwise I wins.

(S5) (X, $\neg A, d$): The game eontinues from (X, A, d^*).

(S6) (X, A V B, 11): Player 11 ehooses Y and Z sueh that $X \subset Y \cup Z$. Then Player I ehooses whether the game eontinues from (Y, A, 11) or (Z, B, 11).

(S7) (X, A VB, I): Player 11 ehooses Y and Z sueh that $X \subset Y \cap Z$. Then Player I ehooses whether the game eontinues from (Y, A, I) or (Z, B, I).

(S8) *(X, ◇A,* 11): Player 11 ehooses a set Y sueh that every node in *X* has an extension in Y. The next position is *(Y, A, 11).*

(S9) *(X, ◇A,* I): The next position is *(Y, A,* I), where Y consists of every extension of every node in *X.*

An easy induetion shows that if Player 11 has a winning strategy in position *(X, A, d),* and $Y \subseteq X$, then she has in position *(Y, A, d),* too. From this faet it follows that (86) ean be replaeed by

(S6)' *(X, A* V *B,* 11): Player 11 ehooses Y and *Z* sueh that $X = Y \cup Z$. Then Player I ehooses whether the game eontinues from *(Y, A,* 11) or *(Z, B, 11).*

and (87) ean be replaeed by

(S7)' *(X, A* V *B,* I): Player I ehooses whether the game eontinues from *(X, A,* I) or *(X, B, I).*

Theorem 5.1. If *(X, A,* 11) E *T* (i.e., *A* is true in *X),* then Player 11 has a winning strategy in Gset(A) in position *(X, A, 11).*

Proof. Suppose that $(X_a, Aa,$ 11) E *T.* The strategy of 11 in *Gset*(A_a) is to play in sueh a way that if the play is in *Gset*(A_a) in position $P = (X, A, d)$, then $T(P) = (X, A, d)$ ET. In the beginning the position is $(X_a, Aa,$ 11) and indeed *Aa* is true at X_a. Aftel' this we have different cases before the game ends:

Case 1: $P = (X, \neg A, d)$. By assumption, $T(P) = (X, \neg A, d)$ ET. By (T5) (X, A, d^*) E *T.* Now the game eontinues from position $pi = (t, A, d^*)$ and $T(P^{I}) = (X, A, d^*)$ ET.

Case 2: $P = (X, A \vee B,$ 11). By assumption, $T(P) = (X, A \vee B,$ 11) ET. By (T6) there are Y and *Z* sueh that $X <$ YUZ, *(Y, A,* 11) ET and *(Z, B,* 11) E *T.* 80 11 plays Y and *Z* in Gset(A_a). Now I deeides whether the game eontinues from position *(Y, A,* 11) or from position *(Z, B,* 11). Whiehever the dec ision is, we have *(Y, A,* 11) ET and *(Z, B,* 11) ET.

Case 3: $P = (t, A \vee B,$ I). By assumption, $T(P) = (X, A \vee B,$ I) ET. By (T7)', *(X, A,* I) E Tand *(X, B,* I) E *T.* Now the set game eontinues from position *(X, A,* I) or from position *(Y, B,* I, aeeording to the deeision of I. Whiehever the deeision is, we have *(X, A,* I) ET and *(X, B,* I) ET.

Case 4: $P = (t, ◇A,$ 11). By assumption, $T(P) = (X, ◇A,$ 11) E *T.* By (T8), *(Y, A,* 11) E *T* for some set Y of nodes aeeessibie from nodes in *X.* This set Y is the ehoiee of *11* in Gset(A_a). Now the game eontinues from position $pi = (Y, A,$ 11) and $T(P^{I}) = (Y, A,$ 11) E *T.*

Case 5: P = $(t, \Diamond A, \mathrm{I})$. Ey assumption, $T(P)$ = $(X, \Diamond A, \mathrm{I})$ E T. Ey (Tg), (Y, A, I) E T for the set Y of all nodes accessible from nodes in X. Now the game continues from position pi = (Y, A, I) and $\tau(P')$ = (Y, A, I) E T.

At the end of the game Gset(A_a) we have to check that 11 indeed has won. There are again several cases:

Case 6: P = $(X,p,$ 11). Since $T(P)$ = (X,p,lI) E T, p is true at every t e X by (Tl). So 11 has won.

Case 7: P = (X, p, I). Since $T(P)$ = (X, p, I) E T, -,p is true at every t e X by (T2). So 11 has won.

Case 8: P = $(X, =(\mathrm{Pl}, \ldots, Pn, q),$ 11). Let s, t e X agree about $\mathrm{Pl}, \cdots, Pn$. Since $(X, =(\mathrm{Pl}, \ldots, Pn, q),$ 11) E T, we can conclude from (T3) that s and t agree about q. Player 11 has won.

Case 9: P = $(X, =(\mathrm{Pl}, \ldots, Pn, q),$ I). So $T(P)$ = $(X, =(\mathrm{Pl}, \ldots, P_n, q),$ I) E T. Ey (T4), X = o. Player 11 has won. Q.E.D.

6 Power strategy

We shall describe a strategy in Gsem(A) which is based on playing Gset(A) in the power set of the Kripke model, hence the name *power- stmtegy*. The advantage of playing in the power set is that we can in a sense play many games in parallel and use this to get a uniform strategy in $G_{sem}(A)$ (see Figure 6).

Theorem 6.1. If Player 11 has a winning strategy in Gset(A) in position (X,A,II), then in Gsem(A), she has a uniform winning strategy in X.

Proo]. Suppose σ is a winning strategy of 11 in Gset(A_a) in position $(X_a, Aa,$ 11). The strategy of 11 in $Gsem(A_a)$ is to play so that if the play is in position P = (t, A, d), then 11 is in the game Gset(A_a), playing σ, in position $T(P)$ = (X, A, d) with t EX. In the beginning the position in $Gsem(A_a)$ can be any (s,A_a,II), where s E X_a. In Gset(A_a) the initial position is $(Xa,A_a,$ 11). So whichever P = (s,A_a,lI) the game $Gsem(A_a)$ starts with, we can let $T(P)$ = $(X_a, Aa,$ 11). Aftel' this we have different cases before the game ends:

Case 1: P = $(t, -,A, d)$. Ey assumption, $T(P)$ = $(X, -,A, d)$ with t EX. Now the game continues from position pi = (t, A, d^*) in $G_{sem}(Aa)$ and from position $T(P^I)$ = (X, A, d^*) in Gset(A_a).

Case 2: P = $(t, A$ V$B,$ 11). Ey assumption, $T(P)$ = $(X, A$ V$B,$ 11) such that t EX. By (S6) the strategy σ gives two sets Y and Z such that $X \subseteq Y u Z$, the game Gset(A_a) continues from $(Y, A,$ 11) or $(Z, B,$ 11). Since t E Y U Z, we have either t E Y or t EZ. In the first case 11 lets C = A, U = Y and in the second case C = B, U = Z. Now the game $Gsem(A_a)$ continues from position P' = $(t,$ C, 11) and $T(P^I)$ = $(U,$ C, 11).

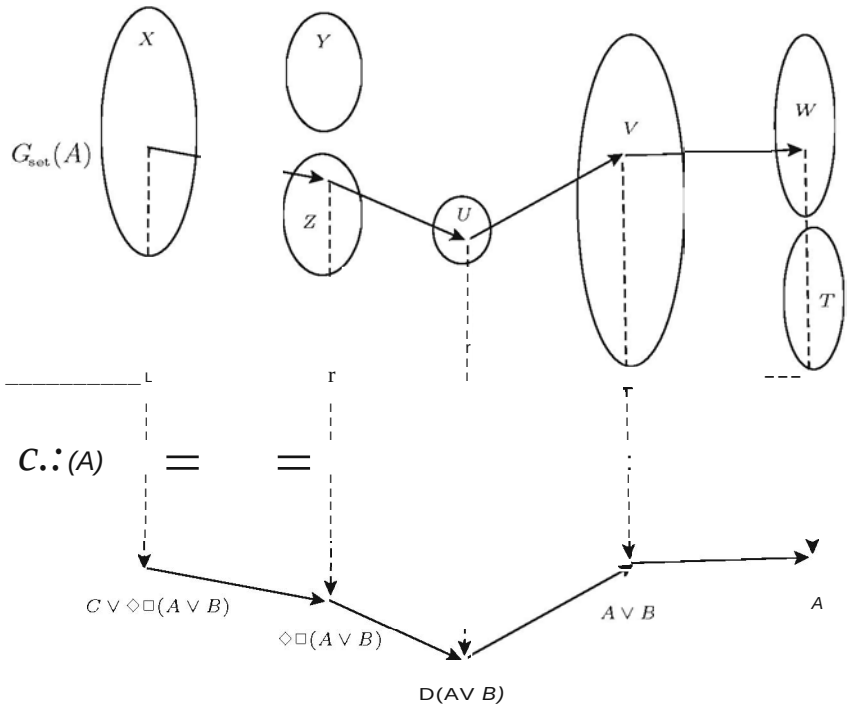


FIGURE 6. Power strategy.

Case 3: $P = (t, A \vee B, \text{I})$. Ey assumption, $T(P) = (X, A \vee B, \text{I})$. Ey (S7)', the game $\text{Gset}(A_a)$ can continue from either (X, A, I) or (X, B, I). Now the game $G_{sem}(Aa)$ continues from position (t, C, I), where $\text{C} = A$ or $\text{C} = B$, according to the choice of I. In either case we let $T(P') = (X, \text{C}, \text{I})$.

Case 4: $P = (t, OA, II)$. By assumption, $T(P) = (\text{X,OA,II})$. By (S8), the strategy σ gives a set Y of nodes accessible from nodes in X and the game $\text{Gset}(A_a)$ continues from (Y, A, II). Since $t \in X$, there is an extension u. of tin Y. This is the choice of 11 in $Gsem(A_a)$. Now the game continues from position $P' = (u, A, \text{II})$ and we define $\tau(P') = (Y, A, \text{II})$.

Case 5: $P = (t, OA, I)$. By assumption, $T(P) = (\text{X,OA,I})$. By (S9), the game $\text{Gset}(A_a)$ continues from position (Y, A, I) for the set Y of all nodes accessible from nodes in X. Since $t \in X$, the extension u. of t chosen by I is bound to be in Y. Now the game continues from position $pi = (u, A, I)$ and we let $\tau(P') = (Y, A, I)$.

At the end of the game $Gsem(A_a)$ we have to check that II indeed has won. There are again several cases:

Case 6: P = $(t,p,$ 11). Since $T(P)$ = $(X,p,$ 11) and σ is a winning strategy, pis true at t. So 11 has won $G_{sem}(A_O)$.

Case 7: P = $(t,p,$ I). Since $T(P)$ = $(X,p,$ I) and σ is a winning strategy, $-,p$ is true at t. So 11 has won $G_{sem}(A_O)$.

Case 8: P = $(t,$ =(Pl, ,$Pn, q),$ 11). Player 11 has won $G_{sem}(A_O)$.

Case 9: P = $(t,=(Pl,$,$Pn,q),I)$. Now $T(P)$ = $(X,=(Pl,$ ... ,$Pn,q),I)$. Since σ is a winning strategy, X = O. On the other hand, by assumption, t EX. So this case simply cannot occur.

Now that we know that this strategy is a winning strategy, we have to show that it is a uniform strategy. Suppose therefore that two plays

$$P_0, \ldots, P_m, \text{where } P_i = (t_i, A_i, d_i)$$

$$P'_0, \ldots, P'_{m'}, \text{where } PI = (t'_i, A'_i, d'_i)$$

end in the same formula $Am = A'_{m'}$ which is of the form =(Pl, ... ,$Pn, q)$ and that the nodes $t-,$ and $t'_{m'}$ give $pi, \ldots .P\text{«}$ the same value. Let

$$T(Pi) = (Xi, Ai, di), \text{i} = 1, \quad , \text{m}$$

$$\tau(P'_i) = (X'_i, A'_i, d'_i), \text{i} = 0, \quad , mi$$

be the corresponding positions in $G_{set}(A_O)$. We show now by induction on i that m = mi, $X;$ = X'_i, Ai = A'_i and di = d'_i. The case i = 0 is deal': Aa = A'_0, X_O = $X6$ and do = d'_0 = 11. The inductive proof is trivial, apart from the case Pi = $(ti, A \vee B, d_i)$, di = 11. Byassumption, $T(P_i)$ = $(Xi, A \vee B,$ 11). The strategy σ has given the two sets Y and Z such that $X \subseteq$ Y U $Z,$ and the game $G_{set}(A_O)$ continues from $(Y, A,$ 11) or $(Z, B,$ 11). Since $t \in$ YUZ, we have either $t \in$ Y or $t \in Z$. In the first case 11 lets C = $A,$ U = Y and in the second case C = $B,$ $U = Z$. Now the game $G_{sem}(A_O)$ continues from position P_{i+1} = $(t,$ C, II) and $\tau(P_{i+1})$ = (U,C,II). Respectively, PI = $(t'_i, A \vee B,$ 11) and $\tau(P'_i)$ = $(Xi, A \vee B,$ 11). The strategy σ (which does not depend on the elements ti and t'_i) has given the same two sets Y and $Z,$ as above, and the game $G_{set}(A_O)$ continues aftel' $T(P_i)$ = $\tau(P'_i)$ from $(Y, A,$ 11) or $(Z, B,$ 11), according to whether t E Y or t EZ. So $XI+l$ = X_{i+1}, A'_{i+1} = A_{i+1} and d'_{i+1} = d_{i+1}.

Thus $t-,$ and t'_m are in X_m and give the same value to $pi, \ldots, Pn$. Because σ is a winning strategy of 11, the nodes $t-,$ and t'_m must give the same value also to q. We have demonstrated the uniformity of the strategy. Q.E.D.

7 The main result

Putting Theorems 4.3, 5.1 and 6.1 together, we obtain:

Theorem 7.1. Suppose A is a sentence of the modal dependence language, and X is a set of nodes of a Kripke structure. The foliowing are equivalent:

1. $(X, A, 11) \in T$ (i.e., A is true in the set X).
2. Player 11 has a uniform winning strategy in $C_{sem}(A)$ in the set X.
3. Player 11 has a winning strategy in $C_{set}(A)$ in X.

Corollary 7.2. Suppose A is a sentence ofthe modal dependence language, and s is a node of a Kripke structure. The foliowing are equivalent:

1. ({s},A,II) ET (i.e., A is true in the set $\{s\}$).
2. Player 11 has a uniform winning strategy in $C_{sem}(A)$ at s.
3. Player 11 has a winning strategy in $C_{set}(A)$ in $\{s\}$.

The proved equivalence leads to easy proofs of the logical consequences and equivalences of Example 1.4. Let us consider, as an example

$$[J(A \to B) \wedge oA \Rightarrow DB.$$

Let X be a set of nodes of a Kripke model. Suppose $O(A \to B)$ and oA are true in X. Let XI be the set of nodes accessible from nodes in X. Thus $A \to B$ and A are true in XI. Then by (T6)', $XI = Y \cup Z$ such that $\neg A$ is true in Y and B is true in Z. Ey Lemma 4.2 and (T7), $A \wedge \neg A$ is true in Y. Ey Lemma 4.2, $Y = 0$. So $XI = Z$ and B is true in XI. We have demonstrated that DB is true in X.

The point of Theorem 7.1 is that the first game Cl with positions of the form (s, A, d) is non-determined and of imperfect information. The set game C_2 is determined and of perfect information. In an obvious sense the two games are *equivalent.* So we have been able to replace a non-determined game of imperfect information with a determined game of perfect information. The cost of this operation is that the determined game of perfect information is played on sets rather than elements. So in a sense there is an exponential cost.

8 Further developments

We can define $=(Pl, \ldots, Pn, q)$ in terms of $=(q)$ if we allow exponential growth of the formula size: $=(Pl, \ldots, Pn, q)$ is true in a set X if and only if the foliowing formula is:

$$\begin{array}{l} (\ Pl \wedge "' \wedge \ Pn \wedge =(q)) \vee \\ (\neg p_1 \wedge \cdots \wedge \ Pn \wedge =(q)) \vee \\ \qquad\qquad \} \ 2^n \text{ disjuncts.} \\ (\neg p_1 \wedge \cdots \wedge \neg p_n \wedge =(q)) \end{array}$$

We can define $=(p)$ if we add to our modal dependence language a Boolean disjunction $A \vee_B B$ with the obvious meaning that $A \vee_B B$ is true in a set iff A is true in the set or Bis, (and $\neg(A \vee_B B)$ is true only' if $X = 0$). **In** terms of the game $Gsem(A_0)$ this means that in position $(s, A \vee_B B, 11)$ Player 11 chooses A or B, and in position $(s, A \vee_B B, I)$ Player I wins. A uniform winning strategy of 11 is required to satisfy the extra condition that player 11 has to make the same move *eoeru* time the position $(s, A \vee_B B, 11)$ is encountered, however many times the game is played. With these conventions $=(p)$ is logically equivalent to $p \vee_B \neg p$.

Merlijn Sevenster (2008) has proved a normal form for modal dependence language and used it to show that the modal dependence language has in fact a translation into basic modal language, but again at exponential cost. He also shows that the satisfaction problem of modal dependence language is NEXP complete.

The finite information logic (Parikh and Väänänen, 2005) is based on dependence formulas of the type $=(A_1, \ldots, A_n, x)$, with the meaning that the value of the variabie x is chosen on the basis of the truth values of the formulas Al, $\ldots, A_n$ only. The formulas Al, $\ldots, A_n$ are assumed to be quantifier free first order formulas (in fact they can be Δ_2 formulas). Quantifiers are allowed only in a "guarded" situation such as $\exists x(=(A_1, \ldots, A_n, x) \wedge B)$ and $\forall x(=(A_1, \ldots, A_n, x) \rightarrow B)$. This is equivalent to the existential-universal fragment of first order logic, but at exponential cost in the length of the formula. The point of this logic is that it captures the concept of *social softsoare* in the sense that people in social situations often make decisions on the basis of finite information about the parameters, indeed on the basis of Lhe truth-val ues of sorne predicates, like "has a valid visa", "speaks Dutch," etc.

In full *dependenee logic* (Väänänen, 2007) first order logic is extended by dependence formulas $=(Yl, \ldots, Yn, x)$ with the meaning that the value of x depends only on the values of $Yl, \ldots, Yn$. This logic is equivalent to the existential second order logic, and is thus quite powerful.

If dependence formulas are added to second order logic, again no proper extension results. We may thus conclude that adding dependence to a logic increases the expressive power in the "middle range" of first order logic, but not in the case of the relatively weak modal logic and the relatively strong second order logics.

Acknowledgments

Research partially supported by grant 40734 of the Academy of Finland. I am indebted to Francien Dechesne, Johan van Benthem and Merlijn Sevenster for helpful comments and suggestions concerning this paper.

1 This follows by symmetry from (T6) and (T7) if we add the condition "one of Y and Z is empty".

References

Hintikka, .1. (1996). *The Principles of Mathematics Revisited.* Cambridge University Press.

Hodges, W. (1997). Compositional Semantics for a Language of Imperfect Information. *Logic Journal of the IGPL, 5(4):539-563.*

Parikh, R. & Väänänen, .1. (2005). Finite information logic. *Annals of Pure and Applied Logic, 134(1):83-93.*

Sevenster, M. (2008). Model-Theoretic and Computational Properties of Modal Dependence Logic. To appeal'.

Väänänen, J. (2007). *Dependence Logic: A New Approach to Independence Friendly Logic,* Vol. 70 of *London Mathematical Society Student Texts.* Cambridge University Press, Cambridge.

Declarations of Dependenee

Francien Dechesne

Department of Mathematics and Computer Science
Technische Universiteit Eindhoven
P.O. Box 513
5600 MB Eindhoven, The Netherlands
f.dechesne@tue.nl

> "Science is the knowledge of consequences, and dependence of one fact upon another."
>
> -Thomas Hobbes

1 It all depends, doesn't it?

Dependence is one of these subtie concepts with so many connotations and usages, that any analysis of its meaning is predestined to fall short of some of its aspects. **In** the first paragraph of the book *Dependenee Logic,* Väänänen (2007), states:

> 'Dependence is a common phenomenon, wherever one looks: ecological systems, astronomy, human history, stock markets. With global warming, the dependence of life on earth on the actions of mankind has become a burning issue. But what is the logic of dependence?'

The book promises a systematic study of the concept, and to show that there is a mathematical theory of dependence.

The paper in this volume (Väänänen, 2008) goes even further. **It** presents a logic of 'possible dependence', where the intended sense of 'dependence' is specified a bit further as follows:

> By *dependence* we mean dependence as it occurs in the following contexts: Dependence of
>
> - a move of a player in a game on previous moves
> - an attribute of a database on other attributes
> - an event in history on other events
> - a variabie of an expression on other variables
> - a choice of an agent on choices by other agents.

In this short comment, we will not discuss the technical properties of Dependence Modal Logic. An excellent analysis of the proposed logic is

Krzysztof R. Apt, Robert van Rooij (eds.). *New Perspectives on Games and Interaction.* Texts in Logic and Games 4, Amsterdam University Press 2008, pp. 255-263.

made by Merlijn Sevenster (2008). Sevenster proves that the expressive power of dependence modallogic does not exceed standard modallogic, but that dependence modal logic can express certain properties more succinctly (witnessed by the fact that adding the dependence atoms to the language increases the complexity from PSPACE to NEXP). Sevenster also proposes an attractive alternative notation for the dependence atoms: dep(xl, ... , x_n ; y) instead of $=(Xl, \ldots, X_n, y)$.

Rather than going into technicalities, we will try to discuss more generally what the dependence atoms, which one can see as *declamtions of dependenee* between variables ('attributes') or propositions ('facts'), do and do not express.

Actually, it was the request to comment on this work at the KNAW-workshop *New perspeciioes on Games and Interaction* that made me realize that the notion of dependence as defined in these atoms does not always coincide with common uses of the notion of dependence. Before the KNAW workshop, I wanted to prepare my oral comment to Professor Väänänen's presentation. I was convineed that what I was going to say, should *depend* on what he would talk about, how could it otherwise count as a comment? So, I prepared my comment only aftel' carefully reading the material sent to me, and left openings for things that would come up during the actual talk.

But in the break before the talk, a more experienced speaker confided me that "if you are asked to comment on somebody's work, just talk about your own work." At first, I was confused by this advice, because it confticted with the dependence I sensed in the concept of 'comment'. But then I realized that this was in Iact LoLally consistent wiLh Lhe mat.hernat.ical Lheory of dependence presented by Väänänen: I could just have prepared a fixed talk about my own work, even without reading the paper. According to Dependence Logic my comment would *depend only* on Jouko's talk: the comment would not depend on anything (would be constant), therefore it would depend on anything!

Fact 1. For any propositions $p, q, qi, \ldots, qn$: if $=(p)$ is satisfied in a set of nodes, then so is $=(q,p)$, and more gene rally: $=(ql, \ldots, qn,p)$.

The conclusion could be that the dependence implicit in the notion of 'to comment', is not just a functional dependence, but maybe a stronger sense of dependence. If you want to express that a different talk should amount to a different comment, one should add an injectivity condition on the function that establishes the dependence: With the usual dependence atoms (on variables) we have: $=(Xl, \ldots, X_n, y)$ is true for a set of valuations (a team) if $y = f(Xl, \ldots, X_n)$ for some f. But it is not necessarily the case that $(Xl, \ldots, Xn) \neq (x'_1, \ldots, x'_n)$ implies $f(Xl, \ldots, xn) \neq f(xi, \cdots, x'_n)$.

Adding injectivity condition would express a stronger property: that the talk *deiermines* the comment (which may be too strong actually for this application).

This discussion shows that the "common phenomenon" that dependence is (quoting Väänänen), is not so easy to capture mathematically. **In** this comment we explore the kind of dependence expressed in the dependence atoms as proposed for Dependence (Modal) Logic a bit further.

2 What is dependence?

In order to clarify a bit what it could possibly mean, we simply retrieve the 'common' meaning(s) of the word *dependenee* from Merriam-Webster's dictionary:

> dependence
> 1: the quality or state of being dependent; *especially:* the quality or state of being influenced or determined by or subject to another 2: reliance, trust 3: one that is relied on 4 a: drug addiction (developed a dependence on painkillers) b: habituation

If one proposes a mathematical theory of dependence, it is good to specify which sense of dependence one intends to formalize. **It** is deal' from the contexts mentioned in the quote above from (Väänänen, 2008), that for example *ontological dependence,* as in "a child is dependent on his parents" (sense 2 and 3) falls outside the scope of Dependence Logic. (For an overview of theories of ontological dependence, see Lowe, 2005.) Obviously, also dependence in the sense of addiction (4) is not within the intended scope of Dependence Logic. The dependence atoms $= (Ql, \ldots, qn, p)$ are **declarations of dependenee** in the sense of 1: one fact is being determined by other facts, or for variables: the value of one variabie is somehow determined by, or at least correlated with the values of other variables. But still one can wonder what we mean exactly by that.

3 A few interesting propositions

In Dependence Modallogic, the dependence atoms will be evaluated in a *set* of nodes. The modal aspect of the logic is that this set of nodes evolves along the accessibility relations in the Kripke model, by taking the modalities as operating on sets of nodes, starting from the set containing the actual node (or nodes). For the evaluation of the dependence atom, it is only the corresponding set of valuations for the propositional variables that matters. **In** fact, the dependence in Dependence Modal Logic is a propositional rather than a modal dependence (this in contrast to the *independenee* in IF-modal logics as discussed in Tulenheimo and Sevenster, 2006).

We discuss some (propositional) facts in Dependence Modal Logic.

3.1 Non-idempotency and bivalence

Let's look at the non-idempotency phenomenon pointed out in section 3 of (Väänänen, 2008): the formula C := O(=(p)V =(p)) is shown to be true in the given model *N (ibid.* Figure 4), while *B* := Op is not.

Note that this does not depend on the chosen model *N,* nor on the O-modality. In fact, for any propositional variabie *p* in any set of nodes (valuations) *W: W* $\models$ (=(p)V =(p)), while *W* $\models$ =(p) only holds if all valuations in *W* agree on *p.* It is easy to see that this fact 'expresses' the two-valuedness of the underlying propositionallogic.

Compare this with (first order) Dependence Logic, where =(x) is true for sets of valuations that give *x* a constant value, (=(x)V =(x)) for sets that give *x* at most 2 values, (=(x)V =(x)V =(x)) at most 3, and so forth. To count the number of elements in a model, we need to put a universal quantifier in front (e.g., $\forall x[=(x)\vee$ =(x)] is true only in models containing at most two elements). This is a difference with the two-valuedness of the propositional part of modal logic: this does not depend on aspecific Kripke model, making the O-modality in a sense irrelevant.

3.2 Does a consequence depend on its cause?

What could be the formal relation between causation and dependence: if one fact *causes* the other, does the one then also *depend* on the ot her? In the absence of a causation-connective, despite the fact that causation and (material) implication are notoriously different, we do a first test by investigating the following implicational formula. Does for each set of nodes *W*

$$W \models (p_0 \to p_1) \to =(p_0, p_1).$$

The answer can be seen to be 'no': take

$$W := \{(p_0; \neg p_1), (\neg p_0; \neg p_1), (\neg p_0; p_1)\}$$

(where we identify the three nodes with their valuations for *Po* and p_1). If we write out the implications, the question boils down to

$$W \overset{?}{\models} (,po \wedge p_1)\vee \ =(pO,Pl).$$

We split *W* at the disjunction. Only the subset *{(po;* 'Pl)} satisfies the first disjunct, so *{(,po;* Pl), *(,po;* $\neg p_1$)} should satisfy the second. But it doesn't, because p: gets a different truth value despite po's truth value being the same.

However, note that both:

$$\models (p_0 \leftrightarrow p_1) \to =(p_0, p_1)$$

$$\models (p_0 \leftrightarrow p_1) \rightarrow =(p_1, p_0)$$

If two atomie propositions are equivalent, then the truth value of the one is a boolean function of the truth value of the other: it is the same. So, equivalent formulas are always mutually dependent!

3.3 Axioms for functional dependenee

One thing a mathematical theory of dependenee could be able to bring, is an axiomatization of the notion of dependence. Armstrong's axioms from database theory, are known to be sound and complete for functional dependence. They are actually not formulated in terms of a *single* attribute that may depend on a *sequence* of attributes, but in terms of *sets* of attributes. If we write $=(X, \mathrm{Y})$ for 'two database records that agree in the attributes in X, also agree in the attributes of Y', they are:

A1 If $\mathrm{Y} \subseteq X$ then $=(X, \mathrm{Y})$ (trivial dependenee or Reflexivity)

A2 If $=(X, \mathrm{Y})$ then $=(X \cup Z, \mathrm{Y} \cup Z)$ (Augmentation)

A3 if $=(X, \mathrm{Y})$ and $=(Y, Z)$ then $=(X, Z)$ (Transitivity)

They are refiected in the following rules for the dependenee atoms in **DL** (cf. also Väänänen, 2008, Example *IA, 10-12):*

1. $=(x, x)$ (trivial dependenee or refiexivity)
2. If $=(x, z)$ then $=(x, y, z)$ (augmentation)
3. $if=(x,y)$ and $=(y,z)$ then $=(x,z)$ (transitivity)
4. If $=(x, y, z)$ then $=(y, x, z)$ (dependenee is order irrelevant)
5. If $=(x, x, y)$ then $=(x, y)$ (dependence is resource insensitive)

Of course, these are axiom *schemes* in the sense that they should be generalized to arbitrary numbers of variables (e.g., also $=(x, y, t, z) \Rightarrow =(x, t, y, z)$). Rules 4 and 5 now appeal' because of the switch from sets to sequences.

These axioms are really for *functional* dependence, and their soundness can be easily checked by writing them out in terms of the existence of functions. However, if we informally read the dependenee atoms as *'depends (only) on'*, not all of them sound completely natura!. For example, is it true that every attribute *depends only on* itself? And also, the augmentation rule 2 amounts to the paradox of Fact 1: a constant (an attribute for which $=(x)$) *depends* on nothing else, but at the same time it depends on anything else ($=$ (Yl, ... , Yn, x)). These considerations make us aware that Dependenee Logic in fact remains a mathematical theory of the already quite mathematical functional sense of dependence.

4 Dependenee versus independenee

We make some remarks on the relationship between dependenee and independence, which is the central concept in Independenee Friendly Logic (Hintikka, 1996). Instead of declaring dependenee in an atomie proposition, the independenee in IF-Iogic is declared at the quantifier: one lists the attributes on which some attribute is not supposed to depend, leaving the attributes on which it may depend to be determined by the context. This context is formed by both the ot her quantifiers within the formula, but also by the domain of the set of valuations (the 'team') for which the formula is evaluated. A detailed study of the effects of this latter part of the context can be found in (Caicedo et al., 2007), where it is shown how equivalence for open formulas can only be soundly defined by fixing a context of variables. We note that the translation procedure from IF-Iogic to Dependenee Logic for sentences, given in (Väänänen, 2007, p. 46), does not work for open formulas for this reason, as the following example shows.

Example 2. Consider the IF-formula $\varphi = \exists y_{/x}[x = y]$. The translation into **DL** would be $\varphi^* := \exists y[=(y) \wedge x = y]$. This formula is only true with respect to sets of valuations in which the value assigned to x is constant (then we can extend every valuation with the same constant assignment to y, thereby both satisfying $=(y)$ and $x = y$). However, this is different for the original IF-formula φ. Consider the set of valuations V consisting of $v_1 = (x \mapsto 0, z \mapsto 0)$ and $v_2 = (x \mapsto 1, z \mapsto 1)$. This is a set in which x is assigned two different values, hence it does not satisfy φ^*. **It** *does* however satisfy the IF-formula φ, because we can let the value we assign to y depend on (by making it equal to) the value of z.

To go from saying *'y must be independent of x'* to *'y may depend on the variables ot her than x'*, one needs to be able to determine the set X of all variables that y could possibly depend on. For open formulas, this set consists not only of variables occurring in the formula, but also on variables in the domain of the valuations that may not occur in the formula itself (like z in the example).

In con neetion to the example, it is interesting to note that ru Ie 3 from the previous section formalizes the issue that makes the validity of Hodges' (1997) formula $\forall x \exists z \exists y_{/x}[x = y]$ from IF-Iogic counterintuitive. On the one hand, y is declared *independent* only from x, not excluding that it may depend on z. But z may depend on x, and then for y to depend on z implies to depend on x by transitivity. This apparently violates the declaration of independenee from x. Dependenee Logic makes this nicely explicit:

$$\forall x \exists z \exists y[=(x, z) \wedge =(z, y) \wedge x = y].$$

Indeed, it follows from the transitivity rule that y also depends on x in this formula. Note that adding an extra conjunction to this translation, trying

to enforce the independenee requirement explicitly, does make the formula false, by contradiction:

$$\forall x \exists z \exists y [=(x, z) \wedge =(z, y) \wedge (\neg =(x, y)) \wedge x = y].$$

But not always do we get such clean-cut contradiction from a negated dependenee atom. As noted in section 5.2 of (Sevenster, 2008), it is not really the case that negations of dependenee atoms declare independence. **In** fact, a negated dependenee atom is only true in the empty set of nodes or valuations, hence will not in itself express independenee in non-empty sets. **In** order to express independence, one needs to jump to the meta-level of second order logic (to express that there is no function witnessing the dependence).

Dependenee and independenee are duals, but not really opposites in the contradictory sense. Therefore, it is not straightforward to make the switch from declaring dependenee to declaring independence.

5 Conclusion

There are several reasons why the topic of Dependenee Logic is at home within the theme *New perspectioes on Games and Interaction.* There is a *new perspeciuoe* in the focus on *dependenee* rather than independence, as in IF-logic.

Like IF-logic, the mathematical theory of dependenee comes with a game theoretical semantics. Where IF-logic enlarged the field of logical games with games of imperfect information, Dependenee Logic adds a uniformity condition on the winning strategies. This is a less standard generalization, in the sense that it is not part of standerd game terminology. The correspondence with database theory, and correlations between attributes is more convincing in our taste. With respect to the **interaction** part: dependenee can be seen as a form of interaction between facts. Note that the concept of (in)dependence does not arise in isolation (cf. Hodges, 2007).

The main conclusion aftel' considering several Dependenee modal formulas, is that the notion of dependenee expressed in the dependenee atoms, is strictly *functional* dependence. **It** allows to talk about functional dependenee while keeping the function implicit (compare, for example $=(y, x)$ with $x = g(y)$ and $=(q,p)$ with $p \leftrightarrow (\neg q)$).

One can wonder what kind of dependenee functional dependenee actually captures. **It** is a bit counterintuitive for a notion of dependenee to have things that are both dependent and independent on anything (viz. constants). **It** seems that functional dependenee expresses some kind of 'correlation', which we sense to be a weaker notion than dependence.

But in the end, we think the historical (technical) evolution of logic provides the clearest view on the motivations for studying dependenee and

independenee between variables and propositions. Dependenee of one variable upon another is already a natural feature in logical languages, by the structure (nesting of quantifiers). The original spark to study dependenee and independenee was the generalizations of the specific syntactic pattern built in in first order logic ('Frege's fallacy' Hintikka, 1996), and to see how they would behave and extend the expressive power of the language. A great motivational insight was Hintikka's idea to link the dependenee of variables to availability of information in semantic games, and thereby independenee to imperfect information. But as many examples have shown (Janssen, 2002; Hodges, 1997), interpretation of the IF-formulas naturally in terrns of what one generally understands as 'independence', is not so straightforward.

Syntactic subtieties in the end turn out to be important spoilers for smooth and elegant results. This is shown in (Caicedo et al., 2007) by the amount of work we need there in order to restore the Prenex Form Theorem for IF-Iogic. But using this syntactic result, we are able to make a syntactic gain for first order logic: by translating a first order sentence into it's IF-prenex form, and then skolemizing it, we avoid unneccesary arguments in the Skolem functions. With these final remarks, I have managed to follow the advice on commenting: to speak about my own work.

References

Caicedo, X., Dechesne, F. & Janssen, T. (2007). Equivalence and Quantifier Rules in Logic with Imperfect Information. Technical Report *ILLG Publications* PP-2007-20, University of Amsterdam.

Hintikka, .1. (1996). *The Principles of Mathematics Revisited.* Cambridge University Press.

Hodges, W. (1997). Compositional Semantics for a Language of Imperfect Information. *Logic Joumal of the IGPL, 5(4):539-563.*

Hodges, W. (2007). Logies of imperfect information: Why sets of assignments? **In** van Benthem, .1., Gabbay, D. & Löwe, R, eds., *Intemctive Logic: Selected Papers [rom the 'lth Augustus de Morgan Workshop, London,* Vol. 1 of *Texts in Logic and Games,* pp. 117-133. Amsterdam University Press.

Janssen, T.M.V. (2002). Independent Choices and the Interpretation of IF-Logic. *Joumal of Logic, Language and Information, 11(3):367-387.*

Lowe, KJ. (2005). Ontological dependence. **In** Zalta, KN., ed., *The Stanford Encyclopedia of Philosophy.* Summer 2005 edition. http://plato.stanford.edu/archives/sum2005/entries/dependence-ontological/.

Sevenster, M. (2008). Model-Theoretic and Computational Properties of Modal Dependenee Logic. To appeal'.

Tulenheimo, T. & Sevenster, M. (2006). On Modal Logic, **IF** Logic, and **IF** Modal Logic. **In** Governatori, *G.,* Hodkinson, I. & Venema, Y., eds., *Advances in Modal Logic, Vol.* 6, pp. 481-501. College Publications.

Väänänen, J. (2007). *Dependence Logic: A New Approach. to Independence Friendlsj Logic,* Vol. 70 of *London Mathematical Society Student Texts.* Cambridge University Press, Cambridge.

Väänänen, .1. (2008). Modal dependenee logic. This volume.

Backward Induction and Common Strong Belief of Rationality

Itai Arieli

Center for the Study of Rationality, Department of Mathematics
Hebrew University of Jerusalem
Jerusalem 91904, Israel
iarie1i@rna.huji.ac.i1

Abstract

We provide a syntactie framework for analyzing extensive-form games. Specifically, we correlate to every such game a "language" and axiomatization. The language and the axiomatization, are presented as a class of sound and complete modeIs, through which we explore the epistemic conditions for the reduction process introduced by Pearce (1984) and Battigalli (1997).

1 Introduction

Economists and game theorists use concepts of knowiedge, belief, and rationality to characterize solution concepts. **In** this work we present an epistemic characterization of the red uction process introduced by Pearce (1984). To do this, we use a syntactic environment in a way that correlates a language to every extensive-form game.

A natural starting point is Aumann's work on backward induction (Aumann, 1995). This work uses the usual knowledge partition setup for anaIyzing interactive rationality in generic perfect information (PI) games.

1.1 Aumann's model

Let Γ be a **PI** game and let Ω be a set of states of the world such that every player $i \in I$ is equipped with an equivalence relation that forms a partition of Ω. Knowledge is derived from this partition in the usual way. Aumann correlates a strategy to every player i in each state $w \in \Omega$ $s_i(w)$ such that every player knows his own strategy in w.

Civen' $h \ll H_i$, player i is defined to be h-rational in state $w \in \Omega$ if it is *noi* the case that he knows he has a better strategy in the sub-game that starts in h than $S_i(W)$. Player i is rational if he is h-rational at each and

1 H_i is the set of non-terminal histories in which player i is required to perform an action.

Krzysztof R. Apt, Robert van Rooij (eds.). *New Perspectives on Games and Interaction.* Texts in Logic and Games 4, Amsterdam University Press 2008, pp. 265-281.

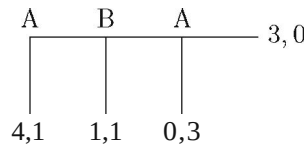


FIGURE 1.

every history $h \in Hi$. Rationality is obtained if all players are rational. The main result of Aumann's work states the following:

Theorem 1.1. In generic **PI** games, common knowledge of rationality is consistent with and entails the backward induction strategy for every player.

That is, if in a state of the world $w \in \Omega$ all players are rational, all players know that they are all rational, and so on ad infinitum, then all players play according to their backward induction strategy.

Aumann's seminal paper provided a formal logical characterization for the solution concept of backward induction and also gave rise to a fruitful debate on the nature of this solution concept in **PI** games.

The main critiques ofthe paper, as weil as ofbackward induction analysis in general, can be explained using the following example: **In** the game depicted in Figure 1, Ann's dominating action is to exit on her first move. **If** for some reason Ann stays, then she should be interpreted by Bob as irrational. The event "Bob is called to play" contradiets Ann's rationality and, in particular, there is no reason to assume that he will stick to his backward induction strategy. That is, the demand for common knowledge of rationality in *eoeru* subgame seerns awkward since it does not take into consideration the moves that lead to the history.

The problem, as we see it, sterns from the fact that the assumption of interactive rationality in the model ignores past history. A player does not update his knowledge (or beliefs) in the course of the game with respect to a possible deviation. The reason for that is that the model does not take into account the interactive nature of the game and does not allow for counterfactual reasoning.

To circumvent this inconsistency we are going to weaken some restrictive assumptions in Aumann's model. First, the language to be constructed includes a belief with probability one operator (rather than knowiedge) for every player $i \in I$ in every one of his histories $h \in Hs$, and the belief revision is carried out using Bayes rule.

Second, we replace the interactive knowledge operator by a *sironq belief* operator. Player i strongly believes in formula 9 if he believes in 9 at each

and every one of his histories that are *logically consistent* with *g.* Using this forward-induetion reasoning eliminates the consistency problem.

This work has some similarities with (Battigalli and Siniscalchi, 2002). Battigalli and Siniscalchi assign to every game in extensive form a mathematical object called a *belief-complete type space.* This object includes *all* possible Bayesian updating belief hierarchies that satisfy some coherence conditions. As they show in their paper, the use of belief-complete type space is crucial in conducting a comprehensive analysis independently of a specific belief type space.

By using a belief with probability one operator rather than probability measures we avoid the complication involved in constructing a belief-complete type space (see Battigalli and Siniscalchi, 1999). Moreover, the use of syntax and logical consistency replace the completeness demand in the type space construction and makes the analysis much more transparent.

2 Framework

Start with a generic PI game/ G. The game consists of: player i's histories *Hi,* i.e, histories in which player i is active and the empty history, for each player i; *terminal histories Z;* and a payoff to each player at each terminal history. A *stTategy* for i is a function that assigns to each of i's histories $h \in H;$ an action at *h.* Each strategy s_i of i determines a set $H(S_i)$ of histories of i, namely, those that s_i *allows* (does not preclude by an action at a previous history). A plan of i is the restriction' to $H(S_i)$ of a strategy s_i.

We now construct a formal language whose building blocks are the following:

Atomie sentences.

These have the form "player i uses plan *Pi,"* denoted simply *Pi.*

Left parentheses and right parentheses.

Connectives and operators of the propositional calculus.

As is known, it is sufheient to take just "or" (V) and "not" $(\neg)$ as primitives and in terms of them to define "and" $(1\backslash)$ and "implies" $(\rightarrow)$.

Belief operator.

For each player i and history $h \in Hs,$ there is a conditional probability one belief operator, denoted *bh.* Informally, $bh\,9$ means that if player i were to observe history *h,* he would ascribe probability one to *g.* Players are not permitted to condition on their own actions. We will return to this demand when we represent the axiomatization.

---

2 For simplicity we introduce the framewerk for PI games, but in fact our results hold for a more general case to be discussed in Section 7.

3 Plans are sometimes called "strategies". Here we do not want a strategy to be defined at the histories that it excludes.

A formula is a finite string obtained by applying the following two rules in some order finitely often:

- Every atomie sentenee is a formula.
- If f and gare formulas, so are (J) V (g), $,(J)$ and $bh(J)$ for every non-terminal history h.

The set of all formulas, ealled *syntax,* is denoted by X (for the game under eonsideration).

Let hand hl be histories, where $h \succ hl$ means that h follows hl in the game tree (or that hl is a prefix of h). Let a be an action at history $h \ll H$; and "i plays a" (or a for short) is a formula, namely VPi, where the disjunction ranges over all plans of i that either preclude h or eall for him to play a at h. Also, "h is reaehed" (or simply h) is a formula, namely Ad, where the conjunction ranges over all players j with histories on the path to h, and at those histories, over all those actions d leading to h. If L is a set of histories, then "L is reaehed" (or simply L) is the formula Vh, where the disjunction ranges over all hEL.

Let $h \ll H$, be an h-plan of i that allows h, and denote by $Pi(h)$ the set of i's h plans. An *opposition h-plan* is a conjunction of plans that allow h, one for eaeh player other than i. An h-plan Pi together with an *opposition* h-plan $P\text{-}i$ determine a terminal history of the game tree z, where $z \succ h$ and a payoff $u_i(Pi, p\text{-}i)$ for i. The set of all *opposition* h-plans is denoted by $P_{-i}(h)$ and the formula eorresponds to "all players other than i play aeeording to h" is:

$$h0 = \bigvee_{p\text{-}iEP\text{-}i(h)} P\text{-}i\cdot$$

2.1 Inference

To make our language meaningful we provide an axiomatization that will deseribe the game at hand. We start by introdueing the axioms and rules. Here $t, 9$ represent formulas and $h, \hat{h}$ histories of the same player:

$$\mathsf{T} \tag{1}$$

$\bigvee Pi$ where the disjunction is over all plans of player i. (2.1)

$'(Pi \wedge qi)$ where Pi and qi are different plans of the same player i. (2.2)

$$bh(J \to g) \to bh\, f \to bh\, 9 \quad (Kj. \tag{3.1}$$

$$bh\, f \to ,bh\, ,f \quad (Dj. \tag{3.2}$$

$$b^h f \to b^{\hat{h}} b^h i, \text{ where } h, \hat{h} \ll Hi. \tag{3.3}$$

$$,bh f \to b^{\hat{h}} \neg b^h f. \tag{3.4}$$

$$Pi \leftrightarrow bh\, Pi \text{ for every } h \ll tt; \tag{3.5}$$

$bh\ ha.$ (3.6)

$(bh\ i \wedge \neg b^h \neg \hat{h}^\circ) \rightarrow b^{\hat{h}}\ i,$ where $h \prec i$: (3.7)

From $i \rightarrow 9$ and i infer 9 *(modus ponens).* (4.1)

From i infer $bh\ i$ *(generalization).* (4.2)

Axiom (1) includes all the tautologies, that is, formulas that assess the value true in every possible truth valuation. Axioms (2.1) and (2.2) correspond to the obvious demand that every player execute exactly one plan. Axioms (3.1) and (3.2) represent classical modal beliefaxioms. Axioms schema (3.3) through (3.5) combine versions of the "truth" and "introspection" axioms. Briefly, they say that players are sure of their own plans and beliefs.

bh is interpreted as the belief of i at history *h* regarding the plans and beliefs of players other than himself. Therefore it makes sense to require that player i in *h* believes that players other than himself played in accordance with *h,* that is, *bh hO.*

Axiom (3.7), which deals with the belief revision policy, states that if i believed i at *h,* and also believed that $\hat{h}^\circ$ "could" occur at *h,* then he believes i at t: **It** refiects the fact that players update their beliefs in a Bayesian way.

(4.1) and (4.2) are the two inference rules of the axiom system AX. A proof in **AX** is a finite sequence of formulas each of which is either an axiom or follows from those preceding it through the application of one of the two inference rules. A proof of a formula i is a proof whose last formula is i. i is provable in **AX** if there exists a proof of i written l-AX *f.*

i is inconsistent if its negation is provable; otherwise it is consistent. Formulas $f_1, f_2, \ldots$ are inconsistent if the conjunction of some finite subset of them is inconsistent; otherwise they are consistent. They entail *9* if the conjunction of some finite subset of them entails *g.*

3 The theorem

3.1 Rationality

Instead of defining the rationality of a player in a particular history as in (Aumann, 1995), we define rationality in terms of the plan that a player uses and his beliefs. We replace the utility maximization demand by a weaker one; namely, we require that in every optional active history dictated by the player's plan, it not be the case that he believes that he has a better plan.

Formally, we say that i uses *rationol* plan *Pi* if in every history $h \in H_i(pi)$[4] it is not the case that he believes that he has a better h-plan. A

[4] We denote by *Hi(Pi)* the set of player i's histories that are not precluded by the plan *Pi.*

formula that represents the rationality of a player of course needs to take into account player i's payoff function in the game.

For each pair of different plans of player i narnely, *Pi* and *qi,* we denote by $Q^h_{p_i}(qi)$ the disjunction of opposition h-plans in *P_i(h),* where *qi* yields a higher payoff' than *Pi.* The formula that represents that plan *Pi* is rational for i is the fol!owing:

$$T(Pi) = Pi \rightarrow \bigwedge_{\{hlhEH(Pi)\}} \bigwedge_{\{q_i \in P_i(h) | q_i \neq p_i\}} bh\, Q^h_{p_i}(q_i).$$

Define player i to be rational if he uses a rational plan. That is,

$$r_i = \bigwedge_{Pi} r(p_i).$$

In brief, a player is rational if it is not the case that he believes that he has a better plan in one of the histories that is consistent with his plan. This is in fact a much weaker demand than in (Aumann, 1995), where a player is required not to know that he has a better strategy in *eoeru* one of his histories.

Remark. We do not claim that the above formal definition of "rationality" is the only possible one. We do however claim that any reasonable definition entails the fol!owing: if i is "rational" in any commonly accepted sense (such as utility maximization), then certainly *Ti* obtains. The formula corresponding to al! players being rational is

$$r = \bigwedge_{Ti}.$$

3.2 Strong belief

The st rong belief operator is a substitutional concept for interactive common knowiedge. The operator is external to the language and it is defined in such a way as to resolve the consistency problem discussed in the Introduction.

We say that i *sironqls; believes* a formula *gif,* for each of i's histories *h « U.,* either

(i) i believes *9* at *h,* or

(ii) *9* precludes *h* being reached, or, equivalently, *9* is inconsistent with *h.*

5 I.e., $Q^h_{p_i}(q_i) = V\{P\text{-}i \in P\_i(h) \mid Ui(qi,P\text{-}i) > Ui(Pi,P\text{-}i)\};$ if there are na such *P-i* we set $Q^h_{p_i}(qi) = \bot$.

In words, i continues to believe 9 no matter what happens, unless he reaches a history that is logically impossible under *g.* We can write it as a formula in the following way:"

$$Sbi(g) = \bigwedge_{\{h \in H_i \mid \not\vdash_{Ax} \neg(h \wedge g)\}} bh(g).$$

We say that 9 is *stmngly believed* (or that there is *sironq belief* of *g,* written *sb(g))* if each player strongly believes *g. Mutual stronq belief* of 9 of order *n* (written *sbn(g))* is defined as $sb(sb^{n-1}(g))$; that is, each iteration provides for strong belief of the foregoing iteration (note that the strong belief operator does not commute with conjunction). *Common sironq belief* of 9 comprises all the formulas $sb^n(g)$ for all *n.*

The main result of this paper states the following:

Main Theorem. Common strong belief of rationality is consistent and entails the backward induction outcome in every **PI** game.

4 Soundness and completeness

Before proving the main theorem, we would like to present a class of simple models for our axiomatization that links the syntax to the semantics. The most preferabie way would be to link the syntax to a class of models that characterize it by soundness and completeness relation. The way to do that would be by looking at the canonical model of the language with respect to the logic that our axiomatization defines.

We would first like to introduce some more terminology:
An axiom system Ax is said to be *sound* for a language $\mathfrak{L}$ with respect to a class of models C if every provable formula f is valid with respect to C. An axiom system Ax is said to be *complete* for a language $\mathfrak{L}$ with respect to a class of models C if every valid formula f with respect to C is provable in Ax.

4.1 The canonical model

Definition 4.1. A set of formulas Γ is *maximal consistent* with respect to Ax if it satisfies the following two conditions:

a. Γ is consistent with respect to Ax.

b. Γ is maximal with respect to that property.

It can be seen that maximal sets do exist[7] and satisfy the following properties:

6 If there are no $h \in H_i$ that are consistent with g, put sbi(g) = $\perp$.
7 See (CheIlas, 1984).

1. **ris** closed under *modus ponens (4.1).*
2. **r** contains all the theorems of**Ax.**
3. For every formula φ, φ E r or $\neg\varphi$ E **r.**
4. For every formula φ, ψ, $\varphi \vee \psi$ E r *ij}"* φ E r or ψ E **r.**
5. For every formula φ, ψ, $\varphi \wedge \psi$ E r *ij}"* φ E **rand** ψ E **r.**
6. Every consistent set of formulas can be extended to a maximal consistent set.

Now let Ω be the set of all maximal consistent sets; we call the elements of Ω *states of the uiorld.* For each r E Ω and non-terminal history *h* « *Hi,* we define ***rl*** *h* to be the set of all formulas that player i *h-believes* in **r.** That is,

$$\Gamma/h = \{\varphi \mid \mathrm{b}^h \varphi \in \Gamma\}.$$

For every player i and a non-terminal history *h* E *Hi,* define the usual accessibility binary relation *Rh* over Ω as follows: let **r,** Λ E Ω, r *RhA* iff ***rl*** *h* $\subseteq$ A. Let B^h_Γ be the set of all states of the world that player i considers possible at *h* E *Hi,* that is,

$$B^h_\Gamma = \{\Lambda \in \Omega \mid \Gamma R_h \Lambda\}.$$

Observation 4.2. ***rl*** *h* satisfies the following conditions:

1. ***rl*** *h* is consistent (therefore $B^h_\Gamma \neq 0$).
2. ***rlh*** is closed under (4.1) and (4.2) (if φ **Er,** then *bh* φ E Γ).
3. φ E ***rl*** *h* for every φ such that I-AX φ.

Proo]. Part 2 follows from positive introspection, while part 3 is straightforward from generalization. For part 1, assume by way of contradiction that ***rl*** *h* is not consistent. Then we have φ_1, φ_k E ***rl*** *h* such that **AX** I- $\neg(\varphi_1 \wedge \ldots \wedge \varphi_k)$. By definition, *bh* φ_1, *bh* φ_k E **rand** so from *K* we get $\mathrm{b}^h(\varphi_1 \wedge \ldots \wedge \varphi_k)$ E r but from part 3 *bh* $\neg(\varphi_1 \wedge \ldots \wedge \varphi_k)$ **Er,** a contradiction to *D.* Q.E.D.

Let r E Ω; we would now like inductively to define a truth assessment over the set Ω, where $\|\varphi\|$ is the set of states in which φ holds:

- for atomie sentences, r $\models$ *Pi ij}"* *Pi* E I';
- for formula φ, r $\models \neg\varphi$ *ij}"* r $\not\models \varphi$;
- for formulas φ and ψ, r $\models \varphi \vee \psi$ *ij}"* r $\models \varphi$ or r $\models \psi$;

- for formula $\varphi = \mathrm{b}^h \psi$, $h \in H_i$, $\Gamma \models \mathrm{b}^h \varphi$ *iff* $B_\Gamma^h \subseteq \|\varphi\|$.

Proposition 4.3. For every r EΩ and every formula φ, r $\models \varphi$ *ijl'* φ Er.

Proo]. We wili prove the proposition using induction on the depth of the formula.

For formulas of depth zero the proof is immediate from the properties of maximal consistent sets and the truth assessment policy. We prove the proposition first for formulas of the form $\varphi = \mathrm{b}^h \psi$, where ψ is from depth *n-1* $\geqslant 0$. The general case foliows from the properties of maximal consistent sets and the truth assessment policy.

$\Leftarrow$: If φ Er then by definition of r/*h*, ψ Er/*h;* therefore ψ EΛ for every Λ EB_Γ^h by the induction hypothesis $B_\Gamma^h <$ $\|\psi\|$; therefore r]- φ.

$\Rightarrow$: If r $\models \varphi$ then $B_\Gamma^h \subseteq \|\psi\|$ so ψ EΛ for every Λ such that *r/h* $\subseteq$ r; therefore r/*h* I-AX ψ for otherwise we could have constructed a maximal consistent set Λ′ such that r/*hu* {¬ψ} < Λ′. But because r/*h* contains ali the theorems ofAX and is closed under 4.1 and 4.2 we get that ψ E r/*h* and therefore φ E r. Q.E.D.

Thus Proposition 4.3 leads to the foliowing theorem:

Theorem 4.4. Ω is a sound and complete model with respect to AX.

We would like to state a few properties of the set B_Γ^h.

Lemma 4.5. For every r E Ω player i, and history *h* E *Hi,* B_Γ^h satisfies Lhe foliowing properties:

1. $B_\Gamma^h \neq$ o.
2. Let *h* Ett, such that $\hat{h} \succ h$ and $B_\Gamma^h \cap \|\hat{h}^\circ\| \neq$ 0; then $B_\Gamma^{\hat{h}} \subseteq B_\Gamma^h \cap \|\hat{h}^\circ\|$.

Proo]. Part 1 is a consequence of part 1 in the previous observation. For part 2, assume B_Γ^h n $\|\hat{h}^\circ\| \neq$ 0; then from the truth assessment policy we get r $\models$ -ib^h ,*ha,* but by Lemma (4.1) that means that -i b^h ,*ha* Er. If for some formula *j, bhj* E r, then, because r is a maximal consistent set, *bhj* Λ ,*bh* ¬$\hat{h}^\circ$ Er; therefore from (3.6) and (4.2) $\mathrm{b}^{\hat{h}}$ *t c t:*

We show the foliowing: if *bhj* Er then $\mathrm{b}^{\hat{h}}$ *j* Er. Therefore r/*h* < r/$\hat{h}$ and in particular $B_\Gamma^{\hat{h}} \subseteq B_\Gamma^h$. This foliows from the fact that if r/*h* $\subseteq \Lambda$ then obviously $\Gamma/\hat{h} \subseteq$ A. From $B_\Gamma^{\hat{h}} \subseteq \|\hat{h}^\circ\|$ (3.4) we get the desired result. Q.E.D.

We would like to consider a class of models with the following properties for our game G. Let Ω be a set of states of the world. Each player i is equipped with an equivalence relation $\sim_i$ over Ω and a plan *Pi = Pi(W),* for

each wE Ω. For every equivalence class *IIi(w)* and history *h* E *Hi*, consider a nonempty set $B^h_{\Pi_i(\omega)}$ C Ω. Let **llpi ll** be the event in which player i executes the plan *Pi*. We would like the fol!owing properties to hold in the model:

1. *If* $W \sim_i io'$ *then Pi(W)* = *Pi(W')*.
2. $B^h_{\Pi_i(\omega)}$ < *IIi(w)* n **llh**[8]
3. For $\hat{h}, h$ E *tt*, such that $\hat{h} \succ h$, if $B^h_{\Pi_i(\omega)}$ n **llholl** $\neq$ 0 then $B^{\hat{h}}_{\Pi_i(\omega)}$ C $B^h_{\Pi_i(\omega)}$ n **llholi**.

We can think of the equivalence relation $\sim_i$ over Ω as arelation that defines knowiedge. That is, *w* $\sim_i$ *wI* if player i cannot distinguish between these two states. The first requirement represents the demand that each player knows his own strategy.

The second requirement entails that each player know his beliefs and that if the game gets to history *h* E *H;* player i wil! assign probability one to the set in which players other than himself played in accordance with *h*. The third requirement relates to the belief revision policy.

We cal! the above class of models for the game C, *M(C)*. The truth assessment in those models is as in the canonical case. We are now in a position to state the main result of this chapter:

Theorem 4.6. The class of modeIs *M(C)* is sound and complete with respect to Ax.

Proo]. The soundness part is omitted. Completeness is a straightforward consequence of Lemma 3.1 and Proposition 1. **It** suffices to note that the canonical model satisfies al! the above requirements. The only thing left to define is the equivalence relation for every player i. Let r, **l"** E Ω. We define r $\sim_i$ **l"** if, for some history *h* E *H;* r / *h* = *r' / h*, we have to show that B^h_Γ depends only in the equivalence class of r. But it turns out to be pretty clear, assume that r / *h* = *r' / h* for some *h* E *H;* and let *hl* E *Hi*. **If** *f* E r / *hl* then *bh' f* E **rand** from the positive introspeetion property (3.3) $b^h b^{hl} f$ E r it fol!ows that $b^h b^{hl} f$ E **l"**. Therefore $b^{hl} f$ E **l"** and *f* E *rI/hI*, and vice versa. Therefore r / *hl* = *rI/hI* for every history *hl* E *Hs*, and so $B^{h'}_\Gamma = B^{h'}_\Gamma$ for every history *hl* E *Hi*. Q.E.D.

The set $B^h_{\Pi_i(\omega)}$ is the set of states that player i considers possible in *h*, or, equivalently, the support of player i's measure in *h*. From property 3 we can see that the belief update is not strictly Bayesian; that is, we have

8 1111,0 ll is the event in Ω where players other than i play in accordance with history *h*.

$B^{\hat{h}}_{\Pi_i(\omega)} < B^{h}_{\Pi_i(\omega)}$ n IIholl rather than equality. If we want strict Bayesian updating we must add the following axiom:

$$\mathrm{b}^{\hat{h}} f \rightarrow \mathrm{b}^{h}(f \vee \neg \hat{h}^{\circ}) \text{ where } h \prec \hat{h}. \tag{3.8}$$

Denote by Ax! the axiom system Ax with the addition of (3.8) and by *M+(G)* the class of models with equality in property 3. We then get the foliowing theorem:

Theorem 4.7. *M+(G)* is sound and complete with respect to Ax".

Proo]. It suffices to note that in the canonical model axiom (3.8) entails that $B^{\hat{h}}_{\Gamma} \supseteq B^{h}_{\Gamma}$ n IIholl for every r E Ω. Assume by way of contradiction that $B^{\hat{h}}_{\Gamma} \not\supseteq B^{h}_{\Gamma}$ n $\|\hat{h}^{\circ}$ 11; then we have Λ E Ω such that Λ E B^{h}_{Γ} n $\|\hat{h}^{\circ}\|$ but $\Lambda \notin B^{\hat{h}}_{\Gamma}$. By definition of $B^{\hat{h}}_{\Gamma}$, $\Gamma/\hat{h} \not\subseteq \Lambda$ and so there exists f such that $\mathrm{b}^{\hat{h}} f$ E rand f ∉ Λ and so from maximality of Λ , f E A. Then, from (3.8), b^{h}(J V ,*ĥO)* E rand so f V ,*ĥo* E Λ but *h0* E A, a contradiction to the consistency of A. Q.E.D.

5 Proof of the Main Theorem

Consider the foliowing inductively defined series of plans: For every player i, let p_i^O = *Pi*, P^0_{-i} = $\Pi_{j\neq i} P^0_j$ and *pa* = $\Pi_j P^0_i$. Let *n* > 0 and assume P_l^{n-1} to be defined for every player i, and let *Hn-l* be those non-terminal histories that are reachable by profiles of plans from *P'r:',* Now *Pi* E P^n_i if it satisfies the foliowing requirement:

1. *Pi* ■P^n i^1

2. For every history *h* « *H(pi)* n *Hn-l* there exists *P-i* E $P^{n-1}_{-i}(h)$ such that $u_i(p_i, p_{-i}) \geq u_i(q_i, p_{-i})$ for every *qi* E *Pi(h).*

In words, a plan *Pi* E P_l^{n-1} is in P^n_i if and only if, for every non-terminal history *h* that is consistent with *P"* and *Pi,* there exists some opposition h-plan *P-i* E P_l^{n-1} such that *Pi* is a best reply to ali plans *qi* E *Pi(h).*

Lemma 5.1. For every generic PI game there exists m such that, for every player i, $P^n_i = P^m_i$ for every *n* > m. And, every profile *P* E *P'"* leads to the unique backward induction outcome.

Proo]. See (Battigalii, 1996). Q.E.D.

Lemma 5.2. A plan *Pi* of player i is consistent with the set of formulas *{T,* Sb(T), ... ,*Sbn(T)}* iff *Pi* E P^{n+1}_i.

Proo]. We prove the lemma using induetion on *n*.

Case $n = 0$.

$\Rightarrow$: Let $Pi \in P_i^0$ sueh that $Pi \notin P_i^1$. By the definition of P_i^1 we have a history $h \in H;$ sueh that, for every opposition h-plan $P\text{-}i$, there exists $qi \in Pi(h)$ with $u_i(p_i, p_{-i}) < u_i(q_i, p_{-i})$. Let $a = pi(h)$, the action prescribed by Pi in h. We argue that there exists an action $b \neq a$ in h sueh that the maximal payoff in $T(h,a)$ [9] is smaller than the minimal payoff in $T(h,b)'$ Otherwise we eould find an appropriate opposition h-plan $P\text{-}i$ sueh that Pi would be a best reply to $P\text{-}i$. Let $qi \in Pi(h)$ sueh that $qi(h) = b$. Then, for every opposition h-plan $P\text{-}i \in P\text{-}i$ we have $u_i(p_i, p_{-i}) < u_i(q_i, p_{-i})$. Therefore by the definition of $Q^h_{p_i}(q_i)$ (the set of all the opposition h-plans sueh that, for player i, qi yields a higher payoff than Pi), we have $Q^h_{p_i}(qi) = hO$. From the definition of rationality r we have $\vdash_{\mathrm{Ax}} r \wedge Pi \rightarrow T(Pi) \wedge Pi$. But from the definition of $T(Pi)$, $\vdash_{\mathbf{Ax}} (T(Pi) \wedge Pi) \rightarrow \neg \mathrm{b}^h(Q^h_{p_i}(q_i))$, whieh together with axiom (3.6) contradiets (3.2).

$\Leftarrow$: Let $P \in t»$, we have to show that P is consistent with T. For every i, $Pi \in P_i^1$, therefore, by definition, for every history $h \in H(pi)$ we ean find an opposition h-plan, $p^h_{-i} \in P\_i(h)$ sueh that $u_i(p_i, p^h_{-i}) \geq u_i(q_i, p^h_{-i})$ for every h-plan qi. Let $\Omega = \mathrm{TIiEf}\, Pi$, where $\hat{p} \sim_i pi$ iff $\hat{p}_i = p'_i$. Therefore the equivalenee class $IIi(w)$ is determined by player i's plan.

Start with $h \ll H(pi)$ and then put $B^h_{p_i} = (p_i, p^h_{-i})$. Now for the other $\hat{h}$, if $\hat{h} \in H(pi) \cap H(p^h_{-i})$, put $B^{\hat{h}}_{p_i} = (Pi, p^h_{-i})$, or else ehoose $p^{\hat{h}}_{-i}$ and put $B^{\hat{h}}_{p_i} = (Pi, p^{\hat{h}}_{-i})$ and so forth until all the histories $H(pi)$ are covered. For $h \notin H(Pi)$ we ean let $B^h_{p_i}$ be arbitrary. We repeat the procedure for all the players. One ean see that our model belongs to the class $M(G)$.

To show that $P \models T$, it would be suffieient to show that $P \models T(Pi)$ for every player i.[1a] This follows from the faet that for every history $h \in H(pi)$ and $qi \in Pi(h)$, $p^h_{-i} \notin Q^h_{p_i}(q_i)$.

Case $n > 0$.

At this point we have the basis for the induetion. Now assume that the theorem is obtained for every $0 \leq k < n$. We prove it for n as follows.

$\Rightarrow$: Let Pi be a plan of player i sueh that $Pi \notin P_l^{n+l}$. If there exists a $k \leq n - 1$ sueh that $Pi \notin Pik^{+1}$, then by the induetion hypothesis we are done. Assume that $Pi \in P_i^n$. Then for some history $h \in H(pi) \cap H''$ and for every $P\text{-}i \in P^n_{-i}$ $nP\text{-}i(h)$ there exists a strategy $qi \in Pi(h)$ sueh that $u_i(p_i, p_{-i}) < u_i(q_i, P_{-i})$. Denote $(a = pi(h));$ as in the $n = 0$ case there exists an action b in h sueh that the maximal payoff for player i in $T^n_{(h,a)}$ is

[9] The subtree $T_{\hat{h}}$ corresponds to the subgame starting at history $\hat{h}$.

[10] Obviously for $q; \neq Pi$, $P \models r(qi)'$

smaller than the minimal payoff in $T^n_{(h,b)}$.[11] Set $r'' = \wedge_{0\le k\le n}$ sb" r (where sb$^0 r$ = r); by the induction hypothesis the only plans of players other than i that are consistent with r^{n-1} are those in P^n_{-i} and therefore $\vdash_{AX} r^{n-1} \rightarrow$ $V_{p_{-i}\in P^n_{-i}} p\text{-}i$. h « H"; so by the definition of sb"> = sb(sbn-1r) we have:

$$\vdash_{AX} \text{sbnr} \rightarrow \text{bh(sbn-l r}\}$$

Therefore $\vdash_{AX} r'' \rightarrow$ bh[$V_{p_{-i}\in P^n_{-i}} p\text{-}i$] denote by $P^n_i(h)$ the set of plans of player i in P^n_i that are consistent with the history h. By axioms (3.4) and (3.5) we have $\vdash_{AX} r'' \rightarrow$ bh[V $P^n_{-i}(h)$]. Let qi be a strategy of player i that allows hand $qi(h)$ = b. So from both the assumption and the definition we get $P^n_{-i}(h) \subseteq Q^h_{p_i}(q_i)$ and $\vdash_{AX} Pi \wedge r(pi) \rightarrow \neg$ bh[V $Q^h_{p_i}(q_i)$]. Therefore $\vdash_{AX} Pi \wedge r(pi) \rightarrow \neg$ bh[V $P^n_{-i}(h)$]. Thus, Pi is not consistent with r.

$\Leftarrow$: Let $P \in pn+l$. We have to show that P is consistent with [r; ... , sbn(r)}. From the fact that $Pi \in P^{n+1}_i$, as in the n = 0 case, for every history $h \in H$; n H'' there exists an opposition h-plan p^h_{-i} such that the following properties obtain:

1. $p^h_{-i} \in P^n_{-i}$.
2. $u_i(p_i, p^h_{-i}) \ge u_i(q_i, p^h_{-i})$ for every $qi \in Pi(h)$.

Let Ω = Di Pi; inductively we can construct $B^h_{\hat{p}_i}$ such that the following property will be satisfied:

$$for\ k \le n \quad \hat{p} \in P^{k+1} \Rightarrow \hat{p} \models r^k.$$

Now we can change $B^h_{p_i}$, so that $B^h_{p_i}$ = (Pi, p^h_i) and obviously (*) would still be satisfied. By the induction hypothesis, for every player i and $h \in HnnH_i$, for the appropriate p^h_{-i} we get $(Pi, p^h_{-i}) \models r^{n-1}$. From the modification of $B^h_{p_i}$ we have $P \models bh(r^{n-1})$. But again using the induction hypothesis sb(r^{n-1}) = $\wedge_{hEH_n} bh(r^{n-1})$ and therefore $P \models$ sb" rand therefore $P \models r^n$. Q.E.D.

6 Battigalli and Siniscalchi

In this section we present the model of Battigalli and Siniscalchi, and discuss the conneetion between our model and theirs. We start with a few definitions related to their work.

For a given measure space $(X, \mathfrak{X})$ and a nonempty collection of events $\mathcal{B}$ such that $0 \notin \mathcal{B}$:

11 The subtree T''' is the subtree of the original game that is compatible with P':

Definition 6.1. A *conditional probability system* (CPS) on $(X, \mathfrak{X}, \mathcal{B})$ is a mapping $\mu(\cdot \mid \cdot) : \mathfrak{X} \times \mathcal{B} \to [0,1]$ such that, for all B, $\emptyset \neq B \in \mathcal{B}$ and $A \in \mathfrak{X}$, (1) $\mu(B \mid B) = 1$, (2) $\mu(\cdot \mid B)$ is a probability measure on $(X, \mathfrak{X}, B)$, and (3) $A \subset B \subset C$ implies that $\mu(A \mid B)\mu(B \mid C) = \mu(A \mid C)$.

The set of the conditional probability system on $(X, \mathfrak{X}, \mathcal{B})$ can be regarded as a (closed) subset of $[\Delta(\mathfrak{X})]^{\mathcal{B}}$, denoted by $\Delta^{\mathcal{B}}(\mathfrak{X})$.

Definition 6.2 (Ben-Porath). Given a *(PI)* game Γ, a *type space* on $(\mathcal{H}, S(\cdot), I)$ is a tuple $\mathfrak{T} = (\mathcal{H}, S(\cdot), I, (\Omega_i, T_i, g_i)_{i \in I})$ such that for every $i \in I$, T_i is a compact topological space and

1. Ω_i is a closed subset of $S_i \times T_i$ such that $\mathrm{proj}_{S_i} \Omega_i = S_i$.
2. $g_i = (g_{i,h})_{h \in \mathcal{H}} : T_i \to \Delta^{\mathcal{H}}(\Omega_{-i})$ is a continuous mapping.

For any $i \in I$, $g_{i,h}(t_i)$ denotes the beliefs of epistemic type t_i conditional on h.

Definition 6.3. A *belief-complete type space* on $(\mathcal{H}, S(\cdot), !)$ is a type space $\mathfrak{T} = (\mathcal{H}, S(\cdot), I, (\Omega_i, T_i, g_i)_{i \in I})$ such that for every $i \in I$, $\Omega_i = S_i \times T_i$ and the function g_i maps T_i onto $\Delta^{\mathcal{H}}(\Pi_{j \neq i} S_j \times T_j)$.

Battigalli and Siniscalchi (1999) showed that for finite games, a belief-complete type space can always be constructed.

6.1 Semantical conditional belief operators

Let $\mathcal{A}_i$ denote the sigma-algebra of events $E \subset \Omega$ such that $E = \Omega_{-i} \times \mathrm{proj}_{\Omega_i} E$. A_i is similarly defined. The *conditional belief* operator for player i given history $h \in \mathcal{H}$ is a map $B_{i,h} : A_i \to \mathcal{A}_i$ defined as follows:

$$\forall E \in A_i \qquad B_{i,h}(E) = \{(s,t) \in \Omega \mid g_{i,h}(t_i)(\mathrm{proj}_{\Omega_{-i}} E) = 1\}.$$

Like the syntactical operator $\mathrm{b}^h(\cdot)$, the semantical operator $B_{i,h}(E)$ has the meaning of "player i ascribes probability one that his opponents' strategies and epistemic types are consistent with Γ, when he observes history h."

We say that player i *strongly believes* that an event $E \neq \emptyset$ is true if and only if he is certain of E at all histories consistent with E. Formally, for any type space $\mathfrak{T}$, define the operator $\mathrm{SB}_i : A_i \to \mathcal{A}_i$ as follows: $\mathrm{SB}_i(\emptyset) = \emptyset$ and

$$SB_i(E) = \bigcap_{h \in \mathcal{H}:\, E \cap [h] \neq \emptyset} B_{i,h}(E)$$

for all events $E \in A_i \setminus \{\emptyset\}$ and $[h]$ is the event "history h occurs."

Along the lines of their syntactic approach, B&S define what they call "the mutual strong belief" operator. For any Borel subset $E \subseteq \Omega$ such that $E = \text{II}_{i \in I} \text{proj}_{\Omega_i} E$, let

$$SB(E) = \bigcap_{i \in I} \text{SB}_i(\Omega_i \times \text{proj}_{\Omega_i} E).$$

To adjust this model to inclusion in the class $M^+(G)$, we have to specify a plan, an equivalence relation, and a set $B^h_{\Pi_i(\omega)}$ for every player i and history $h \ll H$; in each state of the world w.

Obviously, for $W = (s,t)$, $Pi(W)$ would be the plan induced by the strategy Si. Define $W \sim_i WI$ iff $ti(w) = ti(W^I)$ and $Si(W) = Si(W^I)$. For every $h \in H$; let $B^h_{\Pi_i(\omega)} = (Si(W), ti(w)) \times supp\{g_{i,h}(t_i(\omega))\}$, where $Si(W)$ and $ti(w)$ are the strategy and the type of player i respectively in the state of the world wand $supp\{g_{i,h}(t_i(\omega))\}$ is the support of the measure $gi,h(ti(W))$ in Ω_{-i}.

It remains to show that the sets $B^h_{\Pi_i(\omega)}$ satisfy the requirement stated for the $M^+(G)$ modeis.

Lemma 6.4. Let $i: h \in tt$, such that $h \succ \hat{h}$; if $B^{\hat{h}}_{\Pi_i(\omega)} \cap \text{Ilholl} \neq 0$ then $B^h_{\Pi_i(\omega)} = B^{\hat{h}}_{\Pi_i(\omega)} \cap \text{Ilholl}$.

Proof. $\text{Ilholl} = Si \times S\_i(h) \times \text{Tand}$, in particular, $\text{proj}_{\Omega_{-i}} \|h^\circ\| = S\_i(h) \times T\_i$, is an open set. From the fact that $B^{\hat{h}}_{\Pi_i(\omega)} \cap \text{Ilholl} = (ti(w), Si(W)) \times supp\{g_{i,\hat{h}}(t_i(\omega))\} \cap \text{Ilholl} \neq 0$ we get $gi,h(ti(w))(\text{llhOII} \cap \Omega_{-i}) > 0$ and by property 3 in definition 6.2 we get the result. Q.E.D.

Theorem 6.5. The model obtained from the *complete belief type spa*ce for a game is a canonical model with respect to.Ax" .

Proof. The soundness part follows from the lemma; as for the completeness part, the proof is a bit more complicated and it relies heavily on the completeness of the type space. Q.E.D.

7 Discussion

One more important aspect in extensive-form games yet to be discussed counterfactual reasoning. To test the stability of an equilibrium concept, one must consider questions like "what would be the case if..." That is, to provide an epistemic foundation for a solution concept one must determine whether a player thinks that he could gain by deviating from the strategy prescribed by the equilibrium.

Consider the following version of the centipede game (Figure 2). Backward induction reasoning leads to the profile of strategies in which Ann and Bob exit at each and every one of their decision points. But what if Ann

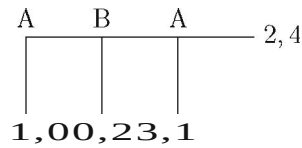


FIGURE 2.

stays on her first move? Bob might interpret this move as irrational, in which case he would expeet Ann to be irrational also at her second decision point, and so he expects to gain by staying. Arm, by the above reasoning, thinks that by staying she might be interpreted as irrational by Bob who, as a rational player, will stay and so when reaching her second decision point she will exit and get 3 instead of 1.

This paradox can be resolved using Battigalli's best rationalization principle (Battigalli, 1996) which states that "players, beliefs conditional upon observing history h are consistent with the highest degree of strategie sophistication of their opponents." That is, when a player observes deviation from an equilibrium path by another player, he assigns to that deviation the maximal degree of belief with rationality.

The main theorem in fact states that the plans that are consistent with common strong belief of rationality are consistent with the principle that we have just described. That is, if *Pi* is a plan of i in a generic **PI** game that is consistent with common strong belief of rationality, then in every history $h \ll H(Pi)$, *Pi* maximizes his payoff function according to the highest rationality degree assumption. Formally, let $h \in H(Pi)'$ Assume that h is consistent with $\{T, \ldots, \mathrm{Sbm(T)}\}$ but not with $\{T, \ldots, \mathrm{Sbm(T)}, \mathrm{Sb}^{m+1(T)}\}$. Then the reduction of the strategy *Pi* to the subgame starting at h maximizes player i's payoff function with respect to the assumption $\{\mathrm{b}^h(r), \ldots, \mathrm{bh(sbm(T))}\}$.

Thus, if players stick to the best rationalization principle, and that is common strong belief, then no player will believe that he can gain by deviating from his plan.

7.1 General extensive games

We reduced the analysis to **PI** extensive-form games but in fact it is equally valid for general finite extensive games with perfect recall. The way to adjust the framework for this case is fairly obvious. Again we use plans rather than strategies, except that now H, is the collection of information sets of i; indeed, in the **PI** case it is identical to the set of histories of player i.

The axiomatization stays the same but here we have a belief with a probability one operator for every player's information set rather than for every history. The rationality definition and the strong belief operator re-

main unchanged. The sound and complete models are the same once we replace histories with information sets.

The way in which interactive rationality prunes strategies remains unchanged. That is, let P_i^0 = *Pi* and P_i^n is inductively defined as before. Let *Hn-l* be those information sets that are reachable by profiles of plans from *P'r:: Pi* ∈ P_i^n iff *Pi* ∈ P_i^{n-1} and for every information set *h* ∈ *H(pi) nHn-l* there exists *P-i* ∈ P_{-i}^{n-1} such that $u_i(p_i, p_{-i}) \geq u_i(q_i, p_{-i})$ for every *qi* ∈ *Pi(h).* Now we get the following version of (5.2):

Lemma 7.1. A plan *Pi* of player i is consistent with the set of formulas {r, sb(r), ... ,sbn(r)} iff *Pi* ∈ P_i^{n+1}.

In this case, unlike the generic **PI** case, there could be many optional outcomes in *P?',*

References

Aumann, R.J. (1995). Backward Induction and Common Knowledge of Rationality. *Games and Economie Behaoior, 8(1):6-19.*

Battigalli, P. (1996). Strategie Rationality Orderings and the Best Rationalization Principle. *Games and Economie Behaoior, 13(1):178-200.*

Battigalli, P. (1997). On rationalizability in extensive games. *Journalof Economie Theory,* 74(1):40-6l.

Battigalli, P. & Siniscalchi, M. (1999). Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games. *Journalof Economie Theory,* 88(1):188-230.

Battigalli, P. & Siniscalchi, M. (2002). Strong Belief and Forward Induction Reasoning. *Journalof Economie Theoru,* 106(2):356-39l.

Chellas, B. (1984). *Modal Logic: an Introduetion.* Cambridge University Press.

Pearce, D. (1984). Rationalizable Strategie Behavior and the Problem of Perfection. *Econome irica,* 52(4):1029-1050.

Efficient Coalitions In Boolean Games

Elise Bonzon
Marie-Christine Lagasquie-Schiex
Jérórne Lang

Institut de Recherche en Informatique de TouIouse (IRIT)
Université Paul Sabatier
118 Route de Narbonne
31062 TouIouse Cedex 04, France
{bonzon,lagasq,lang}@irit.fr

Abstract

Boolean games are a logical setting for representing strategie games in a succinct way, taking advantage of the expressive power and conciseness of propositional logic. A Boolean game consists of a set of players, each of whom controis a set of propositional variables and has a specific goal expressed by a propositional formula. We show here that Boolean games are a very simple setting, yet sophisticated enough, for studying coalitions. Due to the fact that players have dichotomous preferences, the following not ion emerges naturally: a coalition in a Boolean game is efficient if it guarantees that the goal of each member of the coalition is satisfied. We study the properties of efficient coalitions, and we give a characterization of efficient coalitions.

1 Introduction

Boolean games (Harrenstein et al., 2001; Harrenstein, 2004; Dunne and van der Hoek, 2004; Bonzon et al., 2006b) are a logical setting for representing strategie games in a succinct way, taking advantage of the expressive power and conciseness of propositional logic. Informally, a Boolean game consists of a set of players, each of whom controls a set of propositional variables and has a specific goal expressed by a propositional formula[1]. Thus, a player in a Boolean game has a dichotomous preferenee relation: either her goal is satisfied or it is not. This restrietion may appeal' at first glance unreasonable. However, many concrete situations can be modelied as games where agents have dichotomous preferences (we give such an example in the paper). iVIoreover, due to the fact that players have dichotomous preferences, the foliowing simple (yet sophisticated enough) notion emerges

1 We refer here to the version of BooIean games defined in (Bonzon et al., 2006b), that generalizes the initiaI proposaI by Harrenstein et al. (2001).

Krzysztof R. Apt, Robert van Rooij (eds.). *New Perspectives on Games and Interaction.* Texts in Logic and Games 4, Amsterdam University Press 2008, pp. 283-297.

naturally: a coalition in a Boolean game is efficient if it guarantees that all goals of the members of the coalition are satisfied. Our aim in the following is to define and characterize efficient coalitions, and see how they are related to the well-known concept of core.

Aftel' recalling some background of Boolean games in Section 2, we study in Section 3 the properties of effectivity functions associated with Boolean games. In Section 4 we study in detail the notion of efficient coalitions. We give an exact characterization of sets of coalitions that can be obtained as the set of efficient coalitions associated with a Boolean game, and we relate coalition efficiency to the notion of core. Related work and further issues are discussed in Section 5.

2 n-player Boolean games

For any finite set $V = \{a, b, \ldots\}$ of propositional variables, L_V denotes the propositional language built up from V, the Boolean constants $\mathbf{T}$ and $\bot$, and the usual connectives. Formulas of L_V are denoted by φ, ψ etc. A *literal* is a variabie x of V or its negation. A *term* is a consistent conjunction of literals. A *clause* is a disjunction of literals. If $\varphi \in L_V$, then $Var(\varphi)$ (resp. $Lit(a)$) denotes the set of propositional variables (resp. literals) appearing in φ.

2^V is the set of the interpretations for V, with the usual convention that for $M \in 2^V$ and $x \in V$, M gives the value *true* to x if $x \in M$ and *false* otherwise. $\models$ denotes the consequence relation of classical propositional logic. Let $V' \subseteq V$. A VI-interpretation is a truth assignment to each variabie of V', that is, an element of $2^{V'}$. VI-interpretations are denoted by listing all variables of V', with a - symbol when the variabie is set to false: for instance, let $V' = \{a, b, d\}$, then the VI-interpretation $M = \{a, d\}$ assigning a and d to true and b to false is denoted by abd. If $Var(\varphi) \subseteq X$, then $Mod_X(\varphi)$ represents the set of X-interpretations satisfying φ.

If $\{V_1, \ldots, V_p\}$ is a partition of V and $\{M_1, \ldots, M_p\}$ are partial interpretations, where $M_i \in 2^{V_i}$, $(M_1, \ldots, M_p)$ denotes the interpretation $M_1 \cup \ldots \cup M_p$.

Given a set of propositional variables V, a Boolean game on V is an n-player game", where the actions available to each player consist in assigning a truth value to each variabie in a given subset of V. The preferences of each player i are represented by a propositional formula φ_i formed using the variables in V.

Definition 2.1. An n-player Boolean game is a 5-tuple (N, V, tt, Γ, Φ), where

2 In the original proposal (Harrenstein et al., 2001), Boolean games are two-players zero-sum games. However the model ean easily be generalized to n players and non necessarily zero-sum games (Bonzon et al., 2006b).

- $N = \{1, 2, \ldots, n\}$ is a set of players (also called agents);
- V is a set of propositional variables;
- $tt : V \to N$ is a control assignment function;
- $\Gamma = \{\gamma_1, \ldots, \gamma_n\}$ is a set of constraints, where each γ_i is a satisfiable propositional formula of $L_{\pi(i)}$;
- $\Phi = \{\varphi_1, \ldots, \varphi_n\}$ is a set of goals, where each φ_i is a satisfiable formula of L_V.

A 4-tuple (N, V, tt, Γ), with N, V, tt, Γ defined as above, is called a pre-Boolean game.

The control assignment function tt maps each variabie to the player who controls it. For ease of notation, the set of all the variables controlled by a player i is written π_i such as $\pi_i = \{x \in V \mid \pi(x) = i\}$. Each variabie is controlled by one and only one agent, that is, $\{\pi_1, \ldots, \pi_n\}$ forms a partition of V.

For each i, γ_i is a constraint restricting the possible strategy profiles for player i.

Definition 2.2. Let $G = (N, V, tt, \Gamma, \Phi)$ be a Boolean game. A strategy[3] for player i in G is a 1fi-interpretation satisfying γ_i. The set of strategies for player i in G is $S_i = \{S_i \in 2^{\pi_i} \mid S_i \models \gamma_i\}$. A strategy profile s for Gis a n-tuple $s = (s_1, s_2, \ldots, s_n)$ where for all i, $s_i \in S_i$. $S = S_1 \times \ldots \times S_n$ is the set of all strategy profiles.

Note that since $\{\pi_1, \ldots, \pi_n\}$ forms a partition of V, a strategy profile s is an interpretation for V, i.e., $s \in 2^V$. The following notations are usual in game theory. Let $s = (s_1, \ldots, s_n)$ be a strategy profile. For any nonempty set of players $I \subseteq N$, the projection of s on I is defined by $s_I = (s_i)_{i \in I}$ and $s_{-I} = s_{N \setminus I}$. If $I = \{i\}$, we denote the projection of s on $\{i\}$ by s_i instead of $s_{\{i\}}$; similarly, we note s_{-i} instead of $s_{-\{i\}}$. π_I denotes the set of the variables controlled by I, and $tt_{-I} = \pi_{N \setminus I}$. The set of strategies for $I \subseteq N$ is $S_I = \times_{i \in I} S_i$, and the set of goals for $I \subseteq N$ is $\Phi_I = \bigwedge_{i \in I} \varphi_i$.

If s and s_i are two strategy profiles, (s_{-I}, s'_I) denotes the strategy profile obtained from s by replacing s_i with s'_i for all $i \in I$.

The goal φ_i of player i is a compact representation of a dichotomous preference relation, or equivalently, of a binary utility function $u_i : S \to \{0, 1\}$ defined by $u_i(s) = 0$ if $s \models \neg\varphi_i$ and $u_i(s) = 1$ if $s \models \varphi_i$. s is at least as good as s_i for i, denoted by $s \succeq_i s_i$, if $u_i(s) \geq u_i(s')$, or equivalently,

3 In this paper, only pure strategies are considered.

if $s \models \neg\varphi_i$ implies $s_i \models \neg\varphi_i$; s is strictly better than s_i for i, denoted by $s \succ_i s'$, if $u_i(s) > u_i(s')$, or, equivalently, $s \models \varphi_i$ and $s_i \models \neg\varphi_i$.

Note that this choice of binary utilities clearly implies a loss of generality. However, some interesting problems, as in Example 4.3, have preferences that are naturally dichotomous, and Boolean games allow to represent these problems in a compact way. Furthermore, Boolean games can easily be extended so as to allow for non-dichotomous preferences, represented in some compact language for preferenee representation (see Bonzon et al., 2006a).

3 Coalitions and effectivity functions In Boolean games

Effectivity functions have been developed in social choice to model the ability of coalitions (Moulin, 1983; Abdou and Keiding, 1991; Pauly, 2001). As usual, a coalition C is any subset of N. N is called the grand coalition. Given a set of alternatives S from which a set of agents N have to choose, an effectivity function Eff: $2^N \to 2^{2^S}$ associates a set of subsets of S with each coalition. $X \in$ Eff(C) is interpreted as "coalition C is effective for X".

Definition 3.1. A coalitional effectivity function is a function Eff: $2^N \to 2^{2^S}$ which is monotonie: for every coalition C $\subseteq N$, $X \in$ Eff(C) implies Y $\in$ Eff(C) whenever $X < Y < S$.

The function Eff associates to every group of players the set of outcomes for which the group is effective. We usually interpret $X \in$ Eff(C) as "the players in C have a joint strategy for bringing about an outcome in X".

In (Pauly, 2001), the meaning of "effective" is precised in the framework of strategie games by defining "a-effectivity": a coalition C $\subseteq N$ is α-effective for $X \subseteq S$ if and only if players in C have a joint strategy to achieve an outcome in X *na matter-* what strategies the other players choose.

As Boolean games are a specific case of strategie games, we would like to define a-effectivity functions in this framework. One of the features of Boolean games is the definition of individual strategies as truth assignments of a given set of propositional variables. We might wonder how restrictive this specificity is. In this section we study Boolean games from the point of view of effectivity functions. Clearly, the definition of S_i as $Mod_{\pi_i}(\hat{I}i)$ induces some constraints on the power of players. Our aim is to give an exact characterization of a-effectivity functions induced by Boolean games. Since in Boolean games the power of an agent i is her goal φ_i, it suffices to consider pre-Boolean games only when dealing wit.h effectivity Iunctions. A pre-Boolean game Ginduces an a-effectivity function Effe as follows:

Definition 3.2. Let G = (N, V, tt, Γ') be a pre-Boolean game. The coalitional a-effectivity function induced by G is the function Effe: $2^N \to 2^{2^S}$

defined by: for any $X \subseteq S$ and any $C \subseteq N$, $X \in \text{Effe}(C)$ if there exists $Sc \in Sc$ sueh that for any $s_{-C} \in S\text{-}C$, $(Sc, s_{-C}) \in X$.[4]

This definition is a partieular ease of the a-effeetivity function indueed by a strategie game (see Pauly, 2001, Chapter 2). Therefore, these functions satisfy the following properties (cf. Pauly, 2001, Theorem 2.27): (i) $\forall C \subseteq N$, $o \notin \text{Eff}(C)$; (ii) $\forall C < N$, $S \in \text{Eff}(C)$; (iii) for all $X < S$, if $\bar{X} \notin \text{Eff}(0)$ then $X \in Eff(N)$; (iv) Eff is superadditive, that is, if for all C, $Cf \subseteq N$ and $X, Y \subseteq S$, $X \in \text{Eff}(C)$ and $Y \in Eff(C^{f})$, then $X \cap Y \in \text{Eff}(C \cup Cf)$. An effeetivity function satisfying these four properties is ealled strongly playable. Note that strong playability implies regularity and eoalition-monotonieity (Pauly, 2001, Lemma 2.26).

However, pre-Boolean games are a speeifie ease of strategie game forms, therefore we would like to have an exaet eharaeterization of those effeetivity functions that correspond to a pre-Boolean game. We first have to define two additional properties. Define $At(C)$ as the minimal sets in $\text{Eff}(C)$, that is, $At(C) = \{X \in \text{Eff}(C) \mid \text{there is no } Y \in \text{Eff}(C) \text{ sueh that } Y \subseteq X\}$.

Atomicity: Eff satisfies *atomicity* if for every $C \subseteq N$, $At(C)$ forms a partition of S.

Decomposability: Eff satisfies *decomposability* if for every $I, J \subseteq N$ and for every $X \subseteq S$, $X \in Eff(I \cup J)$ if and only if there exist $Y \in Eff(I)$ and $Z \in Eff(J)$ sueh that $X = Y \cap Z$.

Note that deeomposability is a strong property that implies superadditivity.

Proposition 3.3. A eoalitional effeetivity function Eff satisfies (1) strong playability, (2) atomieity, (3) deeomposability and (4) $Eff(N) = 2^5 \setminus 0$ if and only if there exists a pre-Boolean game $G = (N, V, tt, \Gamma)$ and an injeetive function $\mu : S \to 2^V$ sueh that for every $C \subseteq N$: $\text{Effe}(C) = \{\mu(X) \mid X \in \text{Eff}(Cn)$.

Sketch of PTOOf.[5] The $(\Leftarrow)$ direetion does not present any diffieulty: we ean easily prove than Effe satisfies strong playability (from Pauly, 2001, Theorem 2.27), atomieity, deeomposability and $Effe(N) = 2^5 \setminus 0$. As μ is a bijeetion between S and $\mu(S)$, these properties transfer to Eff.

For the $(\Rightarrow)$ direetion, we first show that for every $s \in S$, there exists a unique $(Z1, \ldots, Zn)$ sueh that for every i, $Z; \in At(i)$ and $Z1 \cap \ldots nZn = \{s\}$. Then, we build G from Eff as follows:

4 Note that elfectivity functions induced by pre-Boolean games may be equivalently expressed as mappings $\text{Elf} : 2^N \to 2^{L_V}$ from coalitions to sets of logical formulas: $\varphi \in \text{EIf(I)}$ if $Mod_{ç}(\varphi) \in \text{EIf(I)}$. This definition obviously implies syntax-independence, that is, if $\varphi \equiv \psi$ then $\varphi \in \text{Elf}(I)$ iff $\psi \in \text{Elf}(I)$.

5 A complete version of this proof can be found in (Bonzon et al., 2007).

- for every i, number *At(*i): let r_i be a bijective mapping from *At(*i) to {O, 1, ... , *IAt(i)1* - I}. Then create $Pi = \lceil \log_2 |At(i)| \rceil$ propositional variables $x_i^1, \ldots, x_i^{p_i}$. Finally, let V = $\{x_i^j$ I i E N, $1 \leq j \leq p_i\}$;
- for each i: π_i = $\{x_i^1, \ldots, x_i^{p_i}\}$;
- for each i and each $j \leq Pi$, let $\varepsilon_{i,j}$ be the jth digit in the binary representation of *Pi.* Note that ε_{i,p_i} = 1 by definition of *Pi.* If x is a propositional variabie then we use the following notation: $O.X = \neg x$ and $Lx = x$. Then define

$$\gamma_i = \bigwedge_{j \in \{2,\ldots,p_i\}, \varepsilon_{i,j}=0} \left(\bigwedge_{1 \leq k \leq j-1} \varepsilon_{i,j}.x_i^k \rightarrow \neg x_i^j \right)$$

- finally, for each s E S, let μ(s) E 2^V defined by: x_i^j E μ(s) if and only if the jth digit of the binary representation of r_i*(Zi (s))* is 1.

For every i E N and every Z E *At(i),* let k = *Ti(Z)* and *Si(Z)* the strategy of player in i in G corresponding to the binary representation of k using *{Xil, ... , $x_i^{p_i}$}, xiI* being the most significant bit. For instance, ifpi = 3 and *Ti(Zi)* = 6 then *Si(Z)* = *(xiI, Xi2,'Xi3).* Q.E.D.

Note. To follow the proof, it may be helpful to see how this construction works on an example. Let N = {I, 2, 3}, S = {I, 2, 3, 4, 5, 6,7,8,9, *A, B, C},* *At(l)* = {1234, 5678, *9ABC}, At(2)* = *{13579B,2468AC}, At(3)* = *{12569C, 3478AB}* (parentheses for subsets of S are omitted-1234 means {I, 2, 3, 4} and so on). By decomposability, we have *At(12)* = {13, 24, 57, 68, *9B,AC}, At(13)* = *{12,34,56,78,9C,AB},* and *At(23)* = *{159,37B, 26C, 48A}. IAt(I)1* = 3, therefore p: = 2. *IAt(2)1 = IAt(3)1* = 2, therefore *P2 = P3* = 1. Thus, $V = \{x_1 1, x\,12, x21, x31\}$. Let *At(l) = {Zo,* Z1, *Z2},* that is, r_1(1234) = 0, r_1(5678) = 1 and r_1*(9ABC)* = 2. Likewise, *T2(13579B)* = 0, *T2(2468AC)* = 1, *T3(12569C)* = 0 and *T3(3478AB)* = 1. Consider s = 6. We have s = 5678 n *2468AC* n *12569C,* therefore s_c = $\mu(s)$ = $(\neg x_1 1, x_1 2, x_2 1, \neg x_3 1)$. The constraints are γ_1 = $(x_1 1 \rightarrow$ *,x12),* γ_2 = γ_3 = T.

Then, we show that for every C, Effe*(C)* = $\mu(\mathrm{Eff}(C))$. The proof, though rather long, does not present any particular difficulty. See (Bonzon et al., 2007).

4 Efficient coalitions

4.1 Definitions and characterization

We now consider full Boolean games and define *efficient coalitions.* Informally, a coalition is efficient in a Boolean game if and only if it has the ability to jointly satisfy the goals of all members of the coalition:

Definition 4.1. Let G = $(N, V, \pi, \Gamma, \Phi)$ be a Boolean game. A coaJition $C \subseteq N$ is **efficient** if and only if $\exists s_C \in Sc$ such that $\forall s_{-C}$, $Sc \models \bigwedge_{i \in C} \varphi_i$. The set of all efficient coaJitions of a game G is denoted by EC(G).

Example 4.2. Let G = $(N, V, \Gamma, \pi, \Phi)$ where $V = \{a,b,c\}$, $N = \{1,2,3\}$, γ_i = T for every i, $\pi_1 = \{a\}$, $\pi_2 = \{b\}$, $\pi_3 = \{c\}$, $\varphi_1 = (\neg a \wedge b)$, $\varphi_2 = (\neg a \vee \neg c)$ and $\varphi_3 = (\neg b \wedge c)$.

Observe first that $\varphi_1 \wedge \varphi_3$ is inconsistent, therefore no coaJition containing {1,3} can be efficient. {1} is not efficient, because φ_1 cannot be made true only by fixing the value of a; similarly, {2} and {3} are not efficient either. {1, 2} is efficient, because the joint strategy $S\{1,2\} = \overline{a}b$ is such that $S\{1,2\} \models \varphi_1 \wedge \varphi_2$. {2,3} is efficient, because $S\{2,3\} = \overline{b}\overline{c} \models \varphi_2 \wedge \varphi_3$. Obviously, 0 is efficient", because $\varphi_\varnothing = \bigwedge_{i \in \varnothing} \varphi_i \equiv \mathsf{T}$ is always satisfied. Therefore, EC(G) = {0, {1, 2}, {2, 3}}.

From this simple example we see already that EC is neither downward closed nor upward closed, that is, if C is efficient, then a subset or a superset of C may not be efficient. We also see that EC is not closed under union or intersection: {1, 2} and {2, 3} are efficient, but neither {1, 2} ∩ {2, 3} nor {1,2}∪{2,3} is.

Example 4.3 (Kidney exchange, aftel' Abraham et al., 2007). Consider *n* pairs of individuals, each consisting of a recipient *R;* in urgent need of a kidney transplant, and a donor Di who is ready to give one of her kidneys to save *Ri*. As D_i's donor kidney is not necessarily compatible with *Ri*, a strategy for saving more people consists in considering the graph $\langle \{1, \ldots, n\}, E)$ containing a node i ∈ 1, ... ,*n* for each pair (Di, R_i) and containing the edge *(i,j)* whenever D_i's kidney is compatible with R_j. A solution is any set of nodes that can be partitioned into disjoint cycles in the graph: in a solution, a donor Di gives a kidney if and only if *R;* is given one. An optimal solution (saving a maximum number of Jives) is a solution with a maximum number of nodes. The problem can be seen as the following Boolean game G:

- $N = \{1, \ldots, n\}$;
- $V = \{gij \mid i,j \in \{1, \ldots, n\}\}$; *gij* being true means that Di gives a kidney to R_j.
- $\pi_i = \{gij;\ 1 \leq j \leq n\}$;
- for every i, $\gamma_i = \bigwedge_{j \neq k}$ *,(gij* ∧ *gik)* expresses that a donor cannot give more than one kidney.

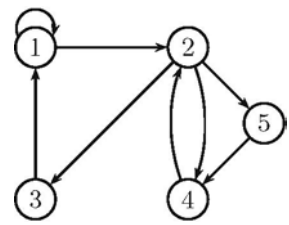
6 One may argue this makes little sense to say that the empty coalition is efficient. Anyway, the definition of an efficient coalition could be changed so as to exclude 0, further notions and results would be unchanged.

- for every i, $\varphi_i = \bigvee_{(j,i)\in E} g_{ji}$ expresses that the goal of i is to be given a kidney that is compatible with R_i.

For example, take $n = 5$ and $E = \{(1,1), (1,2), (2,3), (2,4), (2,5), (3,1), (4,2), (5,4)\}$. Then $G = (N, V, \Gamma, \pi, \Phi)$, with

- $N = \{1,2,3,4,5\}$
- $V = \{g_{ij} \mid 1 \leq i,j \leq 5\}$;
- $\forall i, \gamma_i = \bigwedge_{j\neq k} (g_{ij} \wedge g_{ik})$
- $\pi_1 = \{g_{11}, g_{12}, g_{13}, g_{14}, g_{15}\}$, and similarly for π_2, etc.
- $\varphi_1 = g_{11} \vee g_{31}$; $\varphi_2 = g_{12} \vee g_{42}$; $\varphi_3 = g_{23}$; $\varphi_4 = g_{24} \vee g_{54}$; $\varphi_5 = g_{25}$.

The corresponding graph is depicted below.



Clearly enough, efficient coalitions correspond to solutions. **In** our example, the efficient coalitions are $\emptyset$, {1}, {2,4}, {1, 2, 4}, {1, 2, 3}, {2, 4, 5} and {1,2,4,5}.

We have seen that the set of efficient coalitions associated with a Boolean game may not be downward closed nor upward closed, nor closed under union or non-empty intersection. We find that it is possible to characterize the efficient coalitions of a Boolean game.

Proposition 4.4. Let $N = \{1, \ldots, n\}$ be a set of agents and $SC \subseteq 2^{2^N}$ a set of coalitions. There exists a Boolean game G over N such that the set of efficient coalitions for G is SC (i.e., $EC(G) = SC$) if and only if SC satisfies these two properties:

(1) $\emptyset \in SC$.

(2) for all I, $J \in SC$ such that $I \cap J = \emptyset$, $I \cup J \in SC$.

Thus, a set of coalitions corresponds to the set of efficient coalitions for some Boolean game if and only if (a) it contains the empty set and (b) it is closed by union of disjoint coalitions.

Sketch of proo],' The $(\Rightarrow)$ direetion is proven easily; intuitively, when two disjoint eoalitions I and J are effieient, eaeh one has a strategy guaranteeing its goals to be satisfied, and the joint strategies of I and J guarantee that the goals of all agents in $I \cup J$ are satisfied. As seen in Example 4.2, this is no longer true when I and J interseet. The $(\Leftarrow)$ direetion of the proof is more involved and needs the following Boolean game G to be eonstrueted for eaeh set of eoalitions SC satisfying (1) and (2):

- $V = \{connect(i, j) \mid i, j \in N\}$ (all possible eonneetions between players);
- $\forall i, \gamma_i = \top$;
- $\pi_i = \{connect(i,j) \mid j \in N\}$ (all eonneetions from player i);
- $\varphi_i = \bigvee tese_{,i \in I} F_I$, where

$$r, = \Big(\bigwedge_{j,k \in I} connect(j, k)\Big) \wedge \Big(\bigwedge_{j \in I, k \notin J} \neg connect(j, k)\Big)$$

(player i wants all the players of her eoalition to be eonneeted with eaeh other and diseonneeted from the players outside the eoalition).

We want to show that $EC_G = SC$ (where EC_G is the set of effieient eoalitions for G). We first show that $SC \subseteq EC_G$. Let $I \in SC$. If every agent $i \in I$ plays $\big(\bigwedge_{j \in I} connect(i, j)\big) \wedge \big(\bigwedge_{k \notin I} \neg connect(i, k)\big)$, then φ_i is satisfied for every $i \in I$. Henee, I is an effieient eoalition for G and SC is included in EC(G).

In order to prove that $EC_G \subseteq SC$, we define a *coueruiq of a coalition I by disjoint subsets of SC* as a tuple $\vec{C} = (C_i \mid i \in I)$ of eoalitions sueh that: (a) for every $k \in I$, $Ci \in SC$; (b) for all $Cj, Ci \in C$, either $Cj = Ci$ or $C_j \cap Ci = o$; (e) for every $i \in I$, $i \in Ci$. Let $Cov(SC, 1)$ be the set of all eovering of I by disjoint subsets of SC.

For instanee, if SC {0, 1,24,123, 124} then $Cov(SC, 12)$ {(I, 24), (123, 123), (124,124)}[8], $Cov(SC, 124) = \{(I, 1,24), (1,24,24), (124,124,124)\}$, $Cov(SC, 123) = \{(123, 123, 123)\}$ and $Cov(SC, 234) = Cov(SC, 1234) = 0$.

The proof goes along the following steps:

Ll For any eollection $Col = \{Ci, i = 1, \ldots, q\} \subseteq 2^{2^N}$, $\bigwedge_{1 \le i \le q} F_{C_i}$ is satisfiable if and only if for any $i, j \in \{I, \ldots, q\}$, either $C; = C_j$ or $C, nC_j = o$.

7 A complete version of this proof can be found in (Bonzon et al., 2007).

8 There are two players in $I = \{I, 2\}$, therefore each $\vec{C}$ in $Cov(SC, 12)$ contains 2 coalitions, one for each player, satisfying (a), (b) and (c).

L2 From Ll, we deduee that $\forall I \neq$ o, Φ_I is equivalent to

$$\bigvee_{\text{êECov(SC,I)}} \bigwedge_{i \in T} Fci .$$

L3 From property (2) (assumption of Proposition 4.4) and L2, we ean prove that if $I \subseteq 2^N$, then Φ_I is satisfiable if and only if there exists J E SC sueh that $I < J$.

Let I be an effieient eoalition sueh that $I \notin SC$ (whieh implies $I \neq$ o, beeause by assumption 0 E SC).

- If $I = N$ then there is no J E SC sueh that $I \subseteq J$ (beeause $I \notin SC$), and then L3 implies that Φ_I is unsatisfiable, therefore I eannot be effieient for G.

- Assume now that $I \neq N$ and define the following I-strategy Sr ($I = N \backslash I$): for every i E I, si = {,connect(i,j) Ij E I} (plus whatever on the variables $connect(i,j)$ sueh that $j \notin I$). Let $\vec{C}$ = (Ci,i E I) E $Cov(SC,I)$.

  We first claim that there is a i* E I sueh that C_{i^*} is not eontained in I. Indeed, suppose that for every i E I, C, $\subseteq I$. Then, beeause i E C, holds for every i, we have UiEI C, = I. Now, C, E SC for all i, and any two distinet Ci, C_j are disjoint, therefore, by property (2) we get I E SC, whieh by assumption is false.

  Nuw, let k E $C_{i^*} \setminus I$ (such a k exists beeause C_{i^*} is nut euntained in I). Now, the satisfaetion of $F_{\vec{C}}$ requires $connect(k,$ i*) to be true, beeause both i and k are in Ci. Therefore $s_k \models \bigwedge F_{Cj}$, and a fortiori «$t \models \bigwedge F_{Cj}$, whieh entails «$t \models$ 'ÀiEIFCi.

  This being true for any $\vec{C}$ E $Cov(SC,I)$, it follows that we have «$t \models$ ÀêECov(SC,I) ¬ $\bigwedge_{i \in I} FC_i$ that is, «$t \models \bigvee$ êECov(SC,I) $\bigwedge_{i \in I} Fci$. Together with L2, this entails «$t \models \neg\Phi_I$. Hence, I does not control Φ_I and I eannot be effieient for G.

Q.E.D.

The notion of effieient eoalition is the same as the notion of sueeessful eoalition in qualitative eoalitional games (QCG) introdueed in (Wooldridge and Dunne, 2004), even if, as we diseuss in Seetion 5, QCG and Boolean games are quite different.

4.2 Efficient coalitions and the core

We now relate the notion of efficient coalition to the usual notion of core of a coalitional game. In coalitional games with ordinal preferences, the core is usually defined as follows (see e.g., Aumann, 1967; Owen, 1982; Myerson, 1991): a strategy profile s is in the core of a coalitional game if and only if there exists no coalition C with a joint strategy *Sc* that guarantees that all members of Care better off than with *s*. Here we consider also a stronger notion of core: a strategy profile *s* is in the strong core of a coalitional game if and only if there exists no coalition C with a joint strategy *Sc* that guarantees that all members of C are at least as satisfied as with *s*, and at least one mernber of C is strictly better off than with *s*.

Definition 4.5. Let C be a Boolean game. The (weak) core of C, denoted by WCore(*C*), is the set of strategy profiles $s = (s_1, \ldots, sn)$ such that there exists no eeN and no $Sc \in Sc$ such that for every i E C and every $s_{-C} \in S\text{-}c$, $(sc, s_{-C}) \succ_i s$.

The strong core of a Boolean game C, denoted by SCore(C), is the set of strategy profiles $s = (Sl, \ldots, sn)$ such that there exists no eeN and no $Sc \in Sc$ such that for every i E C and every $s_{-C} \in S\text{-}c$, $(sc, s_{-C}) \succeq_i s$ and there is an i E C such that for every $S\text{-}c \in S\text{-}c$, $(sc, s_{-C}) \succ_i s$.

This concept of weak core is equivalent? to the notion of strong Nash **equilibrium** introduced by Aumann (1959), where coalitions form in order to correlate the strategies of their members. This notion involves, at least implicitly, the assumption that cooperation necessarily requires that players be able to sign "binding agreements": players have to follow the strategies they have agreed upon, even if some of them, in turn, might profit by deviating. However, if players of a coalition C agreed for a strategy *Sc*, at least one player i E C is satisfied by this strategy: we have $\exists i \in C$ such that $s \models \varphi_i$.

The relationship between the (weak) core of a Boolean game and its set of efficient coalitions is expressed by the following simple result. The proofs of following results can be found in (Bonzon et al., 2007):

Proposition 4.6. Let $C = (N, V, \Gamma, tt, \Phi)$ be a Boolean game. Then $s \in$ WCore(C) if and only if *s* satisfies at least one member of every efficient coalition, that is, for every C E EC(*C*), $s \models \bigvee_{i \in C} \varphi_i$.

In particular, when no coalition of a Boolean game C is efficient, then all strategy profiles are in WCore(*C*). Moreover, the weak core of a Boolean game cannot be empty:

9 This equivalence is easily shown: it is just a rewriting of the definition given in (Aumann, 1959).

Proposition 4.7. **For** any Boolean game G, WCore(G) $\neq$ o.

The strong core of a Boolean game is harder to characterize in terms of efficient coalitions. We only have the fol!owing implication.

Proposition 4.8. Let G = (N, V, Γ, tt, Φ) be a Boolean game, and s be a strategy profile. **If** s E SCore(G) then for every C E EC(G) and every i E C, $s \models \varphi_i$.

Thus, a strategy in the st rong core of G satisfies the goal of every mernber of every efficient coalition. The fol!owing counterexample shows that the converse does not hold.

Example 4.9. Let G = (N, V, Γ, tt, Φ) be a Boolean game. We have:

$V = \{a, b, c, d, e, f\}$, $N = \{1, 2, 3, 4, 5, 6\}$, γ_i = T for every i, $\pi_1 = \{a\}$, $\pi_2 = \{b\}$, $\pi_3 = \{c\}$, $\pi_4 = \{d\}$, $\pi_5 = \{e\}$, $\pi_6 = \{f\}$, $\varphi_1 = b \vee d$, $\varphi_2 = a \vee c$, $\varphi_3 = \neg b \vee d$, φ_4 = e, $\varphi_5 = rva. \wedge \neg b \wedge \neg c$ and $\varphi_6 = \neg a \wedge \neg c \wedge <d.$

This game has two efficient coalitions: {1,2} and {2,3}.

Let $s = abcd\bar{e}f$. We have $s \models \varphi_1 \wedge \varphi_2 \wedge \varphi_3 \wedge \neg\varphi_4 \wedge \neg\varphi_5 \wedge \neg\varphi_6$. So, $\forall C$ E EC(G), $\forall i$ E C, $s \models \varphi_i$.

However, $s \notin$ SCore(G): $\exists C' = \{1, 2, 3, 4, 5\}$ C N such that $\exists s_C = abcde \models \varphi_1 \wedge \varphi_2 \wedge \varphi_3 \wedge \varphi_4 \wedge \neg\varphi_5$. So, $\forall s_{-C}$, $(sc, s_{-C}) \succeq_1 s$, $(sc, s_{-C}) \succeq_2 S$, $(sc, s_{-C}) \succeq_3 s$, $(sc, s_{-C}) \succeq_5 s$, and $(sc, s_{-C}) \succ_4 s$. $s \notin$ SCore(G).

Note that the st rong core of a Boolean game can be empty: in Example 4.2, the set of efficient coalitions is $\{0, \{1, 2\}, \{2, 3\}\}$, therefore there is no s E S such that for al! C E EC(G), for al! i E C, $s \models \varphi_i$, therefore, SCore(G) = o. However, we can show than the non-emptiness ofthe st rong core is equivalent to the fol!owing simple condition on efficient coalitions.

Proposition 4.10. Let G = (N, V, Γ, tt, Φ) be a Boolean game. We have the fol!owing:

$Score(G) \neq 0$ if and only if $\bigcup\{C \subseteq N \mid C \in$ EC(G)$\} \in$ EC(G), that is, if and only if the union of al! efficient coalitions is efficient.

5 Conclusion

We have shown that Boolean games can be used as a compact representation setting for coalitional games where players have dichotomous preferences. This specificity led us to define an interesting notion of efficient coalitions. We have given an exact characterization of sets of coalitions that correspond to the set of efficient coalitions for a Boolean game, and we have given several results concerning the computation of efficient coalitions.

Note that some of our notions and results do not explicitly rely on the use of propositional logic. For instance, efficient coalitions can be defined in a more general setting where goals are simply expressed as nonempty sets of states. However, many notions (in particular, the control assignment function π) become much less clear when abstracting from the propositional representation.

Clearly, a limitation of our results is that they apply to dichotomous preferences only. However, as illustrated on Example 4.3, some problems are naturally expressed with dichotomous goals. Moreover, it is always worth starting by studying simple cases, especially when they already raise complex notions-?.

As Boolean games, qualitative coalitional games (QCG), introduced in (Wooldridge and Dunne, 2004), are games in which agents are not assigned utility values over outcomes, but are satisfied if their goals are achieved. A first difference between QCG and Boolean games is that there is no control assignment function in QCG. A second one is that each agent in QCG can have a set of goals, and is satisfied if at least one of her goals is satisfied, whereas each agent in Boolean games has a unique goal. However, QCG's characteristic function, which associates to each coalition C the sets of goals that members of C can achieve, corresponds in Boolean games to the set $W(C) = \{X \subseteq \{\varphi_1, \ldots, \varphi_n\}$ such that $\exists s_C \in Sc : Sc \models \varphi_i\}$[11].

Coalition logic (Pauly, 2001) allows to express, for any coalition C and any formula φ, the ability of C to ensure that φ hold (which is written $[C]\varphi$). **In** Boolean games, the power of agents, expressed by the control assignment function *tt*, is still in the metalanguage. Expressing *tt* within coalition logic would however be possible, probably using ideas Irorn (van der Hoek and Wooldridge, 2005). The next step would then consist in introducing goals into coalition logic. This is something we plan to do in the near future.

References

Abdou, .1. & Keiding, H. (1991). *Ejj"ectivity Functions in Social Choice.* Kluwet.

Abraham, D.J., Bium, A. & Sandholm, T. (2007). Clearing Algorithms for Barter Exchange Markets: Enabling Nationwide Kidney Exchange. **In** MacKie-Mason, .1.*K.*, Parkes, D.e. & Resnick, P., eds., *Proceedinqs Bth*

10 Note that Boolean games can easily be extended with propositional languages for representing compactly nondichotomous preferences (Bonzon et al., 2üü6a).

11 For instanee, we have for Example 4.2 : *W({I})* = *W({3})* = *W({1,3})* *{'P2},* *W({2})* = o, *W({1,2})* = $\{\varphi_1, \varphi_2\}$, *W({2,3})* = $\{\varphi_2, \varphi_3\}$, *W({1,2,3})* $\{\{\varphi_1, \varphi_2\}, \{\varphi_2, 'P3\}\}'$

ACM Conference on Elecironic Commerce (EC-2007), San Diego, California, USA, June 11-15, *2007,* pp. 295-304. ACM.

Aumann, R.J. (1959). Acceptable points in general n-person games. **In** Tucker, A.W. & Luce, R.D., eds., *Conirioui ions to the Theoru of Games IV.* Princeton University Press.

Aumann, R.J. (1967). A Survey of Cooperative Games without Side Payments. **In** Shubik, M., ed., *Essays in Jvlathematical Economics in HonOUT of Oskar Morgenstern.* Princeton University Press.

Bonzon, K, Lagasquie-Schiex, M.-C. & Lang, .1. (2006a). Compact Preference Representation for Boolean Games. **In** Yang, Q. & Webb, G.I., eds., *PRICAI 2006: Trends in Artificial Intelligence, 9th Pacific Rim International Conference on Artificial Intelligence (PRICAI'06),* Vo!. 4099, pp. 41-50. Springer.

Bonzon, K, Lagasquie-Schiex, M.-C. & Lang, .1. (2007). Efficient Coalitions in Boolean Games. Research report IRITjRR-2007-13-FR, Institut de Recherche en Informatique de Toulouse.

Bonzon, K, Lagasquie-Schiex, M.-C., Lang, .1. & Zanuttini, B. (2006b). Boolean Games Revisited. **In** Brewka, G., Coradeschi, S., Perini, A. & Traverso, P., eds., *ECAI2006, 1'lth Europeon Conference on Artificial Intelligence, August* 29 - *September* 1, *2006, Riva del Garda, Italy, Including Prestiqious Applications of Intelligent Systems (PAIS 2006), Proceedinqs,* pp. 265-269. lOS Press.

Dunne, P.K & van der Hoek, W. (2004). Representation and complexity in boolean games. **In** Alferes, .1..1. & Leite, .1.A., eds., *Proceedinqs of Logics in Ariificiol Intelligence, 9th Europeon Conference, JELIA 2004,* Vo!. 3229 of *Leciure Notes in Computer Science,* pp. 347-359.

Harrenstein, P. (2004). *Logic in Conflict.* **PhD** thesis, Utrecht University.

Harrenstein, P., van der Hoek, W., Meyer, .1.-.1. & Witteveen, C. (2001). Boolean games. **In** van Benthem, .1., ed., *Proceedings of the 8th Conference on Theoretncal Aspects of Rationality and Knowledge (TARK-2001), Ceriosa di Pontignano, Uruoersits] of Siena, Italy, July 8-10, 2001,* pp. 287-298. Morgan Kaufmann.

van der Hoek, W. & Wooldridge, M. (2005). On the Logic of Cooperation and Propositional Contro!. *Artificial Intelligence,* 164(1-2):81-119.

Moulin, H. (1983). *The Strategy of Social Choice.* North-Holland.

Myerson, **R.B.** (1991). *Game Theory: Analysis of Conflict.* Harvard University Press.

Owen, G. (1982). *Game Theory.* Academie Press.

Pauly, M. (2001). *Logic [or Social Software.* **PhD** thesis, University of Amsterdam. *ILLC Publications* DS-2001-10.

Wooldridge, M. & Dunne, P. (2004). On the Computational Complexity of Qualitative Coalitional Games. *Artificial !ntel!igence, 158(1):27-73.*

Interpretation of Optimal Signals

Michael Franke

institute for Logic, Language and Computation
Universiteit van Amsterdam
Nieuwe Doelenstraat 15
1012 CP Amsterdam, The Netherlands
ID. franke@uva.nl

Abstract

According to the optimal assertions approach of Benz and van Rooij (2007), conversational implicatures can be calculated based on the assumption that a given signal was optimal, i.e. that it was the sender's best choice if she assumes, purely hypothetically, a particular naive receiver interpretation behavior. This paper embeds the optimal assertions approach in a general signaling game setting and derives the notion of an optimal signal via a series of iterated best responses (cf. .Jäger, 2007). Subsequently, we wil! compare three different ways of interpreting such optimal signaIs. It turns out that under a natural assumption of expressibility (i) the optimal assertions approach, (ii) iterated best response and (iii) strong bidirectional optimality theory (Blutner, 1998, 2000) all prove equivalent. We then proceed to show that, if we take the iterated best response sequence one step further, we can account for M-implicatures (Horn's division of pragmatic labor) standardly in terms of signaling games.

Often we express more with the use of our words than what those words mean literal!y. For example, if you were to say that this observation is not particularly new, I would clearly get the hint and understand that you meant to say that it is more than just not particularly new, indeed a werking standard in linguistic pragmatics. Such *conoersationol implicaiures* were first studied by Grice (1989) and stil! concern the community in various ways. **In** particular, recent years saw an increasing interest in game-theoretic models of conversational implicature calculation, and this study belongs to this line of research. **It** provides a formal comparison of selected previous approaches which extends to a uniform synchronie account of different kinds of conversational implicatures.

The paper is organized as fol!ows. Section 1 briefiy reviews the classification of conversational implicatures into 1-, Q- and M-implicatures. Section 2 introduces a game-theoretical model of implicature calculation: a signaling game with exogenously meaningful signais. We wil! see in Section 2.3 that

Krzysztof R. Apt, Robert van Rooij (eds.). *New Perspectives on Games and Interaction.* Texts in Logic and Games 4, Amsterdam University Press 2008, pp. 299-312.

the standard solution concept for signaling games is not strong enough to account for the empirical observations. The *optimal asserteems approach.* of Benz and van Rooij (2007), which is introduced in Section 3.1, is an attempt to solve this problem. According to the optimal assertions approach, conversational implicatures can be calculated based on the assumption that a given signal was *optima!*. Sections 3.2 and 3.3 then compare three ways of interpreting such optimal signais: (i) the pragmatic interpretation rule of Benz and van Rooij (2007), (ii) iterated best response and (iii) strong bidirectional optimality theory (Blutner, 1998, 2000). **It** turns out that if we assume a sufficiently expressible stock of possible signais, all three approaches prove equivalent. However, it also turns out that M-implicatures (Horn's division of pragmatic labor) cannot be accounted for based solely on the assumption that the received form was optima!. We will conclude that some aid from the refinement literature, in particular Cho's and Kreps' (1987) intuitive criterion, is necessary and sufficient to account uniformly for all 1-, Q- and M-implicatures.

1 Kinds of conversational implicatures

Neo-Gricean pragmatics (Atlas and Levinson, 1981; Horn, 1984) distinguishes I-implicatures (1) and Q-implicatures (2).

(1) John has a very efficient secretary.
$\rightsquigarrow$ John has a very efficient *female* secretary.

(2) John invited some of his friends.
$\rightsquigarrow$ John did not invite all of his friends.

I-implicatures like (1) are inferences to a stereotype: the sentence is associated with the most likely situation consistent with its semantic meaning. Q-implicatures like (2), also called scalar implicatures, are a strengthening of the literal meaning due to the presence of more informative alternatives that were not used: since the speaker only said that some of John's friends were invited, we infer that the compatible stronger claim that all of John's friends were invited-a claim that we may assume relevant if true---does not hold, for otherwise the speaker would have said so-as she is assumed cooperative and informed.

A third kind of implicature, called M-implicature by Levinson (2000), is given in (3).

(3) The corners of Sue's lips turned slightly upwards.
$\rightsquigarrow$ Sue didn't smile genuinely, but faked a smile.

In (3) we naturally infer that something about the way Sue smiled was abnormal, non-stereotypical or non-standard, because the speaker used a long and complicated form where she could have used the simple expression (4).

(4) Sue smiled.

M-implicatures were also discussed by Hom (1984) and have been addressed as *Horn's division of pmgmatic labor* thereafter. It has become customary to assume that both sentences (3) and (4) are semantically equivalent, but, when put to use, the longer form (3) gets to be associated with the non-stereotypical situation, while the short form (4) gets to be associated with the stereotypical situation.

2 Implicatures via signaling games

2.1 Interpretation frames

A fairly manageable set of contextual parameters plays a role in the neo-Gricean ciassification of implicatures: we distinguish various meanings that are more or less stereotypical and we compare different forms with respect to their semantic meaning and complexity. We can then capture any such configuration of contextual parameters that are relevant for the computation of implicatures in an *inierpretation frame.*

Definition 2.1 (Interpretation Frame). An interpretation frame is a tuple

$$\mathcal{F} \stackrel{\text{def}}{=} (W, P, F, \text{c}, [.]])$$

where W is a finite set of worlds or situations, P is a probability distribution over W with the usual properties,[1] F is a set of forms or signals which the sender may send, $\text{c} : F \to \mathbb{R}$ is a cost function and $[\![\cdot]\!] : F \to \mathscr{P}(W)$ is a semantic denotation function mapping forms to subsets of W.

We assume for convenience that $P(w) \neq 0$ for all worlds $w \in W$. We would also like to rule out certain rather pathological situations where there are worlds which simply cannot be expressed by any conventional signal:

Assumption 2.2 (Semantic Expressibility). We only consider interpretation frames in which all worlds are semantically expressible: for all worlds w there has to be a form f such that $w \in [f]]$.

The kinds of implicatures described in the previous section correspond to abstract interpretation frames as follows:

- The *I-frame* is an interpretation frame $\mathcal{F}_I = (W, P, F, \text{c}, [.]])$ where $W = \{w,v\}$, $P(w) > P(v) \neq 0$, $F = \{j,g,h\}$, $cU) < c(g),c(h)$ and $[f]] = W$, $[g] = \{v\}$ and $[h] = \{w\}$. The observed *I-implicature play* is to interpret f as wand to send f in w only.

[1] $P(w) \in [0,1]$, for all $w \in W$; $P(A) = \sum_{w \in A} P(w)$, for all $A \subseteq W$; $P(W) = 1$.

- The *Q-fmme* is an interpretation frame $\mathcal{F}_Q$ = *(W,P,F,c,* [.]) where $W = \{w,v\}$, $P(w) \geq P(v) \neq 0$, $F = \{j,g\}$, $c(j) = c(g)$ and $[f]1 = W$, $[g] = \{v\}$. The observed *Q-implicature play* is to interpret f as wand to send f in w only.

- The *lvI-fmme* is an interpretation frame $\mathcal{F}_{NI}$ = *(W, P, F,* c, [.]]) where $W = \{w,v\}$, $P(w) > P(v) \neq 0$, $F = \{j,g\}$, $c(j) < c(g)$ and [1] = $[g]1 = W$. The observed *lvI-implicature play* is to interpret f as wand to send f in w only, as wel! as to interpret *9* as v and to send *9* in v only.

2.2 **Interpretation** games

Interpretation frames capture the relevant aspects of the situation in which communication takes place. The communication itself can best be imagined as a signaling game: nature selects a world $w \in W$-cal! it the actual world in a given play-with probability $P(w)$ and reveals it to the sender who in turn chooses a form $f \in F$. The receiver does not observe the actual world, but observes the signal f. He then chooses an action A. Sender and receiver receive a payoff based on w, f and A. In the present context, we are interested in *interpretation games:* signaling games in which signals have a conventional, compel!ing meaning that the receiver tries to interpret by choosing an interpretation action $0 \neq A \subseteq W$.

Definition 2.3 (Interpretation Game). An interpretation game is just an interpretation frame to which interpretation actions and utilities for sender and receiver are added, in other words a tuple

$$\mathcal{G} \stackrel{\text{def}}{=} \langle \mathcal{F}, Act, u_S, u_R \rangle$$

where $\mathcal{F}$ = *(W, P, F,* c, [.]]) is an interpretation frame, Act def $\mathscr{P}(W) \setminus 0$ is a set of *interpretation actions* and $u_x : F$ x Act x $W \to \mathbb{R}$ are utility functions of sender and recelver.é

$$u_R(f, A, w) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{|A|} & \text{if } w \in A \text{ and } w \in [f]1 \\ 0 & \text{if } w \notin A \text{ and } w \in [f]1 \\ -1 & \text{otherwise} \end{cases}$$

$$u_S(f, A, w) \stackrel{\text{def}}{=} u_R(f, A, w) - C(j).$$

2 These utilities reftect the mutual desire to communicate which world is actual: the more the receiver narrows down a correct guess the better; miscommunication, on the other hand, is penalized so that ifthe chosen interpretation does not include the actual situation, the payoff is strictly smaller than when it does; a strong penalty is given for communication that deviates from the semantic meaning of messages to enforce the exogenous meaning of signals. (This last point is objectionable, but it is also not strictly necessary. I adept it for ease of exposition since space is limited.)

As usual, we identify the receiver's probabilistic beliefs with the probability distribution $P(\cdot)$. Costs are assumed *nominal:* they are small enough to make a utility difference for the sender for any two different signals i and f' only in case $u_R(f, A, w) = u_R(f', A, w)$.

Definition 2.4 (Strategies). A *sender stmtegy* is a function $\sigma : W \to \mathscr{P}(F) \setminus 0$ that specifies a set $\sigma(w) \subseteq F$ of messages to be sent with equal probability when in world *w.* We call asender strategy σ *truih-respectsnq* iff for all wand i whenever $i \in \sigma(w)$ we have $w \in [i]$. We define also $\sigma^{-1}(f) \stackrel{\text{def}}{=} \{w \in W \mid i \in \sigma(w)\}$. Finally, a *recesoer stmtegy* is a function $p : F \to Act$ specifying an interpretation for each message.

Whether an action is preferabie to another depends on what the other party is doing. If we fix a strategy for the other party we can define the expected utility of each action.

Definition 2.5 (Expected Utilities). Since the sender knows the actual world *w,* his expected utility of sending the form $i \in F$ given that the receiver plays p is actually just his utility in w given i and the receiver's response $p(J)$:

$$EUs(J, p, w) \stackrel{\text{def}}{=} u_S(f, p(J), w).$$

Given that the sender plays σ, the receiver's expected utility of interpreting a form i for which $\sigma^{-1}(J) \neq 0$ as $A \in Act$ is:[3]

$$\mathrm{EU}_R(A, \sigma, f) \stackrel{\text{def}}{=} \sum_{w \in W} P(w|\sigma^{-1}(f)) \times u_R(f, A, w)$$

For a truth-respecting sender strategy this simplifies to:

$$\mathrm{EU}\, R(A, \sigma, j) = \frac{P(A|\sigma^{-1}(f))}{|A|}. \tag{2.1}$$

If the other party's strategy is given, rationality requires to maximize expected utility. A strategy X that maximizes expected utility in all its moves given the other party's strategy Y is called a *best response* to Y. For some sender strategies σ and forms i it may be the case that several actions maximize the receiver's expected utility, and that therefore there is no unique best response. Given Equation 2.1, it is easy to see that all (non-empty) sets that contain only worlds which are maximally likely according to $P(\cdot|\sigma^{-1}(J))$ are equally good interpretations in expectation:"

$$\mathrm{Max}_{A \in Act} \mathrm{EU}\, R(A, \sigma, j) = \mathscr{P}(\mathrm{Max}_{w \in W} P(w|\sigma^{-1}(f))) \setminus o.$$

3 We wiII come back to the question how to interpret messages I in the light of sen der strategies σ that never use I in Sections 3.2 and 3.4. For the time being, assume that $\mathrm{EU}_R(A, \sigma, I) = 0$ is constant for all A if $\sigma^{-1}(I) = 0$.

4 We write MaxxEXF(x) der $\{x \in X \mid \neg\exists x' \in X\colon F(x) < F(x')\}$, for arbit.rary set X and function $F : X \to$ IR.

Assumption 2.6 (Preferred Interpretation). We assume that the receiver selects as his best response to a truth-respecting σ and f the largest interpretation action $\mathrm{Max}_{w \in W} P(w|\sigma^{-1}(f))$. This is because the receiver should not discard any possible interpretation without reason; one should not gambie on proper understanding.f

The standard solution concept for rational play in a signaling game is a perfect Bayesian equilibrium: a pair of strategies that are best responses to one another.

Definition 2.7 (Perfect Bayesian Equilibrium). A pair of strategies $\langle \sigma, p)$ is a perfect Bayesian equilibrium iff

(i) for all wE W: $\sigma(w)$ E Max!EFEUs(J, p, w)

(ii) for all fE F: $p(J)$ E MaxAEActEU $R(A, \sigma, J)$.

2.3 Pragmatics & the problem of equilibrium selection

It is easy to verify that 1-, Q- and M-implicature play are all perfect Bayesian Equilibria (PBEs) in the corresponding interpretation games, but not uniquely so. Indeed, the straight-forward signaling games approach to implicature computation faces a *problem of equilibrium selection:* why is it that particular PBEs are observed and not others?

A natural way of answering this question is to formulate refinements of the assumed solution concept. An interesting proposal along these lines is given by van Rooij (2008) who observes that the Q-implicature play can be singled out as the unique *neologism proof* **PBE** (Farrell, 1993) and that the M-implicature play can be singled out with the help of Cho's and Kreps' *intuitive criterion* (Cho and Kreps, 1987). We will piek up this latter idea in Section 3.4. Notice, however, that van Rooij's approach deviates from a standard signaling game analysis, because in order to arrive at the desired prediction for the M-frame, van Rooij considers a transition from an interpretation frame with just the cheaper message i, to which at a later stage the more costly message *9* is added. The question remains whether we cannot account for the observed implicature plays in more conservative terms.

3 Association-optimal signaling

A recent framework that seeks to give a positive answer to this question is Benz and van Rooij's (2007) *optimal assertions approach.* The basic idea is that the receiver may compute implicatures based on the assumption that the signal he received was an *optimal aseertion.* An optimal assertion in turn is the best response to a naive, hypothetical interpretation of messages

5 This assumption replaces the tie-break rule of Benz and van Rooij (2007).

that takes into account only the semantic meaning of the message and the probabilities of worlds. Benz and van Rooij describe their set-up as a sequence of dec ision problems: on the hypothesis that the receiver interprets signals in a certain, naive way, the sender will choose signals that are optimal given this receiver strategy and the receiver can then interpret messages as optima!.

Another way of looking at this process is as a sequence of *iteraied best responses* (cf. Jäger, 2007). To point out the connection, I will spell out the details of the optimal assertions approach in terms of iterated best responses in Section 3.1. I will then, in Section 3.2, show that Benz's and van Rooij's interpretation rule deviates slightly from the former iterated best response logic in general, but that for a natural subclass of interpretation frames-including 1- and Q-frames-the two approaches fall together. **In** Section 3.3, finally, I will conneet both the optimal assertion and the iterated best response approach with strong bidirectional optimality theory.

3.1 Association optimality

We start with the assumption that the sender says something true:

$$\sigma_0(w) = U \in F \mid w \in [f]\} .$$

We also assume that, given that the sender says something true, the receiver will interpret messages as true; in other words, as the sender starts with a naive 'truth-only' strategy σ_0, the receiver maximizes his expected utility based on that strategy and plays (as σ_0 is truth-respecting):

$$\begin{aligned} po(J) &= \mathrm{Max}_{w \in W} P(w|\sigma_0^{-1}(f)) \\ &= MaxwEWP(wl\ [fJl)\cdot \end{aligned}$$

We could think here of a spontaneous, first associative response to the message *f:* the most likely worlds in which f is true are chosen as the first interpretation strategy, because these are the worlds that spring to mind first when hearing f. We therefore call *Po* the receiver's *association response.*

The association response *Po* is of course a bad interpretation strategy. **In** fact, it is not a pragmatic interpretation strategy at all, for it leaves out all considerations about the interpretation game except $[\![\cdot]\!]$ and $P(\cdot)$: receipt of message f is treated as if it was the observation of the event *[fJl.* But still the association response *Po* is the rational response to the-admittedly non-pragmatic-sender strategy σ_0. The guiding conviction here is that pragmatic reasoning takes semantic meaning as a starting point: if I want to know what you meant by a given linguistic sign, I first feed into the interpretation machine the convention al meaning of that sign. Therefore, as σ_0 is a natural beginning, so is the association response *PO.6*

6 An anonymous reviewer asks for the difference between Jä.ger's (2007) evolutionary

But if this truly is the most reasonable beginning for pragmatic interpretation, the sender may anticipate the receiver's association response P_0 and choose a best response to it:

$$\sigma_1(w) = \text{MaX!EFEUs(J, } P_0, w) = \{f \in F \mid \neg\exists f' \in F : EU_S(J, P_0, w) < \mathrm{EU}_S(f', P_0, w)\}$$

Forms in σ_1 are optimal forms given the receiver's association response. We could therefore call them *association optimal,* or, for short, *optimal:* a form $i \in F$ is (association) optimal in a world w iff $i \in \sigma_1(w)$.

How should the receiver interpret an optimal signal? We'll next consider and compare three possible answers to this question.

3.2 **Optima1** assertions **and** iterated best response

Given semantic expressibility as stated in Assumption 2.2, association optimality is equivalent to Benz's and van Rooij's (2007) notion of an optimal assertion. Although the latter notion requires truth of a message for its optimality, it is easy to see that semantic expressibility and optimality entail truth.

Observation 3.1. Given semantic expressibility, σ_1 is truth-respecting.

Proof. Let some $i \in F$ be false in $w \in W$. From semantic expressibility there is a message $t' \in F$ which is true in w. But then $-1 = u_S(f, PO(J), w) < 0 \leq u_S(f', PO(JI), w)$, so that i is not association optimal in w. Q.E.D.

If the sender sends an association optimal signal, i.e. if the sender sticks to σ_1, the receiver can again interpret accordingly. Benz and van Rooij propose the following interpretation rule based on the assumption that the received signal was an Optimal Assertion: $\rho_1^{\mathrm{OA}}(J) = \{w \in [ij] \mid i \text{ is optimal in } w\}$. Thich simplifies under Observation 3.1 to

$$\rho_1^{\mathrm{OA}}(f) = \sigma_1^{-1}(f). \tag{3.1}$$

Notice, however, that this may not be a well-defined receiver strategy in our present set-up, for it may be the case that $\sigma_1^{-1}(f) = 0$, which is not a feasible interpretation action. The same problem also occurs for the best response to σ_1. It is clear what the best response to σ_1 is for messages that may be optimal somewhere: if $\sigma_1^{-1}(f) \neq 0$, we have

$$\rho_1^{\mathrm{BR}}(f) = \mathrm{Max}_{w \in W} P(w|\sigma_1^{-1}(f)). \tag{3.2}$$

model, which also uses best response dynamics, and the present synchronie approach. One obvious difference is that the present model assumes that at each turn a best response is selected with probability 1. Another difference is the starting point: in Jä.ger's model it is the sender, while in the present model it is the receiver who responds first to a strategy that is given by the semantic meaning of the signals.

But how should a best response to σ_1 interpret messages that are never optimal? Since we defined (tentatively, in Footnote 3) expected utilities as constant for all $A \in Act$ whenever $\sigma^{-1}(J) = o$, any $A \in Act$ is an equally good interpretation for a non-optimal f. For our present purpose-the comparison of frameworks-it is not important what to choose in this case, as long as we choose consistently. We therefore adopt the following assumption and refiect on it in Section 3.4 where it plays a crucial role.

Assumption 3.2 (Uninterpretability Assumption). We assume that the receiver resorts to the mere semantic meaning in case a message is uninterpretabie: if $\sigma_1^{-1}(f) = o$, then $\rho_1^{OA}(f) = \rho_1^{BR}(f) = [f]$.

With this we can show that $\rho_1^{BR}(f)$ entails $\rho_1^{OA}(f)$ for arbitrary f and interpretation frames. Moreover, ρ_1^{OA} also entails ρ_1^{BR}, if we assume *strorui expTessibility:*

Definition 3.3 (Strong Expressibility). An interpretation frame satisfies strong expressibility if each world is immediately associated with some message: for each world w there is a form f such that $w \in po(J)$.

Observation 3.4. Under strong expressibility, association optimality implies inclusion in the association response: if f is association optimal in w, then $w \in po(J)$.

Proo]. Assume strong expressibility. If $w \notin po(J)$, there is a form f' for which $w \in PO(JI)$. But then $0 = u_S(f, po(J), w) < u_S(f', PO(JI), w)$. So f is not association optimal in w. Q.E.D.

Proposition 3.5. For arbitrary interpretation frames it holds that $\rho_1^{BR}(f) \subseteq \rho_1^{OA}(f)$. For interpretation frames satisfying strong expressibility it holds that $\rho_1^{BR}(f) = \rho_1^{OA}(f)$.

Proo]. We only have to look at the non-trivial case where $\sigma_1^{-1}(f) \neq o$. Let $w \in \rho_1^{BR}(f)$. Since all worlds have non-zero probabilities we can conclude that $w \in \sigma_1^{-1}(f)$. Hence, $wE\ \rho_1^{OA}(f)$.

Let $w \in \rho_1^{OA}(J)$ and assume strong expressibility. Then $w \in [f]$ and $f \in \sigma_1(w)$. From Observation 3.4 we then know that $w \in Po(J)$. That means that there is no $u_{i'}$ for which $P(wll\ [f]) > P(wl\ [f]l)$. But since, by Observation 3.1, we know that $\sigma_1^{-1}(f) \subseteq [f]l$, we also know that there is no $u_{i'}$ for which $P(w'|\sigma_1^{-1}(f)) > P(w|\sigma_1^{-1}(f))$. Hence $w \in \rho_1^{BR}(f)$. Q.E.D.

3.3 Strong bidirectional optimality theory

A similar conneetion holds with strong Bi-OT (Blutner, 1998,2000). At first sight, Bi-OT looks rather different from game-theoretic modeis, because in Bi-OT we compare form-meaning pairs $\langle f, w \rangle$ with respect to a preference

order. The idea is that to express a given meaning w with a form i, the form-meaning pair $\langle f, w\rangle$ has to be strongly optimal. Likewise, a form f will be associated with meaning w if and only if $\langle f, w\rangle$ is strongly optimal.

Definition 3.6 (Strong bidirectional optimality). A form-meaning pair $\langle f, w\rangle$ is strongly optimal iff it satisfies both the Q- and the I-principle, where:

(i) $\langle f, w\rangle$ satisfies the Q-principle iff $\neg\exists f' : \langle f', w\rangle > \langle f, w\rangle$

(ii) $\langle f, w\rangle$ satisfies the I-principle iff $\neg\exists w' : \langle f, w'\rangle > \langle f, w\rangle$

How should we define preference relations against the background of an interpretation game? Recall that the Q-principle is a sender economy principle, while the I-principle is a hearer economy principle. We have already seen that each interlocutor's best strategy choice depends on what the other party is doing. So, given σ_0 and P_0 as a natural starting point we might want to define preferences simply in terms of expected utility:

$$\langle f, w\rangle > \langle f, w\rangle \quad \text{iff} \quad \mathrm{EU}_S(f', P_0, w) > \mathrm{EU}_S(f, P_0, w)$$
$$\langle f, w\rangle > \langle f, w\rangle \quad \text{iff} \quad \mathrm{EU}_R(\{w'\}, \sigma_0, f) > \mathrm{EU}_R(\{w\}, \sigma_0, f)$$

This simplifies to:[7]

$$\langle f', w\rangle > \langle f, w\rangle \quad \text{iff} \quad u_S(f', \rho_0(f'), w) > u_S(f, \rho_0(f), w)$$
$$\langle f, w\rangle > \langle f, w\rangle \quad \text{iff} \quad P(w \mid [\![f]\!]) > P(w' \mid [\![f]\!]).$$

Observation 3.7. Interpretation based on optimal assertions $\rho_1^{\mathrm{OA}}(f)$ is strong Bi-OT's Q-principle: a form-meaning pair $\langle f, w\rangle$ satisfies the Q-principle iff $\sigma_1^{-1}(f) \neq \emptyset$ and $w \in \rho_1^{\mathrm{OA}}(f)$.

Proof. A form-meaning pair $\langle f, w\rangle$ satisfies the Q principle iff there is no f' such that $\mathrm{EU}_S(f', P_0, w) > \mathrm{EU}_S(f, P_0, w)$ iff f is association optimal in w iff $\sigma_1^{-1}(f) \neq \emptyset$ and $w \in \rho_1^{\mathrm{OA}}(f)$. Q.E.D.

Let's capture interpretation based on strong optimality in an interpretation operator for ease of comparison. If $\sigma_1^{-1}(f) = \emptyset$, the uninterpretability assumption holds, and we take $\rho_1^{\mathrm{OT}}(f) = [\![f]\!]$; otherwise: $\rho_1^{\mathrm{OT}}(f) = \{w \in W \mid \langle f, w\rangle$ is strongly optimal$\}$, which is equivalent to:

$$\rho_1^{\mathrm{OT}}(f) = \{w \in \mathrm{Max}_{v \in W} P(v \mid [\![f]\!]) \mid f \in \sigma_1(w)\}. \tag{3.3}$$

7 Originally, Blutner (1998) defined preferences in terrns of a function C that maps forrn-meaning pairs to real numbers, where $C(\langle f, w\rangle) = c(f) \times -\log_2 P(w \mid [\![f]\!])$. Form-meaning pairs were then ordered with respect to their C-value. Om formulation here amounts basically to the same, but further integrates the present assumption that casts are nominal and only sender relevant.

Proposition 3.8. For arbitrary interpretation frames it holds that $\rho_1^{\mathrm{OT}}(\mathsf{J}) \subseteq \rho_1^{\mathrm{OA}}(f)$. For interpretation frames satisfying strong expressibility it holds that $\rho_1^{\mathrm{OT}}(\mathsf{J}) = \rho_1^{\mathrm{OA}}(J)$.

Proo]. The first part is an immediate consequences of Observation 3.7. So assume st rong expressibility and let $\sigma_1^{-1}(\mathsf{J}) \neq 0$ and $w \in \rho_1^{\mathrm{OA}}(J)$, so that $f \in \sigma_1(w)$. From Observation 3.4 we know that therefore $w \in po(J)$. So there is no ui' for which $P(wll\ [\mathrm{j}]) > P(wl\ [f]])$. But that means that $\langle f, w)$ also satisfies the I-principle, and therefore $w \in \rho_1^{\mathrm{OT}}(f)$. Q.E.D.

Proposition 3.9. For arbitrary interpretation frames it holds that $\rho_1^{\mathrm{OT}}(\mathsf{J}) \subseteq \rho_1^{\mathrm{BR}}(f)$. For interpretation frames satisfying strong expressibility it holds that $\rho_1^{\mathrm{OT}}(\mathsf{J}) = \rho_1^{\mathrm{BR}}(J)$.

Proo]. Let $\sigma_1^{-1}(f) \neq 0$ and $w \in \rho_1^{OT}(f)$. Then $w \in MaxvEWP(vl\ [f]l)$ and $f \in \sigma_1(w)$. Suppose that there was a $ui' \in W$ with $P(w'|\sigma_1^{-1}(f)) > P(w|\sigma_1^{-1}(f))$. Then $ui' \in \sigma_1^{-1}(f)$, but $ui' \notin [f]$ This contradiets Observation 3.1. The rest follows from Propositions 3.5 and 3.8. Q.E.D.

3.4 Interpretation of optima1 signals

The results of the last sections are graphically represented in Figure 1. What do these results teil us about the respective interpretation rules? **In** particular, what are the conceptual differences between the approaches? Can we conclude that one is better than the other? A quick glance at Equations 3.1, 3.2 and 3.3 reveals that the only difference between frameworks lies in the treatment of probabilities.f The optimal assertions approach does not take probabilities into account, iterated best response chooses the most likely interpretations where the received message was optimal and Bi-OT chooses all those most likely interpretations given the semantic meaning of the message where that message was optima!.

The simplest case where predictions differ is where the to be interpreted message f is true in three worlds, $[f]l = \{w, v, zz\}$, and optimal in two worlds, $\sigma_1^{-1}(f) = \{v, u\}$, with varying degree of probability: $P(w) > P(v) > P(u)$. **In** this case, the optimal assertions approach selects $\rho_1^{\mathrm{OA}}(\mathsf{J}) = \sigma_1^{-1}(\mathsf{J}) = \{v, u\}$, iterated best response selects $\rho_1^{\mathrm{BR}}(f) = \{v\}$, while Bi-OT selects $\rho_1^{\mathrm{OT}}(f) = \mathsf{O}$.

This seems to speak for iterated best response, maybe for optimal assertions, but somehow against Bi-OT. On the other hand, we might also credit Bi-OT for its strict continuation of the idea that probabilities eneode stereotypes in an associative salienee ordering: upon hearing f the associations $po(J)$ spring to mind and those are checked for optimality, so that, if

8 Clearly then, for uniform probability distributions strong expressibility collapses into semantic expressibility and all frameworles behave the exact same.

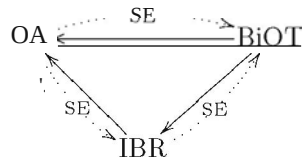


FIGURE 1. Conneetion between (i) optimal assertions (OA), (ii) iterated best response (IER) and (iii) (strong) bidirectional optimality theory (BiOT): a straight arrow indicates inclusion of interpretations of signals while a dotted arrow with label SE indicates inclusion given strong expressibility.

the received message is not optimal in any ofthe associated worlds in *po(J),* then the receiver is stuck-at least for the time being; he might re-associate in a further step.

Can we then make an empirical case for or against any candidate? A first observation is that all three approaches predict the 1- and Q-implicature play equally wel!. **In** particular, since 1- and Q-frames satisfy strong expressibility, the predictions for these cases are exactly the same for all three approaches. The M-frame, on the other hand, does not satisfy strong expressibility, but nevertheless doesn't help judge frameworks, because all of the present candidates mispredict in this case. Take the M-frame as defined above. We then get:

$$\rho_1^{\mathrm{OA}}(f) = \{w,v\} \qquad \rho_1^{\mathrm{OA}}(g) = \{w, v\}$$
$$\rho_1^{\mathrm{BR}}(f) = \{w\} \qquad \rho_1^{\mathrm{BR}}(g) = \{w,v\}$$
$$\rho_1^{\mathrm{OT}}(f) = \{w\} \qquad \rho_1^{\mathrm{OT}}(g) = \{w, v\}$$

The problem is that none of the interpretation rules that we considered handles the long form *9* correctly. Can we fix this problem?

The most obvious idea to try is further iteration. So what would the sender's best response σ_2 be to the receiver's strategy *PI?* The answer to this question now crucially depends on the uninterpretability Assumption 3.2. It is easy to verify that as long as $v \in PI(g)$, the sender's best response will be to send *f* in wand to send *9* in *v.* (Remember that costs are nomina!.) To this, in turn, the receiver's best response is the inverse of the sender strategy. The resulting play is indeed the M-implicature play. This is a noteworthy result in the light of the problem of equilibrium selection: iterated best response starting from a 'truth-only' sender strategy *can* account for 1-, Q- and M-implicatures for *some* versions of the uninterpretability assumption,

but not for others. (To wit, if $Pl(g) = \{w\}$ iteration of best responses has reached a fixed-point different from the M-implicature play).

So is the uninterpretability assumption in 3.2 defensible? **It** does not have to be, since at present it suffices to defend that $Pl(g) \neq \{w\}$, which implies that $v \in Pl(g)$ as desired. And that $Pl(g) \neq \{w\}$ can be argued for based on Cho's and Kreps' (1987) intuitive criterion, as has been demonstrated by van Rooij (2008) (see also the short discussion in Section 2.3). **In** simplified terms, the intuitive criterion gives a strong rationale why the receiver should not believe that a sender in w would send g: she has a message f that, given $Pl(J)$, is always better in w than signal 9 *na matterhaw* the receiver might react to g. (The signal 9 is *equilibr-ium-daminated* for w.) This reasoning establishes that $w \notin Pl(g)$, which gives us the M-implicature play immediately. **If** we adopt a weaker version and only require that $Pl(g) \neq \{w\}$, we can account for M-implicatures aftel' another round of iteration.

4 Conclusion

Taken together, we may say that, with only little help from the refinement literature, the present version of iterated best response provides a uniform, synchronie account of 1-, Q- and M-implicatures. It also subsumes, as a standard game-theoretical model, the optimal assertions approach and strong Bi-OT. This does not discredit either of these latter approaches. For the optimal assertions approach is actually more general than presented here: its predictions were here only assessed for a special case, but the framework is not restricted to a sender who knows the actual world and a receiver who chooses interpretation actions. Similarly, strong optimality is not all there is to Bi-OT: there is also the notion of weak bidirectional optimality which also handles M-implicatures. The conneetion between weak optimality and iterated best response is not obvious and remains an interesting topic of future research. At present, we may safely conclude that, if game-theoretic standards are a criterion for our selection of models of implicature calculation, then iterated best response fares best in the neo-Gricean terrain.

Acknowledgments

Thanks to Anton Benz, Reinhard Blutner, Tikitu de Jager, Gerhard Jäger, Robert van Rooij and an anonymous referee for very helpful comments and discussions.

References

Atlas, J.D. & Levinson, S. (1981). It-clefts, Informativeness, and Logical-Form. **In** Cole, P., ed., *Radical Pmgmatics*, pp. 1-61. Academie Press.

Benz, A. & van Rooij, R. (2007). Optimal assertions and what they implicate. a uniform game theoretic approach. *Topoi, 26(1):63-78.*

Blutner, R. (1998). Lexical Pragmatics. *Joumal of Semantics, 15(2):115-*162.

Blutner, R. (2000). Some Aspects of Optimality in Natural Language Interpretation. *Joumal of Semantics, 17:189-216.*

Cho, I. & Kreps, D.M. (1987). Signaling Games and Stabie Equilibria. *The Quarterly Journal of Economics,* 102(2):179-22l.

Farrell, .J. (1993). Meaning and Credibility in Cheap-Talk Games. *Games and Economie Behaoior,* 5(4):514-53l.

Grice, H.P. (1989). *Studies in the Way of WOTds.* Harvard University Press, Cambridge, Mass.

Hom, L.R. (1984). Towards a new taxonomy for pragmatic inference: Q-based and I-based implicatures. **In** Shiffrin, D., ed., *Meaning, Form, and Use in Context,* pp. 11-42. Georgetown University Press, Washington.

.Jäger, G. (2007). Game dynamics connects semantics and pragmatics. **In** Pietarinen, A.-V., ed., *Game Theoru and Linguistic Meaning,* pp. 89-102. Elsevier.

Levinson, S.c. (2000). *Presumptioe Meanings. The Theoru of Genemlized Conoersational lmplicaiure.* **MIT** Press, Cambridge, Massachusetts.

van Rooij, R. (2008). Games and Quantity Implicature. *Joumal of Economie Methodology.* To appeal'.

A Criterion for the Existence of Pure and Stationary Optimal Strategies In Markov Decision Processes

Hugo Gimbert

Laboratoire Bordelais de Recherche en lnformatique (LABRI)
351, cours de la Libération
33405 Talence Cedex, France
hugo.girnbert@labri.fr

Introduction

Markov decision processes model situations where a controller wishes to control optimally a system, taking her decisions in a sequential way and facing stochastic behaviour of the system. Step aftel' step, the Markov decision process goes through a sequence of states $s_0, s_1, \ldots$ from a set of states S. At each step, the controller chooses an action $a \in A$, which causes the process to change from state s to new state t with fixed probability $p(t|s, a)$. The probability that the decision process stops is 0, i.e., $\sum_{t \in S} p(t|s, a) = 1$ and the time horizon is not bounded hence the decision process never stops. A history is an infinite sequence $h = s_0 a_1 s_1 \ldots$ such that at each step $n \in \mathbb{N}$, the controller has chosen the action a_{n+1}, knowing the sequence $s_0, s_1, \ldots, s_n$ of previous states.

With each history h is associated a payoff $\varphi(h) \in \mathbb{R}$ given to the controller, and the goal of the controller is to maximize the expected value of her payoff.

We are especially interested in these Markov decision processes where the controller can play *optimally* without having to memorize information about the history of the play: in this case the choice of an action a_{n+1} at step n by the controller only depends on the current state s_n of the Markov decision process, the controller is said to play with a *pure and stationary* strategy.

In this paper, we present a criterion about the payoff function which guarantees the existence of *optimal* strategies which are pure and stationary, in any Markov decision process with finitely many state and actions.

Our result still holds in the broader framework of *zero-surn perfect-information stochastic games,* where the controller plays against an adversary which also chooses actions and tries to minimize the expected payoff. This was proven in (Gimbert, 2006a). However, we restriet here to the case

Krzysztof R. Apt, Robert van Rooij (eds.). *New Perspectives on Games and Interaction.* Texts in Logic and Games 4, Amsterdam University Press 2008, pp. 313-328.

one-player games, i.e., Markov decision processes, since this framework is sufheient for giving some interesting applications of our result.

Performances of reactive discrete-event systems can be measured in several ways and similarly Markov decision processes can be equipped with various payoff functions. Some well-known examples are the parity, the mean-payoff or the discounted payoff function, which were initially introduced in the broader context of zero-sum games. **In** Markov decision processes equipped with the *discounted payofJ function,* with each state s is associated a daily payoff r_s and at each step the controller earns the daily payoff corresponding to the current state. Moreover there is an infiation phenomena: future payoffs are multiplied by a discount factor $0 \leq \lambda < 1$, and for a stream $r_{s_0}, r_{s_1}, \ldots$ of daily payoffs, the total payoff received by the controller is:

$$\sum_{n \in \mathbb{N}} \lambda^n r_{s_n}.$$

This payoff was introduced by Shapley (1953). Gilette (1957) considered the case where the controller seeks to maximize the average value of the stream of payoff, i.e.,

$$\liminf_{n \in \mathbb{N}} \frac{r_{s_0} + \ldots + r_{s_n}}{n+1},$$

which defines the *mean-payofJ function.*

Whereas discounted and mean-payoff games are used for economie modelling, *paTity games* were introduced for totally different purposes: they appeared in the context of theoretical computer science as a tooi for studying relations between a logic called the μ-calculus and a class of computation models called tree automata (Emerson and Jutla, 1991; Grädel et al., 2002). The payoff computed by the parity payoff function depends on the set of states visited infinitely often. Other examples of payoff function are the limsup, liminf (Maitra and Sudderth, 1996) and the total (Thuijsman and Vrieze, 1987) payoff functions.

Surprisingly, all these examples of payoff functions share a common nontrivial property. Indeed, in any Markov decision process equipped with one of these functions there exist optimal strategies of a very simple kind: they are at the same time *pure* and *stationarij.* A strategy is pure when the controller plays in a deterministic way and it is stationary when choices of the controller depend only on the current state, and not on the full past history. For the sake of concision, pure stationary strategies are called *positional* strategies, and we say that a payoff function itself is positional if in any Markov decision process equipped with this function, there exists an optimal strategy which is positional.

Existence of positional optimal strategies has a strong algorithmic interest, since it makes the computation of optimal strategies easy. Indeed,

the class of positional strategies is finite, henee a naive algorithm eonsists in enumerating all the positional strategies and seleeting the strategy that gives the highest expeeted payoff. This algorithm is quite ineffieient in praetiee, sinee its running time is at least exponential in the size of the Markov deeision proeess, in the ease where expeeted payoffs are eomputable in polynomial time. However, for the diseounted, mean-payoff and parity funetions, the existenee of positional optimal strategies has been used to design polynomial time algorithms (Puterman, 1994; Filar and Vrieze, 1997).

Content of the paper. This paper addresses the following question: what is the eommon property between the diseounted, mean, limsup and parity payoff whieh explains why all of them are positional? For answering this question, we introduee the class of *submixing* payoff funetions, and we prove that a payoff function whieh is submixing and prefix-independent is also positional (ef. Theorem 2.3). This result partially solves our problem, sinee the parity, limsup and mean-payoff functions are prefix-independent and submixing (cf. Proposition 3.1). Our result has several interesting eonsequenees. First, it unifies and shortens disparate proofs of positionality for the parity, limsup and mean payoff function (Seetion 3). Seeond, it allows us to generate a buneh of new examples of positional payoff functions (Seetion 4).

Plan. This paper is organized as follows. In Seetion 1, we introduee notions of eontrollable Markov ehain, payoff funetion, Markov deeision proeess and optimal strategy. In Seetion 2, we state our main result: prefix-independent and submixing payoff funetions are positional (cf. Theorem 2.3). In the same seetion, we give elements of proof of Theorem 2.3. In Seetion 3, we show that our main result unifies various disparate proofs of positionality. In Seetion 4, we present new examples of positional payoff funetions.

1 Markov decision processes

Let S be a finite set. The set of finite (resp. infinite) sequenees on S is denoted S* (resp. SW). A *probabilits; distriouiior:* on S is a function $\delta : S \to \mathbb{R}$ sueh that $\forall s \in S$, $0 \leq \delta(s) \leq 1$ and $\sum_{s \in \mathbf{S}} \delta(s) = 1$. The set of probability distributions on S is denoted D(S).

1.1 Controllabie Markov chains and strategies

Definition 1.1. A eontrollable Markov ehain $A = (S, A, (A(S))_{sES'P})$ is eomposed of:

- a finite set of states S and a finite set of aetions A,
- for eaeh state $s \in S$, a set $A(s) \subseteq A$ of aetions available in s,
- transition probabilities $p : S \times A \to D(S)$.

When the current state of the chain is *s,* then the controller chooses an available action $a \in A(s)$, and the new state is t with probability $p(t|s,a)$.

A triple $(s,a,t) \in S \times A \times S$ such that $a \in A(s)$ and $p(t|s,a) > 0$ is called a transition.

A *historu* in A is an infinite sequence $h = s_0 a_1 s_1 \cdots \in (SA)^\omega S$ such that for each n, (s_n, a_{n+1}, s_{n+1}) is a transition. State s_0 is called the souree of h. The set of histories with souree s is denoted $\mathbf{P}^\omega_{\mathcal{A},s}$. A *finite historu* in A is a finite prefix of a history. The set of finite histories (resp. of finite histories with souree s) is denoted $\mathbf{P}^*_{\mathcal{A}}$ (resp. $\mathbf{P}^*_{\mathcal{A},s}$).

A *strategy* in A is a function $\sigma : \mathbf{P}^*_{\mathcal{A}} \to D(A)$ such that for any finite history $h \in \mathbf{P}^*_{\mathcal{A}}$ with target $t \in S$, the distribution $\sigma(h)$ puts non-zero probabilities only on actions that are available in t, i.e., $(\sigma(h)(a) > 0) \implies (a \in A(t))$. The set of strategies in A is denoted $\Sigma_{\mathcal{A}}$.

Certain types of strategies are of particular interest for us, these are *pure* and *stationari]* strategies. A strategy is *pure* when the controller plays in a deterministic way, i.e., without using any dice, and it is *staiionars]* when the controller plays without using any memory, i.e., her choices only depend on the cutrent state of the Markov decision process, and not on the entire history of the play. Formally:

Definition 1.2. A strategy $\sigma \in \Sigma_{\mathcal{A}}$ is said to be:

- *pure* if $\forall h \in \mathbf{P}^*_{\mathcal{A}}$, $(\sigma(h)(a) > 0) \implies (\sigma(h)(a) = 1)$,
- *staiionarij* if $\forall h \in \mathbf{P}^*_{\mathcal{A}}$ with target t, $\sigma(h) = \sigma(t)$,
- *positional* if it is pure and stationary.

In the definition of a stationary strategy, $t \in S$ denotes both the target state of the finite history $h \in \mathbf{P}^*_{\mathcal{A}}$ and also the finite history $t \in \mathbf{P}^*_{\mathcal{A},t}$ of length 1.

1.2 Probability distribution induced by a strategy

Suppose that the controller uses some strategy σ and that transitions between states occur according to the transition probabilities specified by $p(\cdot|\cdot,.)$. Then intuitively the finite history $s_0 a_1 \ldots a_n s_n$ occurs with probability

$$\sigma(s_0)(a_1) \cdot p(s_1|s_0, a_1) \cdots \sigma(s_0 \cdots s_{n-1})(a_n) \cdot P(s_n|s_0 \cdots s_{n-1}, a_n).$$

In fact, it is also possible to measure probabilities of sets of infinite histories. For this purpose, we equip $\mathbf{P}^\omega_{\mathcal{A},s}$ with a σ-field and a probability measure. For any finite history $h \in \mathbf{P}^*_{\mathcal{A},s}$, and action a, we define the sets of infinite plays with prefix h or ha:

$$\mathcal{O}_h = \{s_0 a_1 s_1 \cdots \in \mathbf{P}^\omega_{\mathcal{A},s} \mid \exists n \in \mathbb{N}, s_0 a_1 \cdots s_n = h\},$$
$$\mathcal{O}_{ha} = \{s_0 a_1 s_1 \cdots \in \mathbf{P}^\omega_{\mathcal{A},s} \mid \exists n \in \mathbb{N}, s_0 a_1 \cdots s_n a_{n+1} = ha\}.$$

$\mathbf{P}^{\omega}_{\mathcal{A},s}$ is equipped with the cr-field generated by the collection of sets $\mathcal{O}_h$ and $\mathcal{O}_{ha}$. A theorem of Ionescu Tulcea (cf. Bertsekas and Shreve, 1978) implies that there exists a unique probability measure $\mathbb{P}^{\sigma}_{s}$ on $\mathbf{P}^{\omega}_{\mathcal{A},s}$ such that for any finite history $h \in \mathbf{P}^{*}_{\mathcal{A},s}$ with target t, and for every $a \in \mathbf{A}(t)$,

$$\mathbb{P}^{\sigma}_{s}(\mathcal{O}_{ha} \mid \mathcal{O}_h) = \sigma(h)(a),$$
$$\mathbb{P}^{\sigma}_{s}(\mathcal{O}_{har} \mid \mathcal{O}_{ha}) = p(t, a).$$

1.3 Payoff functions

With each history is associated a real value, called the payoff and the controller seeks to maximize this payoff. Payoffs are computed by *payoff functions*, in this subsection we give several examples of such functions: the mean-payoff, the discounted and the parity payoff function.

1.3.1 Mean payoff

The mean-payoff function has been introduced by Gilette (1957). It is used to evaluate average performances of a system. Each transition (s, a, t) is labelled with a *daily payoff* $r(s, a, t) \in \mathbb{R}$. A history $s_0 a_1 s_1 \cdots$ gives rise to a sequence $r_0 r_1 \cdots$ of daily payoffs, where $r_n = r(s_n, a_{n+1}, s_{n+1})$. The controller receives the following payoff:

$$\gamma_{\text{mean}}(r_0 r_1 \cdots) = \limsup_{n \in \mathbb{N}} \frac{1}{n+1} \sum_{i=0}^{n} r_i. \tag{1.1}$$

1.3.2 Discounted payoff

The discounted payoff has been introduced by Shapley (1953) and is used to evaluate short-term performances. Each transition (s, a, t) is labelled not only with a daily payoff $r(s, a, t) \in \mathbb{R}$ but also with a discount factor $0 \leq \lambda(s, a, t) < 1$. The payoff associated with a sequence $(r_0, \lambda_0)(r_1, \lambda_1) \cdots \in (\mathbb{R} \times [0, 1[)^{\mathbb{N}}$ of daily payoffs and discount factors is:

$$\gamma^{\lambda}_{\text{disc}}((r_0, \lambda_0)(r_1, \lambda_1) \cdots) = r_0 + \lambda_0 r_1 + \lambda_0 \lambda_1 r_2 + \cdots . \tag{1.2}$$

The discounted payoff has an intuitive interpretation. Suppose that each time a transition (s, a, t) occurs, a biased coin is tossed to know if the system may halt or proceed. The system proceeds with probability $\lambda(s, a, t)$, and halts with probability $1 - \lambda(s, a, t)$. If the system halts at step n then the payoff is the sum $r_0 + \cdots + r_n$ of rewards seen so far. With that interpretation, $\gamma^{\lambda}_{\text{disc}}$ is exactly the expected sum of daily payoffs before the system halts.

1.3.3 Limsup and liminf payoff

The limsup and liminf payoff functions can be used to measure the peak performances of a system. Let $C \subseteq \mathbb{R}$ be a finite set of real numbers, and

$c_0 c_1 \ldots \in C^\omega$. Then

$$\varphi_{\limsup}(c_0 c_1 \ldots) = \limsup_n c_n$$
$$\varphi_{\liminf}(c_0 c_1 \ldots) = \liminf_n c_n .$$

1.3.4 Parity payoff

The parity payoff function is used to encode temporal logic properties (Grädel et al., 2002). Each transition (s, a, t) is labelled with some priority $c(s, a, t) \in \mathbb{N}$. The controller receives payoff 1 if the highest priority seen infinitely often is odd, and 0 otherwise. For $c_0 c_1 \ldots \in \mathbb{N}^\omega$,

$$\varphi_{\text{par}}(c_0 c_1 \ldots) = \begin{cases} 0 & \text{if } \limsup_n c_n \text{ is even,} \\ 1 & \text{if } \limsup_n c_n \text{ is odd.} \end{cases} \tag{1.3}$$

Remark that since we only consider controllable Markov chains with finitely many states, $\limsup_n c_n$ is always finite, hence the parity payoff function is well-defined.

1.3.5 General payoffs

In the sequel, we will give other examples of payoff functions. Observe that in the examples we gave above, the transitions were labelled with various kinds of data: real numbers for the mean-payoff, couple of real numbers for the discounted payoff and integers for the parity payoff.

We wish to treat those examples in a unified framework. For this reason, we consider in general that a controllable Markov chain A comes together with a finite set of colours C and a mapping col: $S \times A \times S \to C$, which colours transitions.

In the case of the mean payoff, transitions are coloured with real numbers hence $C \subseteq \mathbb{R}$, whereas in the case of the discounted payoff colours are couples $C \subset \mathbb{R} \times [0, 1[$ and for the parity game colours are integers $C \subset \mathbb{N}$.

For a history (resp. a finite history) $h = s_0 a_1 s_1 \ldots$, the colour of the history h is the infinite (resp. finite) sequence of colours

$$col(h) = \text{col}(s_0, a_1, s_1)\, \text{COl}(s_1, a_2, s_2) \cdots$$

Aftel' a history h, the controller receives payoff cp(col(h)), where pay is a payoff function which associates a payoff with each infinite sequence of colours:

Definition 1.3. Let C be a finite set. A payoff function on C is a function $\varphi : C^\omega \to \mathbb{R}$ which is bounded and rneasurable for Lhe cr-field generaled by the sets $\{uC^\omega, u \in C^*\}$.

Boundedness and measurability of payoff functions guarantee that the expected payoff is well-defined.

1.4 Values and optimal strategies in Markov decision processes

Definition 1.4. A Markov decision process is a couple (A, φ), where A is a controllabie Markov chain coloured by a set C and φ is a payoff function on C.

Let us fix a Markov decision process $\mathcal{M} = (A, \varphi)$. Aftel' history h, the controller receives payoff $cp(col(h)) \in \mathbb{R}$. We extend the definition domain of φ to $\mathbf{P}^{\omega}_{\mathcal{A},s}$:

$$\forall h \in \mathbf{P}^{\omega}_{\mathcal{A},s}: \quad \varphi(h) = cp(col(h)).$$

The expected value of φ under the probability $\mathbb{P}^{\sigma}_{s}$ is called the *expected payofj"* of the controller and is denoted $\mathbb{E}^{\sigma}_{s}[\varphi]$. It is well-defined because φ is measurable and bounded. The *value of a state* s is the maximal expected payoff that the controller can get:

$$\mathrm{val(M)(s)} = \sup_{\sigma \in \Sigma_{\mathcal{A}}} \mathbb{E}^{\sigma}_{s}[\varphi].$$

A strategy σ is said to be *optimal* in $\mathcal{M}$ if for any state s ∈ S,

$$\mathbb{E}^{\sigma}_{s}[\varphi] = \mathrm{val(M)(s)}.$$

2 Optima1 positiona1 contro1

We are interested in those payoff functions that ensure the existence of positional optimal strategies. These are defined formally as follows.

Definition 2.1. Let C be a finite set of colours. A payoff function φ on CW is said to be *positional* if for any controllabie Markov chain A coloured by C, there exists a positional optimal strategy in the Markov decision process (A, φ).

Our main result concerns the class of payoff functions with the following properties.

Definition 2.2. A payoff function φ on CW is *prefix-irulependeni* if for any finite word u. ∈ C* and infinite word v ∈ CW, $\varphi(uv) = \varphi(v)$. Payoff function φ is *submixing* if for any sequence of finite non-empty words $u_0, v_0, u_1, v_1, \ldots \in C^*$,

$$\varphi(u_0 v_0 u_1 v_1 \cdots) \leq \max\{\varphi(u_0 u_1 \cdots), \varphi(v_0 v_1 \cdots)\}.$$

The notion of prefix-independence is classica!. The submixing property is close to the notions of *fairly-mixing* payoff functions introduced in (Gimbert and Zielonka, 2004) and of *concave* winning conditions introduced in (Kopczyriski, 2006). We are now ready to state our main result.

Theorem 2.3. Any prefix-independent and submixing payoff function is positional.

The proof of this theorem is based on the 0-1 law and an induction on the number of actions and can be found in (Gimbert, 2006b). We do not repeat this proof here as we prefer to present some of its applications, in the next two sections.

3 Unification of classical results

Thanks to Theorem 2.3, it is possible to give a unified proof of positionality for the limsup, the liminf, the parity and the mean-payoff function. Indeed the positionality of these payoff functions is a simple coroliary of the foliowing proposition.

Proposition 3.1. The payoff functions φ_{lsup}, φ_{linf}, φ_{par} and φ_{mean} are submixing.

Proo]. Consider a finite set of real numbers $C \subseteq \mathbb{R}$ and a sequence of finite non-empty words $u_0, va, u_1, \mathrm{Vl}, \ldots \in C^*$. Let $u. = u_0u_1\cdots \in C^\omega$, $v = vovl \ldots \in C^\omega$ and $w = u_0v_0u_1\mathrm{Vl} \ldots \in C^\omega$. The foliowing elementary fact immediately implies that φ_{lsup} is submixing:

$$\varphi_{\mathrm{lsup}}(w) = \max\{\varphi_{\mathrm{lsup}}(u), \varphi_{\mathrm{lsup}}(v)\}. \tag{3.1}$$

In a similar way, φ_{linf} is submixing since

$$\varphi_{\mathrm{linf}}(w) = \min\{\varphi_{\mathrm{linf}}(u), \varphi_{\mathrm{linf}}(v)\}. \tag{3.2}$$

Now suppose that $C = \{a, \ldots, d\}$ happens to be a finite set of integers and consider function φ_{par}. Remember that $\mathrm{CPpar}(w)$ equals 1 if $\mathrm{cplsup}(W)$ is odd and 0 if $\mathrm{cplsup}(W)$ is even. Then using (3.1) we get that if $\mathrm{CPpar}(w)$ has value 1 then it is the case of either $\varphi_{\mathrm{par}}(u)$ or $\mathrm{CPpar}(v)$. **It** proves that φ_{par} is also submixing.

Finally, we show that function φ_{mean} is submixing. Again $C \subseteq \mathbb{R}$ is a finite set of real numbers. For $i \in \mathbb{N}$ let $c_i \in C$ be the ith letter of word w. Since word w is a shuffle of words $u.$ and v, there exists a partition (Ia, I_1) of $\mathbb{N}$ such that $u. = (c_i)_{i\in Io}$ and $v = (c_i)_{i\in I_1}$. For any $n \in \mathbb{N}$, let $I_0^n = Ia \cap \{a, \ldots, n\}$ and $I_1^n = I_1 \cap \{a, \ldots, n\}$. Then for $n \in \mathbb{N}$,

$$\begin{aligned}\frac{1}{n+1}\sum_{i=0}^{n} c_i &= \frac{|I_0^n|}{n+1}\left(\frac{1}{|I_0^n|}\sum_{i\in I_0^n} c_i\right) + \frac{|I_1^n|}{n+1}\left(\frac{1}{|I_1^n|}\sum_{i\in I_1^n} c_i\right)\\ &\leq \max\left\{\frac{1}{|I_0^n|}\sum_{i\in I_0^n} c_i, \frac{1}{|I_1^n|}\sum_{i\in I_1^n} c_i\right\}.\end{aligned}$$

The inequality holds since $\frac{|I_0^n|}{n+1} + \frac{|I_1^n|}{n+1} = 1$. Taking the superior limit of this inequality, we obtain $CPmean(w) \leq \max\{\varphi_{\mathrm{mean}}(u), \varphi_{\mathrm{mean}}(v)\}$. **It** proves that φ_{mean} is submixing. Q.E.D.

Since φ_{lsup}, φ_{linf}, φ_{par} and φ_{mean} are clearly prefix-independent, Proposition 3.1 and Theorem 2.3 imply that those four payoff functions are positional.

This technique gives a uniform proof of several disparate results, and we compare it to existing proofs.

The case of Markov decision processes equipped with a parity criterion was treated in (Courcoubetis and Yannakakis, 1990). The proof is by inspeetion of strongly connected components whose maximal priority is odd. This proof is far more simple than the result of Gimbert (2006b) used in our argument.

The case of limsup and liminf Markov decision processes was treated in (Maitra and Sudderth, 1996) in the broader framework of stochastie games with infinitely many states. **In** (Maitra and Sudderth, 1996) values Markov decision processes equipped with sup and limsup payoff functions are characterized as fixpoints of some operators. The existence of pure and stationary optimal strategies in the case of finitely many states is derived from this characterization. The proof of Maitra and Sudderth and the proof given here use radically different techniques. The proof of Gimbert (2006b) is shorter since we do not rely on a fine study of values of limsup and liminf games.

For mean-payoff Markov decision process, there basically exists two proofs of the existence of pure and stationary optimal strategies. The first approach, that can be found for example in (Neyman, 2003), consists in proving existence of such strategies in discounted Markov decision processes, and using the fact that values of discounted Markov decision processes are a rational function of discount factors. This implies existence of pure stationary strategies that are optimal for every small values of discount factors, a phenomenon called Blackwell optimality. **In** particular, pure and stationary strategies that are Blackwell optimal are optimal in the limit mean-payoff Markov decision process.

Another approach, in two steps, consists in first considering a weak form of mean-payoff Markov decision processes, where payoffs are computed taking the average value of expectations of rewards rather than the expectation of the average value of rewards (see Puterman, 1994 for example). Using simple matrix calculation, it can be shown that for this weak form of mean-payoff Markov decision processes, there exists pure and stationary optimal strategies. Then one can conclude using a non-trivial result of Bierth (1987) that these strategies are also optimal for (not weak) mean-payoff Markov decision processes.

In our case, we directly prove the existence of pure and stationary optimal strategies, under very weak hypothesis that include as special cases the classes of limsup, liminf, mean-payoff and parity Markov decision processes. The proof of Gimbert (2006b) is more elementary than the proofs above, since it uses only elementary probabilities and measure theory. The theorem of Gimbert (2006b) is also more powerful: in the next section we present several examples of payoff functions that can be proven to be positional thanks to our result, but we do not know if it is possible to do so using existing techniques of Maitra and Sudderth (1996), Puterman (1994), Bierth (1987), and Neyman (2003).

4 New examples of positional payoff functions

In this section we present two new examples of positional payoff functions, namely the weighted payoff and the compromise payoff. We also present three operations on payoff functions, namely mixing with the liminf payoff function, approximation and hierarchical product. These operations have a nice property: the class of submixing and prefix-independent payoff functions is stabie under these operations, hence these operations can be used to generate numerous new examples of positional payoff functions.

4.1 The weighted payoff

Weighted payoff functions were recently introduced by Gimbert and Zielonka (2007). Each transition is labelled with a couple of rewards and weights, the last being a strictly positive real number: the set of colours is $C \subset \{(T,W) \mid r \in \mathbb{R}, w \in \mathbb{R}, w > 0\}$. A history $sOalsl\cdots$ gives rise to a sequence $(Ta, WO)(Tl, w_1)\cdots$ of rewards and weights and the controller receives the payoff:

$$\text{CPweight}((\text{TO}, WO)(Tl, w_1)\cdots) = \limsup_{n\in\mathbb{N}} \frac{1}{\sum_{i=0}^{n} w_i} \sum_{i=0}^{n} w_i \cdot T_i \qquad (4.1)$$

This function generalizes the mean-payoff function: if all weights are taken to be 1, then values of the mean-payoff function and the weighted payoff function do coincide.

Intuitively, weights used in the definition of the weighted payoff functions can be considered as time lengths, and rewards as instant performances. With this interpretation, CPweight computes the average performances over time.

The weighted payoff function is positional. Indeed, it is sub-mixing, the proof is very similar to the proof of Proposition 3.1. Moreover, CPweight is also clearly prefix-independent, hence Theorem 2.3 implies that CPweight is posi tiona1.

Weighted Markov decision processes have strong connections with discounted Markov decision processes with multiple discount factors (Gimbert and Zielonka, 2007), that extend the well-known connections between discounted and mean-payoff Markov decision processes (Gilette, 1957).

4.2 The compromise payoff function

The compromise function was introduced in (Gimbert and Zielonka, 2004), and is defined as follows. We fix a factor $\lambda \in [0,1]$, a finite set $C \subseteq \mathbb{R}$ and for $u. \in C^w$, we define

$$\varphi^{\lambda}_{\mathrm{comp}}(u) = \lambda \cdot \varphi_{\mathrm{lsup}}(u) + (1-\lambda) \cdot \varphi_{\mathrm{linf}}(u). \tag{4.2}$$

This function generalizes the limsup and liminf payoff functions, which correspond to the case where $\lambda = 0$ or $\lambda = 1$.

Intuitively, peak performances of a system can be evaluated using the limsup payoff, whereas its worst performances are computed using the liminf payoff. The *compmmise payofj"* function is used when the controller wants to achieve a trade-off between good peak performances and not too bad worst performances.

4.3 Mixing with the liminf payoff

Not only the limsup function φ_{lsup} but any payoff function φ may be mixed with the liminf function in a way similar. The nice property of this operation is that the submixing property is preserved, as stated in the next proposition.

Proposition 4.1. Let $C < \mathbb{R}$, $0 < \lambda < 1$ and φ be a payoff function on C. Suppose that φ is prefix-independent and submixing. Then the payoff function

$$\lambda \cdot \varphi + (1 - \lambda) \cdot \varphi_{\mathrm{linf}} \tag{4.3}$$

is also prefix-independent and submixing.

Proo]. Let $C \subseteq \mathbb{R}$ be a finite set of real numbers and $u_0, va, u_1, \mathrm{VI}, \in C^*$ be a sequence of finite non-empty words over C. Let $u. = u_0 u_1 \in CW$, $v = vovI \ldots \in CW$ and $w = u_0 v_0 u_1 \mathrm{VI} \ldots \in CW$. Then since φ is submixing, $\varphi(w) \leq \max\{\varphi(u), \varphi(v)\}$ and moreover $\mathrm{CPlinf(W)} = \min\{\varphi_{\mathrm{linf}}(u), \mathrm{CPlinf(V)}\}$, This proves that $\lambda \cdot \varphi + (1 - \lambda) \cdot \varphi_{\mathrm{linf}}$ is submixing. Q.E.D.

In particular, when in (4.3), φ is either $\varphi_{\mathrm{mean}}, \varphi_{\mathrm{par}}$ or φ_{lsup}, we obtain several new examples of positional payoff function.

4.4 Approximating a payoff function

Approximation of a payoff function $\varphi : CW \to \mathbb{R}$ consists in composing φ with a non-decreasing function $f : \mathbb{R} \to \mathbb{R}$. For example, if f is the threshold

function $\mathbf{1}_{\geq 0}$, which associates 0 with strictly negative real numbers and I with positive number, then $f \circ \varphi$ indicates whether φ has positive or negative value.

If φ is positional then of course $f \circ \varphi$ also is, since a positional optimal strategy for some Markov decision process (A, φ) will be optimal for the Markov decision process $(A, f \circ \varphi)$ as weil. **In** fact, it is straightforward to check that approximation not only preserves positionality but also the submixing property.

4.5 The hierarchical product

Now we define a binary operator between payoff functions, which stabilizes the farnily of prefix-independent and submixing payoff functions. We call this operator the *hnerarchical product.*

Let φ_0, φ_1 be two payoff functions on sets of colours Co and C1 respectively. We do not require Co and C1 to be identical nor disjoints.

The hierarchical product $\varphi_0 \rhd \varphi_1$ of φ_0 and φ_1 is a payoff function on the set of colours COUC1 and is defined as follows. Let $u. = \mathrm{COC1} \ldots \mathrm{E} (\mathbf{C}_0 \cup \mathbf{C}_1)^\omega$ and u_0 and u_1 the two projections of $u.$ on Co and C1 respectively. Then

$$(\varphi_0 \rhd \varphi_1)(u) = \begin{cases} \varphi_0(u_0) & \text{if } u_0 \text{ is infinite,} \\ \varphi_1(u_1) & \text{otherwise.} \end{cases}$$

This definition makes sense: although each word u_0 and u_1 can be either finite or infinite, at least one of them must be infinite.

The parity payoff function has an alternative definition in term of the hierarchical product. For $e \in \mathbb{N}$, let Oe and Ie be the payoff functions defined on the one-Ietter alphabet $\{e\}$ and constant equal to 0 and I respectively. Let d be an odd number, and φ_{par} be the parity payoff function on $\{0, \ldots, d\}$. Then

$$\varphi_{\mathrm{par}} = Id \rhd Od\text{-}I \rhd \ldots \rhd 1_1 \rhd 00.$$

The submixing property is stabie under the hierarchical product:

Proposition 4.2. Let φ_0 and φ_1 be two payoff functions. **If** φ_0 and φ_1 are prefix-independent and subrnixing, then $\varphi_0 \rhd \varphi_1$ also is.

Hierarchical products of weighted payoff functions are tightly linked with the discounted payoff function. Indeed, in a Markov decision process equipped with a discounted payoff function, when the discount factors converge to 0, then under weak hypothesis the values converge to the values of the same controllabie Markov chain equipped with a hierarchical product of weighted payoff functions (Gimbert and Zielonka, 2007).

5 Conclusion

We introduced the class of prefix-independent and submixing payoff functions and proved that this class enjoys a nice property: in any Markov decision process equipped with one of these payoff functions, there exists optimal strategies that are pure and stationary. Moreover this class is robust since it contains several payoff functions that are central tools in economics (mean-payoff and limsup functions) and computer science (parity function). Based on these results, we were able to exhibit several new examples of positional payoff functions.

The results of the last section give rise to natural algorithmic questions. For Markov decision processes equipped with mean, limsup, liminf, parityor discounted payoff functions, the existence of optimal positional strategies is the key for designing algorithms that compute values and optimal strategies in polynomial time (Filar and Vrieze, 1997). For examples generated with the mixing operator and the hierarchical product, it seems that values and optimal strategies are computable in exponential time, but we do not know the exact complexity. Also it is not clear how to obtain efficient algorithms when payoff functions are defined using approximation operators.

Another interesting direction for future work is the definition of a good *quantitative* specification language for optimal controller synthesis of *reactive pTogmms,* i.e., programs interacting with their environment. Temporal logies such as CTL* may be used for specifying logical properties about the behaviour of reactive programs, indeed CTL* enjoys nice properties such as closure under boolean operations and computational tractability. When the environment is modelled by stochastic and non-deterministic transitions, computing the minimal probability for a reactive program to satisfy a CTL* specification amounts to solving a Markov decision process equipped with a parity payoff function. The designer of a program may be less interested in the *correctmess* of the program than in its *performances,* for example the expected RAM use. There does not exists a language for specifying quantitative properties of programs with the nice properties of CTL*, although there have already been several steps in this direction (Baier and Clarke, 1998; McIver and Morgan, 2002; de Alfaro, 2003; Chatterejee et al., 2004; de Alfaro et al., 2004; Lluch-Lafuente and Montanari, 2005). Results of the present paper is another effort towards the definition of such a quantitative specification language.

To conclude, we formulate the following conjecture about positional payoff functions: "Any payoff function which is prefix-independent and positional for the class of non-stochastic one-player games is positional for the class of Markov decision processes".

Acknowledgments

The author thanks Wieslaw Zielonka for numerous enlightening discussions about Markov decision processes, and Igor Walukiewicz for his remarks leading to significant improvements of the presentation of this paper.

References

de Alfaro, L. (2003). Quantitative verification and control via the mu-calculus. **In** Amadio, R.M. & Lugiez, D., eds., *CONCUR 2003 - Concurrencsj Theoru, 14th International Conference, Marseille, France, September 3-5, 2003, Proceedinqs,* Vol. 2761 of *Leciure Notes in Computer Science,* pp. 102-126. Springer.

de Alfaro, L., Faella, M., Henzinger, T.A., Majumdar, R. & Stoelinga, M. (2004). Model checking discounted temporal properties. **In** Jensen, K & Podelski, A., eds., *Tools and Alqorithme fOT the Constr-uction and Analysis of Systems, 10th International Conference, TACAS 2004, Held as Pari of the Joint Europeon Conferences on Theoru and Praciice of Softioare, ETAFS 2004, Barcelona, Spain, Moreli* 29 - *APTil* 2, *2004, Proceedinqs,* Vol. 2988 of *Leciure Notes in Computer Science,* pp. 77-92. Springer.

Baier, C. & Clarke, E. (1998). The Algebraic Mu-Calculus and MTBDDs. **In** de Queiroz, R. & Finger, M., eds., *Proceedinqs of the Sth Workshop on Logic, Language, Information and Computation (WoLLIC'98),* pp. 27-38.

Bertsekas, D. & Shreve, S. (1978). *Stochastic Optimal ContTOI: The Discrete-Time Case.* Academie Press.

Bierth, K-.J. (1987). An Expected Average Reward Criterion. *Stochastic Processes and Applications, 26:133-140.*

Chatterejee, K, Jurdziriski, M. & Henzinger, T.A. (2004). Quantitative stochastic parity games. **In** Munro, .l.I., ed., *Proceedinqs of the Fifteenth Annual ACM-SIAM Symposium on Discrete Alqorithme, SODA 2004, New Orleans, Louisiana, USA, Januaru* 11-14, *2004,* pp. 121-130. SIAM.

Courcoubetis, C. & Yannakakis, M. (1990). Markov Decision Processes and Regular Events. **In** Paterson, M., ed., *Automata, Languages and Proqramming, 1'lth Intemational Colloquium, ICALF90, WaTwick Uruoersits), England, July 16-20, 1990, Proceedinqs,* Vol. 443 of *Leciure Notes in Computer Science,* pp. 336-349. Springer.

Emerson, A. & .Jutla, C. (1991). Tree Automata, Mu-Calculus and Determinacy. **In** Sipser, M., ed., *Proceedinqs of 32nd Annual IEEE Symposium on Foundations of Computer Science (FOCS'91),* pp. 368-377.

Filar, .1. & Vrieze, K. (1997). *Competitive Markou Decision Processes.* Springer.

Gilette, D. (1957). Stochastie Games with Zero Stop Probabilities. In *Contrioution to the Theoru of Games, vol. JII,* Vol. 39 of *Annals of Mathematics Studies.* Princeton University Press.

Gimbert, H. (2006a). *Jeux Positionnels.* PhD thesis, Université Denis Diderot, Paris.

Gimbert, H. (2006b). Pure stationary optimal strategies in Markov decision processes. Technical Report 2006--02, LIAFA, Université Paris 7.

Gimbert, H. & Zielonka, W. (2004). When Can You Play Positionally? In Fiala, .1., Koubek, V. & Kratochvil, .1., eds., *Mathematical Foundations of Computer- Science 2004, 29th International Symposium, MFCS 2004, Praque, Czech Republic, August* 22-27, *2004, Proceedinqs,* Vol. 3153 of *Lecture Notes in Computer- Science,* pp. 686-697. Springer.

Gimbert, H. & Zielonka, W. (2007). Limits of multi-discounted Markov decision processes. In *22nd IEEE Symposium on Logic in Computer- Science (LICS 2007), 10-12 July 2007, WTOc!aw, Poland, Proceedinqs,* pp. 89-98. IEEE Computer Society.

Grädel, K, Thomas, W. & Wilke, T. (2002). *Automata, Logics and Infinite Games,* Vol. 2500 of *Leciure Notes in Computer- Science.* Springer.

Kopczyriski, K (2006). Half-Positional Determinacy of Infinite Games. In Bugliesi, M., Preneel, B., Sassone, V. & Wegener, 1., eds., *Automata, Languages and Proqramminq, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedinçs, Part JI,* Vol. 4052 of *Leciure Notes in Computer- Science,* pp. 336-347. Springer.

Lluch-Lafuente, A. & Montanari, U. (2005). Quantitative μ-Calculus and CTL defined over Constraint Semirings. *Theoreiical Computer- Science,* 346(1):135-160.

Maitra, A.P. & Sudderth, W.D. (1996). *Discr-ete Gambling and Stochastic Games.* Springer.

McIver, A.K. & Morgan, e.e. (2002). Games, probability and the quantitative μ-calculus $qM\mu$. In Baaz, M. & Voronkov, A., eds., *Logic [or Proqramming, Ariificiol Intelligence, and Reasoning, 9th International Confer-ence, LPAR 2002, Tbilisi, Georgia, October-* 14-18, *2002, Proceedinqs,* Vol. 2514 of *Leciure Notes in Computer- Science,* pp. 292-310. Springer.

Neyman, A. (2003). From Markov chains to stochastic games. **In** Sorin, S. & Neyman, A., eds., *Stochastic Games and Applications,* pp. 9-25. Klmver Academie Publishers.

Puterman, M.L. (1994). *Markou Decision Processee: Discrete Stochastic Dynamic Proqramminq.* John Wiley & Sons.

Shapley, L.S. (1953). Stochastic games. *Froceedings of the National Academy of Sciences of the United States of America, 39(10):1095-1100.*

Thuijsman, F. & Vrieze, 0 ..1. (1987). The bad match, a total reward stochastic game. *OR Spektrum, 9(2):93-99.*