Players' Beliefs in Extensive Form Games

Oliver Board
Brasenose College, Oxford

21st October 2002 [ie 2003]

Phd.

BODLEIAN LIBRARY OXFORD
DEPOSITED
THESIS
24.VI.03

Players' Beliefs in Extensive Form Games

Oliver Board

ABSTRACT

The epistemic program in game theory uses formal models of interactive reasoning to provide foundations for various game-theoretic solution concepts. Much of this work is based around the (static) Aumann structure model of interactive epistemology, but more recently dynamic models of interactive reasoning have been developed, most notably by Stalnaker (*Economics and Philosophy* 1996) and Battigalli and Siniscalchi (*Journal of Economic Theory* 1999), and used to analyze rational play in extensive form games. But while the properties of Aumann structures are well understood, without a formal language in which belief and belief revision statements can be expressed, it is unclear exactly what are the properties of these dynamic models. In chapter 1, "Dynamic Interactive Epistemology", we investigate this question by defining such a language. A semantics and syntax are presented, with soundness and completeness theorems linking the two.

Chapter 2, "Algorithmic Characterization of Ratioalizability in Extensive Form Games", uses the framework of chapter 1 to construct a dynamic epistemic model for extensive form games, which generates a hierarchy of beliefs for each player over her opponents' strategies and beliefs, and tells us how those beliefs will be revised as the game proceeds. We use the model to analyze the implications of the assumption that the players possess common (true) belief in rationality, thus extending the concept of rationalizability to extensive form games

Chapter 3, "The Equivalence of Bayes and Causal Rationality in Games", takes as its starting point a seminal paper of Aumann (*Econometrica* 1987), which showed how the choices of rational players could be analyzed in a unified state space framework. His innovation was to include the choices of the players in the description of the states, thus abolishing Savage's distinction between acts and consequences. But this simplification comes at a price: Aumann's notion of *Bayes rationality* does not allow players to evaluate what would happen were they to deviate from their actual choices. We show how the addition of a causal structure to the framework enables us to analyze such counterfactual statements, and use it to introduce a notion of *causal rationality*. Under a plausible causal independence condition, the two notions are shown to be equivalent. If we are prepared to accept this condition we can dispense with the causal apparatus and retain Aumann's original framework.

In chapter 4, "The Deception of the Greeks", it is argued that the standard model of an extensive form game rules out an important phenomenon in situations of strategic interaction: deception. Using examples from the world of ancient Greece and from modern-day Wall Street, we show how the model can be generalized to incorporate this phenomenon. Deception takes place when the action observed by a player is different from the action actually taken. The standard model does allow imperfect information (modeled by non-singleton information sets), but not deception: the actual action taken is never ruled out. Our extension of extensive form games relaxes the assumption that the information sets partition the set of nodes, so that the set of nodes considered possible after a certain action is taken might not include the actual node. We discuss the implications of this relaxation, and show that in certain games deception is inconsistent with common knowledge of rationality even along the backward induction path.

Oliver Board, Brasenose College

Thesis submitted for D.Phil. in Economics: Michaelmas Term, 2002 [i.e 2003]

In memory of Professor Michael Bacharach

Chapter 1

Dynamic Interactive Epistemology*

Abstract: The epistemic program in game theory uses formal models of interactive reasoning to provide foundations for various game-theoretic solution concepts. Much of this work is based around the (static) Aumann structure model of interactive epistemology, but more recently dynamic models of interactive reasoning have been developed, most notably by Stalnaker [46] and Battigalli and Siniscalchi [6], and used to analyze rational play in extensive form games. But while the properties of Aumann structures are well understood, without a formal language in which belief and belief revision statements can be expressed, it is unclear exactly what are the properties of these dynamic models. Here we investigate this question by defining such a language. A semantics and syntax are presented, with soundness and completeness theorems linking the two.

1 Introduction

It is well established both theoretically and empirically that strategic reasoning requires agents to form not just conjectures about each other's actions, but also about each other's knowledge and beliefs, which can then be used to infer what actions they might take. In particular, the implications of *common knowledge of rationality*, where all the agents are rational, all know they are all rational, all know that they know, and so on, have been extensively analyzed. More recently, epistemic foundations have been provided for game theoretic solution concepts such as Nash equilibrium (Aumann and Brandenburger [3]). Comprehensive surveys of work in this area are provided by Dekel and Gul [17] and Battigalli and Bonanno [5].

Much of this work is based around the *Aumann structure* model (see Aumann [2]), in which each agent's knowledge is represented by an information partition over a set of states, or possible worlds. For the purposes of the game theorist, however, Aumann structures have several important limitations. First, they describe a very strong concept of knowledge. An implication of modelling agents' epistemic states with information partitions is that everything they know is true, and

*Helpful comments from Michael Bacharach, Paolo Battigalli, Meg Gleason, Mamoru Kaneko, Bob Stalnaker, Johan van Bentham, and Timothy Williamson are gratefully acknowledged.

that they have complete introspective access to this knowledge, i.e. they know everything they know (positive introspection), and they know everything they don't know (negative introspection). Negative introspection in particular has widely been considered inappropriate when applied to knowledge. More generally, it has been thought important to analyze agents' beliefs as well as their knowledge. And beliefs, unlike knowledge, can be false. These issues can be dealt with be replacing the information partitions with possibility correspondences (see e.g. Samet [43]). Beliefs modelled by possibility correspondences at their most general do not satisfy any of the properties described above. By imposing certain restrictions on the correspondences we can recover these properties one by one. In the extreme, if we assume that each correspondence partitions the state space we are back where we started.

The second problem with using Aumann structures to model rational play in games is that they are essentially static: the epistemic states that they model are fixed, while in dynamic games[1] agents have a chance to change their beliefs as the game progresses. In particular, conjectures about what strategies one's opponents might be playing can be revised as moves are observed. A stark illustration of the importance of such revisions is given by Reny [42], who shows that once the possibility of belief change is taken into account, the game-theoretic wisdom that common knowledge of rationality implies backward induction in games of perfect information is undermined. As long as the information that an agent learns is consistent with what she already knew or believed, this problem can be handled in the existing framework. The agent's partition (or possibility correspondence) can be refined, in a manner analogous to Bayesian updating of probabilities, to take account of the new information. But, like Bayes rule, this process is not well defined when the information learned is incompatible with the agent's previous beliefs, i.e. she is *surprised*. And modelling the response to such surprises is crucial: to evaluate the rationality of strategies in a dynamic game, we must have a theory about what the players would believe at *every* node in the game, even though some of these nodes will typically be ruled out by the players on the basis of the information they possess at the beginning of the game.

Models of dynamic interactive reasoning have thus been developed. Stalnaker [46] replaces the information partitions of the Aumann structure with *plausibility orderings* on the set of possible worlds, which encode information not just about each agent's current beliefs, but also about how these beliefs will be revised as new information is learned, even if this new information is a surprise

[1] i.e. games in which there is a flow of information as the game proceeds. These games are commonly represented by the extensive form.

(e.g. it takes the form of an unexpected move made by one's opponent). This seems to be a satisfactory resolution to the problem, and models of this kind have been used by Stalnaker and others to analyze rational play in dynamic games.

From a philosophical point of view, however, there is something unsatisfactory about the Aumann structure model and all its extensions, as identified by Aumann [4] himself: "... the whole idea of 'state of the world,' and of a partition structure that reflects the players' knowledge about the other players' knowledge, is not transparent. What *are* the states? Can they be explicitly described? Where do they come from?" (p. 264). Fagin *et al.* [22] elaborate further: "If we think of a state as a complete description of the world, then it must capture all of the agents' knowledge. Since the agents' knowledge is defined in terms of the partitions, the state must include a description of the partitions. This seems to lead to circularity, since the partitions are defined over the states, but the states contain a description of the partitions" (p. 332).

Economists have developed an alternative model of interactive beliefs which seems to avoid this circularity. The hierarchical approach (Mertens and Zamir [37], Brandenburger and Dekel [13]) takes as its starting point a set of *states of nature*, which describe facts of interest about the physical world, such as which strategy profile will be played. Each agent's beliefs about the state of nature is represented by a probability distribution over the set of states of nature; their beliefs about these beliefs are then represented by a probability distribution over these distributions and the set of states of nature; and so on. In this way, we build up an infinite hierarchy of beliefs for each player, called her *type* (after Harsanyi [30]). In contrast to the Aumann structure approach, where the infinite hierarchy of beliefs is generated implicitly by partitions over obscure states of the world, here it is explicitly constructed from levels of probability distributions over clearly defined states of nature.

The question remains, however, as to whether a state of nature together with a description of each agent's type provides a satisfactory description of a state of the world. For it is not clear that an agent's type gives a complete description of her beliefs. Her type specifies what she believes about all the finite-level beliefs of her opponents, but does it actually describe what she believes about their types, what she believes about what they believe about her type, and so on? It turns out that as long as the types satisfy certain coherency conditions, we can answer this question in the affirmative. These coherency conditions amount to assuming that the agents satisfy positive and negative introspection, and guarantee that the belief hierarchies are closed.

Furthermore, the hierarchical model can be extended to deal with the problem of belief revision.

Battigalli and Siniscalchi [6] have shown how to construct hierarchies of *conditional probability systems*; the level-0 probability systems describe each agent's (probabilistic) beliefs about the physical world as before, but they also encode information about how these beliefs are revised. The level-1 systems represent the agents' beliefs over these level-0 systems, and so on. Again, as long as the appropriate coherency conditions are satisfied, these hierarchies are closed and each agent's type describes all of her beliefs.

Any extra clarity these hierarchical constructions might bring, however, is paid for at a price of greatly-increased complexity. The complexity of these models may well be self defeating: Aumann [4] describes them as "cumbersome and far from transparent... In fact, the hierarchy construction is so convoluted that we present it here with some diffidence" (pp. 265, 295). In addition, two more specific problems arise. The first concerns the coherency conditions that are required for closure of the hierarchies. As we have already discussed, it may not always be appropriate to assume that agents' have complete introspective access to their epistemic states; this remains true even if we are dealing with belief rather than knowledge. In the case of conditional probability systems, the coherency assumption becomes even stronger: here it is assumed that agents have complete introspective access to their belief revision schemes as well. Ideally we would like to have a system that is flexible enough to work with or without positive and negative introspection. The second problem arises when we consider the non-probabilistic analogue of these belief hierarchies, where each level in the hierarchy describes simply which members of the previous level the agent considers possible, rather than assigning probabilities to each (the former is not generally derivable from the latter: a world may be considered possible even if it is assigned zero probability). In this case it turns out that, even with the appropriate coherency conditions, the infinite hierarchy does not in general provide a complete description of an agent's uncertainty; that is, it does not tell us which types of her opponents she considers possible (Fagin [19], Heifetz and Samet [31], Brandenburger and Keisler [14]).

Thankfully there is a path between this Scylla and Charybdis, between the obscurity of Aumann structures and the complexity of belief hierarchies. *Epistemic logic* (surveyed briefly in Appendix A) is based on a formal language which can express statements about the world and what agents believe about the world and about each other. The language is built up from a set of primitive formulas by means of an inductive rule. The primitive formulas and each step of the inductive process are entirely transparent. Hintikka [32] showed how *Kripke structures* (Kripke [34]) can be used be provide a *semantics* for this language, i.e. a set of rules for determining the truth or falsity

of every sentence or *formula* in the language. Hence there is no issue about whether or not these structures provide a complete description of the agents' uncertainty: the language itself defines the limits of what we can and cannot say about the agents' beliefs.

There is a very close connection between Kripke structures and Aumann structures: the former are a general version of the latter, where the information partitions are replaced by possibility correspondences (traditionally referred to as *accessibility relations*), plus the addition of an *interpretation* which assigns truth values to the primitive formulas. Kripke structures are general enough to model knowledge or belief, with or without the introspection assumptions. Certain properties of Kripke structures correspond to various axioms and rules governing the behavior of formulas in the language: these axioms and rules, jointly referred to as an *axiom system*, give us a precise characterization of sets of formulas which are true in different types of Kripke structure, and hence an elucidation of the particular concept of knowledge or belief that is being modelled. The axiom system and language form a *syntax* for the logic.

There is however a gap still to be filled. In order to extend the results just described to structures such as Stalnaker's, we must develop a language that is richer than that of epistemic logic. In section 2 of this paper, we define such a language by adding *revised belief* operators to the standard language. Thus, if $B_i\phi$ is a formula of the language, then so is $B_i^{\phi}\psi$, to be interpreted "i believes that ψ on learning that ϕ". We then present a semantics for this language consisting of *belief revision structures*, which look much like a generalized version of Stalnaker's structures. A theorem links these structures to an axiom system which describes how these revised belief operators, and the rest of the language, behave. This axiom system is essentially the most basic axiom system of epistemic logic augmented by additional axioms and rules that correspond to some of the AGM axioms of belief revision (Alchourrón *et al.* [1]). A brief account of the AGM axioms, which form the basis of a well-established (single-agent) theory of belief revision, is given in appendix B. Several extensions to the model, including the introduction of introspection and consistency axioms, and common belief operators, are developed in section 3. Section 4 comments on some issues which are not treated by our formalism, section 5 discusses related literature, and section 6 concludes.

2 Dynamic interactive epistemology

As discussed in the introduction, an important distinction is made in logic between a syntax and a semantics. A syntax consists of a formal language, defined by a set of formulas, and a proof procedure for generating theorems in that language. The proof procedure, usually expressed in the form of an axiom system, is often rather cumbersome: even basic theorems can be very tricky to prove. A semantics is made out of *structures* that give truth conditions for every formula in the language. A structure is a well-defined mathematical object, and usually very easy to work with, but hard to interpret. The task of the logician is the establish a connection between the syntax and the semantics. This can be done by means of *soundness* and *completeness* theorems, which link the theorems generated by the proof procedure with the truth conditions established by the structures. We start by describing the language we shall work with.

2.1 Language

Our language $\mathcal{L}_n(\Phi)$ is built up from a nonempty set Φ of primitive formulas and an inductive rule. The primitive formulas stand for statements expressing basic facts about the world, such as "agent i plays strategy s_i". The inductive rule enables us to build up more complex formulas standing for statements such as "agent i plays strategy s_i and agent j plays strategy s_j", and "agent j believes that agent i plays strategy s_j". Formally, $\mathcal{L}_n(\Phi)$ is defined as the smallest set which satisfies the following conditions:

(*a*) if $\phi \in \Phi$, then $\phi \in \mathcal{L}_n(\Phi)$;

(*b*) if $\phi, \psi \in \mathcal{L}_n(\Phi)$, then $\neg\phi \in \mathcal{L}_n(\Phi)$ and $(\phi \wedge \psi) \in \mathcal{L}_n(\Phi)$;

(*c*) if $\phi, \psi \in \mathcal{L}_n(\Phi)$, then $B_i\phi \in \mathcal{L}_n(\Phi)$ and $B_i^{\phi}\psi \in \mathcal{L}_n(\Phi)$ for $i = 1, \ldots, n$.

For economy of notation, we take Φ and n to be fixed henceforth and omit them from the notation. We also omit parentheses whenever there is no risk of confusion, and use the following standard abbreviations: $\phi \vee \psi$ for $\neg(\neg\phi \wedge \neg\psi)$; $\phi \Rightarrow \psi$ for $\neg\phi \vee \psi$; and $\phi \Leftrightarrow \psi$ for $(\phi \Rightarrow \psi) \wedge (\psi \Rightarrow \phi)$. As discussed in the introduction, $\mathcal{L}$ is the language of epistemic logic augmented by adding modal operators B_i^{ϕ} that tell us what the agents believe after receiving the information that ϕ. Notice that the language cannot express iterated belief revisions; that is, there are no formulas expressing statements such as "agent i believes that χ on learning that ϕ and then learning that ψ". We comment on this restriction in section 4.2 below.

We now present an axiom system and semantics for $\mathcal{L}$.

2.2 Axiom system

An axiom system AX consists of a set of axioms and inference rules. An axiom is simply a formula or set of formulas, and an inference rule allows us to infer one formula from a set of other formulas. A *proof* in AX is a finite sequence of formulas, each of which is either an (instance of) an axiom or follows from some of the preceding formulas by applying an inference rule. A *proof of* ϕ is a proof whose last formula is ϕ. We say that ϕ is provable in AX (or ϕ is a *theorem* of AX), and write $AX \vdash \phi$, if there is a proof of ϕ in AX.

We shall consider the axiom system BRS for $\mathcal{L}$, consisting of the following axioms and inference rules:

Taut	$true$
Dist	$\left(B_i^{\phi}\psi \wedge B_i^{\phi}(\psi \Rightarrow \chi)\right) \Rightarrow B_i^{\phi}\chi$
Triv	$B_i\phi \Leftrightarrow B_i^{true}\phi$;
Succ	$B_i^{\phi}\phi$
IE(a)	$B_i^{\phi}\psi \Rightarrow \left(B_i^{\phi\wedge\psi}\chi \Leftrightarrow B_i^{\phi}\chi\right)$
IE(b)	$\neg B_i^{\phi}\neg\psi \Rightarrow \left(B_i^{\phi\wedge\psi}\chi \Leftrightarrow \left(B_i^{\phi}\chi \vee B_i^{\phi}(\psi \Rightarrow \chi)\right)\right)$
MP	from ϕ and $\phi \Rightarrow \psi$ infer ψ
RE	from ψ infer $B_i^{\phi}\psi$
LE	from $\phi \Leftrightarrow \psi$ infer $B_i^{\phi}\chi \Leftrightarrow B_i^{\psi}\chi$

Note that: any formulas in $\mathcal{L}$ may be substituted for ϕ, ψ, χ; $i \in \{1, \ldots, n\}$; $true$ stands for any propositional tautology, and $false$ stands for $\neg true$. This system is close to the system K of epistemic logic with extra axioms and rules, corresponding roughly to the AGM axioms of belief revision, to describe the behavior of the revised belief operators B_i^{ϕ}.

Taut, Dist (the *distribution axiom*), **MP** (*modus ponens*), and **RE** (the *rule of epistemization*) are familiar from epistemic logic, and need no further comment (but note that **Dist** and **RE** apply only to the *revised* belief operators). Jointly, these correspond to AGM axiom $(K * 1)$[2]. **Triv**, the *triviality axiom*, says that if the information received by an agent is trivial (i.e. if it is a propositional tautology), then she does not revise her beliefs; a corresponding condition is implied by the AGM axioms. This also ensures that ordinary beliefs satisfies the same properties as revised beliefs. **Succ**,

[2] Under the numbering system of Gärdenfors [25], reproduced in appendix B below.

the analogue of $(K*2)$, is the *success axiom*, which guarantees that the information received is indeed believed in the revised belief state. **IE(a)** and **IE(b)** are the axioms of *informational economy*, motivated by the *criterion of informational economy*: our beliefs are not in general gratuitous, and so when we change them in response to new evidence, the change should be no greater than is necessary to incorporate that new evidence. More specifically, **IE(a)** says that if an agent learns something she already knew, she doesn't revise her beliefs at all; and **IE(b)** says that if she learns something consistent with her original beliefs, then her revised beliefs are formed simply by adding the new information to her existing stock of beliefs and closing under *modus ponens*. **IE(b)** corresponds directly to $(K*7)$ and $(K*8)$, and, in the presence of **Triv**, also to $(K*3)$ and $(K*4)$; **IE(a)** is implied by $(K*7)$ and $(K*8)$ in the presence of $(K*5)$. Finally, **LE**, the *rule of logical equivalence*, corresponds to $(K*6)$, and says that logically equivalent formulas should lead to identical belief revisions: it only the *content* of the information and not the way it is expressed that determines how beliefs are revised[3].

2.3 Semantics

The semantics for $\mathcal{L}$ is provided by a *belief revision structure*. This is based on a combination of Grove's [27] *spheres* model and the Kripke structure framework. The Kripkean accessibility relations are replaced by *plausibility orderings* at every world for each agent, with the most plausible worlds for a given agent at a particular world taking the role of the accessible worlds for that agent at that world. But a plausibility ordering for an agent tells us not only her current epistemic state, it also encodes information about her belief revision policy. In turn, the structure generates the other agents' beliefs about this belief revision policy, and so on, thus providing truth conditions for each formula of $\mathcal{L}$. Formally, a *belief revision structure* M over Φ for n agents is a ordered triple $\langle W, \pi, \preccurlyeq \rangle$, where: W is a set of possible worlds; $\pi : W \times \Phi \rightarrow \{\mathbf{true}, \mathbf{false}\}$ is an interpretation; and $\preccurlyeq$ is a vector of binary relations over W, giving the plausibility ordering of each agent at each world. We use $\preccurlyeq_i^w$ to denote the plausibility ordering of agent i at world w. Intuitively, $x \preccurlyeq_i^w y$ means "from the point of view of agent i at world w, world x is at least as plausible as world y".

Belief revision structures are used to give truth conditions to formulas. Formally, truth of formulas is characterized by the $\models$ relation: $(M, w) \models \phi$ means that ϕ *is true at world* w *in structure* M. We use $[\phi]_M$ to denote the set of worlds in which ϕ is true (the *truth set* of ϕ), i.e.

[3] This rules out what psychologists call *framing effects*.

$[\phi]_M = \{w \mid (M,w) \models \phi\}$. If ϕ is true at every world of a given structure, we say that ϕ *is valid in* M, and write $M \models \phi$. Finally, for a given class of structures $\mathcal{C}$, we say that ϕ *is valid with respect to* $\mathcal{C}$, and write $\mathcal{C} \models \phi$, if $M \models \phi$ for all $M \in \mathcal{C}$.

Before giving the formal definition of $\models$, we impose several restrictions on the form of belief revision structures. Define $W_i^w = \{x \mid x \preccurlyeq_i^w y \text{ for some } y\}$, the set of worlds which are *conceivable* to agent i at world w, though not necessarily accessible. Then, we assume that:

R1 for all i, w: $\preccurlyeq_i^w$ is complete and transitive on W_i^w;

R2 for all i, w: $\preccurlyeq_i^w$ is well-founded.

R1 ensures that each plausibility ordering divides all the worlds into ordered equivalence classes; the inconceivable worlds, i.e. those not in W_i^w, are a class unto themselves and are to be considered least plausible. If $\preccurlyeq_i^w$ is *well-founded* (**R2**), then there are no infinitely descending sequences of the form $\ldots w_n \prec_i^w \prec w_{n-1} \prec_i^w \ldots \prec_i^w w_0$ (where $x \prec_i^w y$ if and only if $x \preccurlyeq_i^w y$ and not $y \preccurlyeq_i^w x$). This guarantees that for every set $X \subseteq W$, if $X \cap W_i^w \neq \varnothing$, then $\min_i^w \{X \cap W_i^w\} \neq \emptyset$, where $\min_i^w$ is defined in the usual way (i.e. $\min_i^w (X) = \{x \in X \mid \text{for all } y \in X,\ x \preccurlyeq_i^w y\ \}$); intuitively, it says that if there are any conceivable worlds in a certain set, then there is a most plausible world in that set[4]. Well-foundedness is satisfied automatically in the case where W is finite. We call a belief revision structure *ordered* if it satisfies **R1**, and *focused* if it satisfies **R2**. Let $\mathcal{M}$ denote the class of all belief revision structures that are ordered and focused.

We are now in a position to define $\models$. The definition proceeds by induction on the form of ϕ.

$(M,w) \models \phi$ (for $\phi \in \Phi$) iff $\pi(w)(\phi) = \mathbf{true}$;

$(M,w) \models \phi \wedge \psi$ iff $(M,w) \models \phi$ and $(M,w) \models \psi$;

$(M,w) \models \neg\phi$ iff not $(M,w) \models \phi$;

$(M,w) \models B_i\phi$ iff $(M,x) \models \phi$ for all $x \in \min_i^w (W_i^w)$;

$(M,w) \models B_i^\phi\psi$ iff $(M,x) \models \psi$ for all $x \in \min_i^w \{[\phi]_M \cap W_i^w\}$.

[4] Well-foundedness of $\preccurlyeq_i^w$ is stronger than the limit assumption assumption proposed by Lewis [35]: if $[\phi]_M \cap W_i^w \neq \varnothing$, then $\min_i^w \{[\phi]_M \cap W_i^w\} \neq \emptyset$. Well-foundedness requires that *every* set which has a nonempty intersection with W_i^w has a least element, while the limit assumption applies this condition only to sets which represent formulas. We impose the stronger condition in order to preserve the clean cut between extra-linguistic reality (as represented by the *frame* (W and $\preccurlyeq$)) and the semantics (which maps the language into the frame).

The first three rules are straightforward. The fourth rule gives truth conditions for formulas of the form $B_i\phi$ in much the same way as the Kripke semantics, with the most plausible worlds playing the role of the accessible worlds: agent i believes that ϕ if and only if ϕ is true in all the most plausible worlds. The fifth rule operates similarly: the worlds accessible to the agent when she learns that ϕ are precisely the most plausible worlds that are consistent with ϕ; thus agent i believes that ψ on learning that ϕ if and only if ψ is true in all the most plausible worlds in which ϕ is true[5]. The five rules provide truth conditions for every formula in $\mathcal{L}$.

2.4 Soundness and completeness

Before stating the main theorem of the paper, we need to introduce some more terminology.

An axiom system AX is said to be *sound* for a language $\mathcal{L}$ with respect to a class $\mathcal{C}$ of structures if every formula in $\mathcal{L}$ that is provable in AX is valid with respect to $\mathcal{C}$. The system AX is said to be *complete* for $\mathcal{L}$ with respect to $\mathcal{C}$ if every formula in $\mathcal{L}$ that is valid with respect to $\mathcal{C}$ is provable in AX. We can think of AX as characterizing the class $\mathcal{C}$ if it provides a sounds and complete axiomatization of that class, i.e. for all $\phi \in \mathcal{L}$, we have $AX \vdash \phi$ if and only if $\mathcal{C} \models \phi$. Soundness and completeness provide a tight connection between the syntactic notion of provability, which is hard to use but easy to understand, and the semantic notion of validity, which is easy to use but hard to understand.

It turns out that a precise connection can be made between the axiom system BRS and belief revision structures. The following theorem tells us that every theorem ϕ that is provable in BRS is also true in every world of every belief revision structure:

Theorem 1 *BRS is a sound and complete axiomatization w.r.t. $\mathcal{M}$ for formulas in $\mathcal{L}$.*

The proof of this and all other theorems is given in appendix C.

2.5 The canonical structure

Before moving on to discuss various extensions of the logic presented in this section, we show how to construct a particularly important belief revision structure M^c, called the *canonical structure for BRS*. To understand what the canonical structure is, we need some more definitions.

[5] The purpose of the well-foundedness condition should now be clear: if it does not hold, $B_i^{\phi}\psi$ could be (vacuously) true even though ψ was not true in any sufficiently plausible ϕ-world, because there might be no *most* plausible ϕ-world. Thus we would clearly have the wrong truth conditions for sentences of this form, and in fact for sentences of the form $B_i\phi$.

For a given axiom system AX, we say that a formula ϕ is *AX-consistent* if $\neg\phi$ is not provable in AX. A finite set of formulas $\{\phi_1, \ldots, \phi_k\}$ is AX-consistent exactly if $\phi_1 \wedge \ldots \wedge \phi_k$ is AX-consistent, and an infinite set of formulas is AX-consistent exactly if all its finite subsets are AX-consistent. Finally, at set of formulas $S \subseteq \mathcal{L}$, is a *maximal AX-consistent set* if (a) it is AX-consistent, and (b) for all ϕ in $\mathcal{L}$ but not in S, the set $S \cup \{\phi\}$ is not AX-consistent.

The canonical structure has a world w_S corresponding to every maximal BRS-consistent set S. This structure is analogous to the *universal* type space construction of Battigalli and Siniscalchi [6], who extend the work of Mertens and Zamir [37] and Brandenburger and Dekel [13] to the dynamic setting. In both cases every *allowable* epistemic type is represented: in the canonical structure by sets of formulas describing each agent's beliefs and how these beliefs will be revised; and in the universal type space an infinite hierarchy of conditional probability systems for each agent. Both approaches rule out certain beliefs: according to the former, an epistemic type is allowable only if the formulas describing it are logically consistent according to the axiom system; in the hierarchical construction of Battigalli and Siniscalchi, the representation of beliefs by conditional probability systems allows only beliefs that satisfy an appropriate set of probability axioms, and additional *coherency* conditions are imposed on the hierarchies to ensure that the various levels of each hierarchy agree with each other.

There are, however, important differences between the two approaches. While the universal type space of Battigalli and Siniscalchi describes the probabilistic beliefs of each agent, the canonical structure presented here tells us what the agents consider possible. It is clear that probabilistic beliefs cannot be recovered from the canonical structure. Nor can possibility correspondences be recovered from the universal type space, unless possibility is identified with strictly positive probability. In addition, the conditional probability systems used by Battigalli and Siniscalchi specify beliefs conditional on *observable* events only; in our terminology, this means that information can take the form of propositional formulas only (i.e. primitive formulas and their conjunctions and negations), which describe the physical world[6]. Our language places no restrictions on the kind of information that may be received; in particular, the possibility that one agent may learn another's beliefs is not ruled out.

For the construction of M^c, we introduce some new notation: let $S/B_i^\phi = \left\{\psi \mid B_i^\phi \psi \in S\right\}$, i.e. S/B_i^ϕ is the set of formulas believed by i when she learns that ϕ. Let $M^c = \langle W, \pi, \preccurlyeq \rangle$, where

[6] A similar restriction is imposed by Friedman and Halpern [23] on their logic of belief change. See section 5 for a more detailed discussion of this work.

$$
\begin{aligned}
W &= \{w_S : S \text{ is a maximal } BRS\text{-consistent set}\} \\
\pi(w_S)(\phi) &= \begin{cases} \textbf{true} \text{ if } \phi \in S \\ \textbf{false} \text{ if } \phi \notin S \end{cases} \quad \text{for all } \phi \in \Phi \\
w_T &\preccurlyeq {}^{w_S}_{i} w_U \text{ if there is some } \phi \in T \cap U \text{ such that } S/B_i^{\phi} \subseteq T
\end{aligned}
$$

To show that each world w_S really does correspond to the set S of formulas, we must prove the following proposition:

Proposition 1 *$(M^c, w_S) \models \phi$ if and only if $\phi \in S$.*

Proposition 1 says that S contains exactly those formulas which are true at w_S. The proof is given in appendix C. Another soundness and completeness theorem emerges as a corollary of this proposition.

Corollary 1 *BRS is a sound and complete axiomatization w.r.t. M^c for formulas in $\mathcal{L}$.*

For (soundness) if ϕ is provable in BRS, it must be contained in every maximal BRS-consistent set (see proof of Theorem 1), and by Proposition 1, it is therefore valid with respect to M^c. And (completeness) if ϕ is valid with respect to M^c, Proposition 1 tells us that it must be contained in every maximal BRS-consistent set. If follows that ϕ is provable in BRS: if not, $\neg\phi$ would be BRS-consistent and thus contained in *some* maximal BRS-consistent set (see again proof of Theorem 1); but $\{\phi, \neg\phi\}$ is not BRS-consistent, and so ϕ and $\neg\phi$ cannot be contained in the same maximal BRS-consistent set.

Although it might seem that the soundness part of Corollary 1 follows from Theorem 1, and that the completeness part of Theorem 1 follows from Corollary 1, this is not the case because $M^c \notin \mathcal{M}$. The reason is that some of the $\preccurlyeq_i^{w_S}$ relations are not well-founded (though they are complete and transitive on $W_i^{w_S}$ in each case). Construct a sequence of maximal BRS-consistent

sets containing the following formulas:

T_1	T_2	T_3	T_4	$\cdots$	T_∞
$\neg\phi$	ϕ	ϕ	ϕ		ϕ
$\neg B_i\phi$	$\neg B_i\phi$	$B_i\phi$	$B_i\phi$		$B_i\phi$
$\neg B_iB_i\phi$	$\neg B_iB_i\phi$	$\neg B_iB_i\phi$	$B_iB_i\phi$		$B_iB_i\phi$
$\neg B_iB_iB_i\phi$	$\neg B_iB_iB_i\phi$	$\neg B_iB_iB_i\phi$	$\neg B_iB_iB_i\phi$		$B_iB_iB_i\phi$
$\vdots$					

,

and a maximal BRS-consistent set S such that $S/B_i^{\neg\phi} = T_1$, $S/B_i^{\neg B_i\phi} = T_2$, $S/B_i^{\neg B_iB_i\phi} = T_3, \ldots$, $S/B_i = T_\infty$. Then it follows from Proposition 1 and the definition of $\models$ that $\ldots w_{T_3} \prec_i^{w_S} w_{T_2} \prec_i^{w_S} w_{T_1}$, i.e. we have an infinitely descending sequence and $\preccurlyeq_i^{w_S}$ is not well-founded. Nonetheless Corollary 1 tells us that the tight connection between valid formulas and formulas that are provable in BRS still holds.

The canonical structure is useful for certain game-theoretic applications. Both forward and backward induction are based on the premise that players try to interpret their opponents' strategy choices as rational whenever possible. But what it is rational for a player to do depends on her beliefs. Using a structure that rules out certain beliefs to analyze a game restricts the set of available explanations for a particular action. So if we are interested in which strategies are compatible with rationality and which are not we must work with a structure that includes all possible beliefs. The canonical structure does just this. See Battigalli and Siniscalchi [7] and Board [9] for further elaboration of this point.

We finish this section with a brief comment on the impossibility results of Fagin [19], Heifetz and Samet [31], Brandenburger and Keisler [14] and others, which show that if epistemic types are represented by possibility sets (as they are here) rather than probability distributions, then a structure containing all epistemic types cannot exist. These results would seem to contradict our claim that the canonical structure does contain a representation of every epistemic type. Brandenburger [12] explains how the two can be reconciled: "...completeness is impossible if literally all possibility sets are wanted. But if we make topological assumptions that serve to rule out certain kinds of possibility sets, then a (restrictedly) complete structure may exist" (p. 4). Working with a formal language has precisely this effect. It is formulas of this language and not arbitrary sets of worlds that are the content of beliefs (and of information), and in any given model there may be

sets of worlds that do not represent any formula of the language.

3 Extensions

3.1 Introspection

Consider the following additional axioms:

TPI $\quad B_i^\phi \psi \Rightarrow B_i^\chi B_i^\phi \psi$

TNI $\quad \neg B_i^\phi \psi \Rightarrow B_i^\chi \neg B_i^\phi \psi$

TPI and **TNI** are the axioms of *total positive introspection* and *total negative introspection*, and state that agents have complete introspective access to their own minds, including not only their current beliefs but also how these beliefs would be or would have been revised. Let $BRSI$ be the axiom system formed by the addition of **TPI** and **TNI** to BRS. To illustrate the strength of these axioms, we consider three implications. The first is introspection of current beliefs: $B_i^\phi \psi \Rightarrow B_i^\phi B_i^\phi \psi$ and $\neg B_i^\phi \psi \Rightarrow B_i^\phi \neg B_i^\phi \psi$. The knowledge analogues of these principles emerge as properties of the Aumann structure model discussed in the introduction and widely used in economic theory, but their universal applicability has been questioned by Geanakoplos [26], among others. Second, **TPI** and **TNI** imply that agents have correct beliefs about their future beliefs, whatever information they receive: $B_i^\phi \psi \Rightarrow B_i B_i^\phi \psi$ and $\neg B_i^\phi \psi \Rightarrow B_i \neg B_i^\phi \psi$. Finally, it is implied that agents can recall their prior beliefs: $B_i \psi \Rightarrow B_i^\phi B_i \psi$ and $\neg B_i \psi \Rightarrow B_i^\phi B_i \psi$. This assumption is inappropriate in certain games and decision problems, such as the absent-minded driver paradox of Piccione and Rubinstein [40]. Bonanno [11] provides a careful analysis of this and other memory axioms in the context of extensive form games.

Imposing an additional restriction on the form of the $\preccurlyeq_i^w$ relations provides a semantic characterization of **TPI** and **TNI**:

R3 for all i, w, x, y, z: if $x \in W_i^w$, then $y \preccurlyeq_i^x z$ if and only if $y \preccurlyeq_i^w z$

Intuitively, **R3** says an agent has the same plausibility ordering in every world that is conceivable to her. If a belief revision structure satisfies **R3** we call it *absolute*, and let $\mathcal{A}$ be the set of belief revision structures which satisfy **R1**–**R3**. Then the following result formalizes the link between **TPI** and **TNI**, and absoluteness.

Theorem 2 *$BRSI$ is a sound and complete axiomatization w.r.t. $\mathcal{A}$ for formulas in $\mathcal{L}$.*

3.2 Consistency

The observant reader will have noticed that there is no axiom in BRS corresponding to the AGM *consistency axiom* $(K*5)$. In the AGM system, this axiom ensures that agents' beliefs are logically consistent whenever possible, i.e. whenever the information learned is logically consistent (if the information is not consistent, then $(K*2)$ forces inconsistent beliefs on the agent). But any attempt to axiomatize this in our logic will lead to circularity, since the notion of logical consistency presupposes a particular axiom system. The AGM system and Friedman and Halpern's [23] more expressive logic of belief change avoid this problem by working with two languages, one for describing facts about the world which can be learned, and another for talking about beliefs. Their consistency axioms ($(K*5)$ and **PS** respectively) apply to the second language, and make reference to the logical consistency only of formulas of the first. In addition to the analytical convenience of working with one language rather than two, an advantage of our approach is that no restrictions are imposed on the form that information may take. This issue is discussed in more detail in section 5.

An independent reason to be suspicious of the AGM consistency axiom is that it affords no purely semantic representation. To guarantee the validity of this axiom, we would need to restrict our attention to belief revision structures which contain *enough* worlds: for each logically consistent formula, we need at least one world in which that formula is true. Hence in Grove's [27] semantics for the AGM system, the set of worlds is identified with the set of maximal consistent sets of formulas of the object language, and Friedman and Halpern make the assumption that their structures are *saturated*, i.e. there is a least one minimal world for every consistent formula of the object language. But logicians tend to think of the *frame* (in this case the worlds and the plausibility orderings) as a representation of the extra-linguistic reality, which is mapped onto a formal language by an interpretation and semantic rules. A reality that can described only in syntactic terms seems artificial[7].

In the place of $(K*5)$, we consider the *weak consistency axiom*:

WCon $\quad \phi \Rightarrow \neg B_i^{\phi} false$

WCon says that as long as the information an agent receives is true, her revised beliefs are

[7] It may, however, be reasonable to impose linguistic constraints on belief revision structures for the sake of particular applications. For example, if we wish to model rational play in a game, we may wish to assume that there is at least one world corresponding to every strategy profile, or even that there is a world for every consistent set of beliefs the players might hold (as in the canonical structure of section 2.4). These restrictions represent contingent facts about particular situations, not matters of logic alone.

consistent[8], and is represented by the following assumption, which says that the actual world is always conceivable:

R4 for all i, w: $w \in W_i^w$.

Call a belief revision structure satisfying **R4** *inclusive*. Let $\mathcal{I}$ the class of all inclusive belief revision structures which also satisfy **R1** and **R2**, and let $BRSC$ be the axiom system consisting of BRS and **WCon**. Then:

Theorem 3 *$BRSC$ is a sound and complete axiomatization with respect to $\mathcal{I}$ for formulas in the language $\mathcal{L}$.*

3.3 A simplification

If we are willing to accept the introspection axioms, **TPI** and **TNI**, and consistency axiom, **WCon**, discussed above, the belief revision structures which provide the semantics for $\mathcal{L}$ can be greatly simplified. Let $BRSIC$ be the resulting axiom system (i.e. $BRSIC = BRS$+**TPI**+**TNI**+**WCon**). Theorem 4 says that **R1**–**R4** give a semantic characterization of $BRSIC$:

Theorem 4 *$BRSIC$ is a sound and complete axiomatization with respect to $\mathcal{A} \cap \mathcal{I}$ for formulas in the language $\mathcal{L}$.*

It turns out that if **R1**–**R4** are satisfied, the plausibility orderings of each agent can be replaced by a *single* binary relation, $\trianglelefteq_i$, defined as follows:

$$w \trianglelefteq_i x \text{ if and only if } w \preccurlyeq_i^x x.$$

The intuition is as follows: recall that **R3** says that each agent has the same plausibility ordering at every world conceivable to her, and **R4** says that the actual world is always conceivable. If both conditions are satisfied, the plausibility orderings divide the worlds into distinct subsets, which can then be described by a single relation. Although some information is lost by this transformation, since many different families of plausibility orderings for a given agent map onto the same $\trianglelefteq_i$ relation, all of the *semantically relevant* information is preserved. Truth conditions can be given

[8] For the purposes of modeling rational play in extensive games, replacing the AGM consistency axiom **WCon** is without loss of generality: the information structure of extensive form games is such that the information received is always true, and so **WCon** guarantees that agents maintain consistency of beliefs. For more on this point see Board [10].

in terms of the $\trianglelefteq_i$ relations which for every formula match the standard truth conditions. First observe that the set of conceivable worlds can be defined as follows:

$$W_i^w = \{x \mid x \trianglelefteq_i w \text{ or } w \trianglelefteq_i x\}.$$

To show that this definition is correct, we must prove

Proposition 2 $\{x \mid x \trianglelefteq_i w \text{ or } w \trianglelefteq_i x\} = \{x \mid x \preccurlyeq_i^w y \text{ for some } y\}$.

Next, we show how the $\trianglelefteq_i$ relation can be used to define truth of formulas. The truth conditions for primitive formulas, conjunctions and negations are the same as before; truth of formulas of the forms $B_i\phi$ and $B_i^\phi\psi$ are defined as follows:

$(M, w) \models B_i\phi$ iff $(M, x) \models \phi$ for all $x \in \min_i(W_i^w)$

$(M, w) \models B_i^\phi\psi$ iff $(M, x) \models \psi$ for all $x \in \min_i(\{[\phi]_M \cap W_i^w\})$

where $\min_i(X) = \{x \in X \mid \text{for all } y \in X,\ x \trianglelefteq_i y\}$. The equivalence of these truth conditions and the standard conditions follows immediately from Proposition 3:

Proposition 3 *for all* $X \subseteq W_i^w$, $\min_i(X \cap W_i^w) = \min_i^w(X \cap W_i^w)$.

The structures resulting from this simplification bear a very close resemblance to the belief revision models developed by Stalnaker [46]. Our $\trianglelefteq_i$ relations work in exactly the same way as the reverse of his Q_i relations: if we define wQ_ix if and only $x \trianglelefteq_i w$, then the truth conditions for formulas expressing beliefs and revised beliefs are identical. Furthermore, it can be shown our $\trianglelefteq_i$ relations satisfy the same properties he requires of his Q_i relations (i.e. they are reflexive and transitive, and if two worlds are related (in either direction) to a third world, then those two worlds are related (in some direction) to each other). Thus it follows from Theorem 4 that the axiom system $BRSIC$ provides a precise syntactic characterization of Stalnaker's purely semantic logic.

3.4 Common Belief

One of the strengths of the logic we are developing here, and of epistemic logic in general, is that they give us a complete description of agents' beliefs about agents' beliefs. As we discussed in the introduction, this is particularly useful for game theory since such beliefs are often considered

necessary for sophisticated strategic reasoning. In particular, we can give an account of the notion of *common belief* frequently used by economists.

But common belief cannot be expressed in the language $\mathcal{L}$ defined above, since infinite conjunctions of formulas in $\mathcal{L}$ are not themselves formulas of $\mathcal{L}$. To remedy this problem, we augment the language with the modal operators E ("everyone believes that..."), and C ("it is common belief that..."). Formally, $\mathcal{L}^C$ is defined by adding the following condition to the definition of $\mathcal{L}$ in section 2.1:

(*d*) if $\phi \in \mathcal{L}^C$, then $E\phi \in \mathcal{L}^C$ and $C\phi \in \mathcal{L}^C$.

It is straightforward to extend our axiom system to incorporate the E operator:

E $E\phi \Leftrightarrow \bigwedge_{i\in\{1,\ldots,n\}} B_i\phi$.

E says simply that everyone believes that ϕ if and only if every agent believes that ϕ. Unfortunately, the axiomatic characterization of common belief is trickier. The problem is that although common belief is an infinite concept, our axioms must be finite in length. It turns out that the following axiom-rule pair (familiar from epistemic logic) will serve our purpose:

FP $C\phi \Rightarrow E(\phi \wedge C\phi)$

IR from $\phi \Rightarrow E(\psi \wedge \phi)$ infer $\phi \Rightarrow C\psi$

FP and **IR** which are known as the *fixed-point axiom* and the *induction rule*, are harder to interpret. We shall merely remark that jointly they imply that common belief has all the properties of (individual) belief. For example, if B_i satisfies **TPI**, so too does C. Let BRS^C (respectively, $BRSI^C$, $BRSC^C$, and $BRSIC^C$) denote the axiom system formed by adding **E**, **FP**, and **RI** to BRS (respectively, $BRSI$, $BRSC$, and $BRSIC$).

The definition of truth for the augmented language, $\mathcal{L}^C$ is extended exactly as we would expect. $E\phi$ is true just if everyone believes that ϕ:

$(M, w) \models E\phi$ iff $(M, w) \models B_i\phi$ for all $i \in \{1, \ldots, n\}$;

and $C\phi$ is true if everyone believes that ϕ, everyone believes that everyone that ϕ, and so on. So, letting $E^0\phi$ be an abbreviation for ϕ, and $E^{k+1}\phi$ be an abbreviation for $EE^k\phi$, we have:

$(M, w) \models C\phi$ iff $(M, w) \models E^k\phi$ for $k = 1, 2, \ldots$

The following theorem confirms the equivalence of the syntactic and semantic characterization of common belief:

Theorem 5 *(a) BRS^C is a sound and complete axiomatization w.r.t. $\mathcal{M}$ for formulas in $\mathcal{L}^C$;*

(b) $BRSI^C$ is a sound and complete axiomatization w.r.t. $\mathcal{A}$ for formulas in $\mathcal{L}^C$;

(c) $BRSC^C$ is a sound and complete axiomatization w.r.t. $\mathcal{I}$ for formulas in $\mathcal{L}^C$;

(d) $BRSIC^C$ is a sound and complete axiomatization w.r.t. $\mathcal{A} \cap \mathcal{I}$ for formulas in $\mathcal{L}^C$.

4 Comments

4.1 Knowledge

While our language allows us to make statements about agents' beliefs (and how these beliefs are revised), economists often make assumptions about agents' *knowledge*. Knowledge could be modeled by adding another set of modal operators to our language: K_i ("i knows that...").

As we discussed in the introduction, in the economics literature knowledge is most commonly analyzed using Aumann's [2] information partition model. The properties of this model are well understood. An Aumann structure can be provided with an interpretation and used to provide truth conditions for a language containing knowledge operators. The properties of the knowledge operators can then be precisely described by a set of axioms which are sound and complete with respect to the class of all (enriched) Aumann structures. In addition to the appropriate analogues of **Taut**, **Dist**, **MP** and **RE**, this axiom system contains:

T	$K_i\phi \Rightarrow \phi$
PI	$K_i\phi \Rightarrow K_iK_i\phi$
NI	$\neg K_i\phi \Rightarrow K_i\neg K_i\phi$

T, the *truth* axiom, is uncontroversial: it says simply that what is known must be true; **PI** (*positive introspection*) and **NI** (*negative introspection*), on the other hand, are even more controversial in the context of knowledge than in the context of belief. They say respectively that an agent knows what she knows and knows what she doesn't know. The problem is the following: as long as we accept the truth axiom, the concept of knowledge imposes an external condition on the agent's cognitive state. Thus even if the agent has complete introspective access to what she

believes and doesn't believe, the introspection axioms do to carry over to knowledge through logic alone.

So it seems that we must reject **PI** and **NI**, and take **T** as a starting point in the analysis of knowledge. But philosophers have long argued that true belief, while necessary, is not a sufficient condition for knowledge. For a true belief to be classified as knowledge, it is required in addition that it be somehow justified in an appropriate manner. Reflecting this requirement, Stalnaker [46] appeals to an analysis of knowledge called the *defeasibility analysis*. The idea behind this account is that "if a person has knowledge, then that person's justification must be sufficiently strong that it is not capable of being defeated by evidence that he does not possess" (Pappas and Swain [38]).

Stalnaker uses his (semantic) model of belief revision to formalize this idea, by defining knowledge as follows: an agent knows that ϕ if and only if ϕ is true, she believes that ϕ, and she continues to believe that ϕ if any true information is received. Truth conditions for formulas of the form $K_i\phi$ can be provided as follows[9]:

$(M, w) \models K_i\phi$ iff $(M, w) \models \phi$ and if $(M, w) \models \psi$, then $(M, w) \models B_i^{\psi}\phi$.

But we have been unable to find a finite axiomatization: the direct translation of Stalnaker's definition involves a formula of infinite length. Let $\psi_1, \psi_2, \ldots$ be an enumeration of all the formulas of $\mathcal{L}$. Then the appropriate axiom would be:

Know $\quad K_i\phi \Leftrightarrow \left(\phi \wedge \left(\psi_1 \Rightarrow B_i^{\psi_1}\phi\right) \wedge \left(\psi_2 \Rightarrow B_i^{\psi_2}\phi\right) \wedge \ldots\right)$

(Note that we do not need to include the formula $B_i\phi$ on the right hand side, since, in the presence of **Taut** and **Triv**, it is implied by $\psi \Rightarrow B_i^{\psi}\phi$ when *true* is substituted for ψ.)

An alternative approach would be treat Stalnaker's definition as providing a necessary but not sufficient condition for knowledge:

Know' $\quad K_i\phi \Rightarrow \left(\phi \wedge \left(\psi \Rightarrow B_i^{\psi}\phi\right)\right)$

It is an open question whether the axiom system consisting of BRS and **Know'** is sound and complete given the proposed semantics.

4.2 Iterated Belief Revision

In an extensive form game, of course, beliefs may need to be revised more than once: new information may be received at each round of the game. But the language $\mathcal{L}$ is not rich enough to express

[9] If we assume that **R4** holds, this can be replaced by the much simpler $(M, w) \models K_i\phi$ iff $(M, x) \models \phi$ for all $x \preccurlyeq_i^w w$, which is precisely the condition given by Stalnaker.

iterated belief revisions. Although it would be a simple task to augment the language to allow such iterations, it is not obvious what the axioms governing these iterations should be. Nor is it clear how the semantics should be extended: the rules for revision give us a new set of most plausible worlds after a formula ϕ is learned (that is, the lowest ranking members of $[\phi]$), but we are not told the relative plausibilities of all the other worlds. So we need some way of preserving the relative plausibility data given by the $\preccurlyeq_i^w$ relations, while taking into account the new information ϕ. More precisely, we need to construct a new ordering that represents the revised epistemic state, so that we can re-apply the revision rule as more information is learned. Further discussion of these issues is beyond the scope of this paper. The interested reader is referred to Spohn [44].

These issues can be avoided if we restrict our attention to extensive form games of perfect recall, which are the focus of much of modern game theory. Such games can be analyzed without loss of generality by considering only single revisions if we are willing to accept a plausible assumption. In such games, the sequence of information that the players receive as the game progresses and they observe moves made by their opponents is of a very particular kind: each new piece of information logically implies every previous piece, since the set of strategies consistent with any given information set for a player is a subset of the set of strategies consistent with any predecessor information sets[10]. If ψ logically implies ϕ, it seems reasonable to assume that learning ϕ and then ψ will generate the same beliefs as if one learns ψ at first: in both cases exactly the same information is learned. This assumption is adopted by Board [9].

5 Related Literature

Much of the related literature has been discussed in the main text of this paper. Here we provide a summary and mention some important omissions.

The semantics of our logic bear a close resemblance to those of conditional logic (Lewis [35], Burgess [16]). The belief revision structures considered above are essentially a multi-agent version of Burgess' *models*. **R1** corresponds to his transitivity and connectivity requirements. The counterpart of **R2** is Lewis' *limit assumption* (L), and **R3** and **R4** correspond to *local absoluteness* (A-) and *total reflexivity* (T) respectively. But their models are used to provide truth conditions for conditional formulas, while here we are interested in belief revision; the axioms of conditional logic are very

[10] Board [10] suggests that this information structure is unduly restrictive, even under the assumption of perfect recall, and shows that it rules out certain interesting situations.

different from the axioms of belief revision.

Stalnaker [46] also uses semantic structures very similar to those employed here; the results of section 3.3 show that his models are essentially belief revision structures which satisfy **R1**–**R4**. But he provides no formal language and no syntax. Thus the results of this paper are complementary to his: our axiom system *BRSIC* can be thought of as characterizing his models. Friedman and Halpern's [23] logic of belief change does provide a syntax as well as a semantics for characterizing the belief revision process, with soundness and completeness theorems linking the two. The key difference between their work and our own (as mentioned in section 3.2) is that they use two distinct formal languages, one for describing facts about the physical world which can be learned, and another for talking about beliefs. Thus in their system agents cannot learn about each other's beliefs. Friedman and Halpern suggest that this restriction is necessary to avoid a triviality result established by Gärdenfors [24], but the results of this paper show that this is not the case. Triviality can be avoided as long as the *Ramsey test* ($B_i^{\phi} B_i^{\psi} \chi \Leftrightarrow B_i^{\phi \wedge \psi} \chi$) is not adopted as an axiom. And there are many situations where information *is* of this form: agents seeking consultancy advice pay to learn about the beliefs of others which may or may not be an accurate reflection of reality; similarly expert testimony in the courtroom yields information about the expert's beliefs, and not hard facts about the physical world.

Alternative models of interactive belief revision have been developed by Battigalli and Siniscalchi [6] and Brandenburger and Keisler [15], who show how the hierarchical approach of Mertens and Zamir [37] can be extended to the dynamic setting. The differences between these models and the current work have been discussed in the introduction and in section 2.5.

6 Conclusion

The aim of this paper has been to develop a dynamic model of interactive reasoning which combines analytical simplicity with clarity of interpretation. Belief revision structures are similar to the models used very successfully by Stalnaker to analyze rational play in extensive form games [46], and to shed light on the forward and backward induction procedures [47]. These structures provide truth conditions for a formal language. Soundness and completeness theorems establish tight connections between the formulas that are true in various classes of belief revision structure, and those that are provable in certain axiom systems, thereby giving us a precise understanding of what the structures mean.

References

[1] ALCHOURRÓN, C. E., P. GÄRDENFORS, AND D. MAKINSON (1985), "On the Logic of Theory Change: Partial Meet Functions for Contraction and Revision", *Journal of Symbolic Logic* **50**, 510–530.

[2] AUMANN, R. J. (1976), "Agreeing to Disagree", *Annals of Statistics* **4**, 1236–1239.

[3] AUMANN, R. J. AND A. BRANDENBURGER (1995), "Epistemic Conditions for Nash Equilibrium", *Econometrica* **63**, 1161–1180.

[4] AUMANN, R. J. (1999), "Interactive Epistemology I: Knowledge", *International Journal of Game Theory* **28**, 263–300.

[5] BATTIGALLI, P, AND G. BONANNO (1999), "Recent results on belief, knowledge and the epistemic foundations of game theory", *Research in Economics* **53**, 149–225.

[6] BATTIGALLI, P. AND M. SINISCALCHI (1999), "Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games", *Journal of Economic Theory* **88**, 188-230.

[7] BATTIGALLI, P. AND M. SINISCALCHI (2001), "Strong Belief and Forward Induction Reasoning", *Journal of Economic Theory*, forthcoming.

[8] BOARD, O. J. (1998), "Belief Revision and Rationalizability", TARK VII, Conference Proceedings, ed. by I. Gilboa.

[9] BOARD, O. J. (2002), "Algorithmic Characterization of Rationalizability in Extensive Form Games", working paper, Department of Economics, Oxford.

[10] BOARD, O. J. (2002), "The Deception of the Greeks: Generalizing the Information Structure of Extensive Form Games", working paper, Department of Economics, Oxford.

[11] BONANNO, G. "Memory and Perfect Recall in Extensive Games", working paper, Department of Economics, University of California, Davis.

[12] BRANDENBURGER, A. (2002), "On the Existence of a 'Complete' Possibility Structure", working paper, Harvard Business School.

[13] BRANDENBURGER, A. AND E. DEKEL (1993), "Hierarchies of Beliefs and Common Knowledge", *Journal of Economic Theory* **59**, 189–198

[14] BRANDENBURGER, A. AND H. J. KEISLER (1999), "An Impossibility Theorem on Beliefs in Games", working paper, Harvard Business School.

[15] BRANDENBURGER, A. AND H. J. KEISLER (2002), "Epistemic Conditions for Iterated Admissability", working paper, Harvard Business School.

[16] BURGESS, J. P. (1981), "Quick Completeness Proofs for Some Logics of Conditionals", *Notre Dame Journal of Formal Logic* **22**, 76–84.

[17] DEKEL, E. AND F. GUL (1997): "Rationality and Knowledge in Game Theory", in *Advances in Economics and Econometrics: Theory and Applications: Seventh World Congress, Vol. 1*, ed. by D. M. Kreps and K. W. Wallis. Cambridge University Press, 87–172.

[18] DEKEL, E., B. LIPMAN, AND A. RUSTICHINI (1998), "Standard State-Space Models Preclude Unawareness", *Econometrica* **66**, 159–173

[19] FAGIN, R. (1994), "A Quantitative Analysis of Modal Logic", *Journal of Symbolic Logic* **59**, 209–252.

[20] FAGIN, R., AND J. Y. HALPERN (1994), "Reasoning about Knowledge and Probability", *Journal of the Association for Computing Machinery* **41**, 340–367

[21] FAGIN, R., J. Y. HALPERN, Y. MOSES, AND M. Y. VARDI (1995), *Reasoning About Knowledge*. The MIT Press, Cambridge, MA.

[22] FAGIN, R., J. GEANAKOPLOS, J. Y. HALPERN, AND M. Y. VARDI (1999), "The Hierarchical Approach to Modeling Knowledge and Common Knowledge", *International Journal of Game Theory* **28**, 331–365.

[23] FRIEDMAN, N. AND J. Y. HALPERN (1994), "Conditional Logics of Belief Change", in *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, 915–921. AAAI Press, Menlo Park, CA.

[24] GÄRDENFORS, P. (1986), "Belief Revisions and the Ramsey Test for Conditionals", *The Philosophical Review* **XCV**, 81–93.

[25] GÄRDENFORS, P. (1988), *Knowledge in Flux*. The MIT Press, Cambridge, MA.

[26] GEANAKOPLOS, J. (1992), "Common Knowledge", *Journal of Economic Perspectives* **6**, 53–82.

[27] GROVE, A. (1988), "Two Modellings for Theory Change", *Journal of Philosophical Logic* **17**, 157–170

[28] HALPERN, J. Y. (1998), "Set-Theoretic Completeness for Epistemic and Conditional Logic", *Annals of Mathematics and Artificial Intelligence* **26**, 1–27.

[29] HARPER, W. L. (1977), "Rational Conceptual Change", in *Philosophy of Science Association* vol. 2, 462–494. East Lansing, MI.

[30] HARSANYI, J. (1968), "Games with Incomplete Information Played by 'Bayesian' Players", parts I–III, *Management Science* **14**, 159–182, 320–334, 486–502.

[31] HEIFETZ, A. AND D. SAMET (1998), "Knowledge spaces with arbitrarily high rank", *Games and Economic Behavior* **22**, 260–273.

[32] HINTIKKA, J. (1962), *Knowledge and Belief.* Cornell University Press, Ithaca, NY.

[33] JEFFERY, R. C. (1965), *The Logic of Decision.* McGraw-Hill, NY.

[34] KRIPKE, S. (1963), "A semantical analysis of modal logic I: normal modal propositional calculi", *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* **9**, 67–96.

[35] LEWIS, D. (1973), *Counterfactuals.* Basil Blackwell, Oxford.

[36] MENDELSON, E. (1964), *Introduction to Mathematical Logic.* D. Van Nostrand Company, NJ.

[37] MERTENS, J.-F. AND S. ZAMIR (1985), "Formulation of Bayesian Analysis for Games of Incomplete Information", *International Journal of Game Theory* **14**, 1–29

[38] PAPPAS, G. AND M. SWAIN (eds.) (1978), *Essays on Knowledge and Justification*, Cornell University Press, Ithaca, NY.

[39] PEARCE, D. G. (1984): "Rationalizable Strategic Behavior and the Problem of Perfection", *Econometrica* **52**, 1029–1050.

[40] PICCIONE, M. AND A. RUBINSTEIN, "On the Interpretation of Decision Problems with Imperfect Recall", *Games and Economic Behavior* **20**, 3–24.

[41] RANTALA, V. (1982), "Impossible Worlds Semantics and Logical Omniscience", *Acta Philosophica Fennica* **35**, 106–15.

[42] RENY, P. J. (1993), "Common Belief and the Theory of Games with Perfect Information", *Journal of Economic Theory* **59**, 257–274.

[43] SAMET, D. (1990), "Ignoring Ignorance and Agreeing to Disagree", *Journal of Economic Theory* **52**, 190–207.

[44] SPOHN, W. (1988), "Ordinal Conditional Functions: A Dynamic Theory of Epistemic States", in *Causation in Decision, Belief Change, and Statistics* vol. II, ed. by W. L. Harper and B. Skyrms, Kluwer, Dordrecht, 105–134.

[45] STALNAKER, R. (1988), *Inquiry.* The MIT Press, Cambridge, MA.

[46] STALNAKER, R. (1996), "Knowledge, Belief and Counterfactual Reasoning in Games", *Economics and Philosophy* **12**, 133–163.

[47] STALNAKER, R. (1998), "Belief Revision in Games: Forward and Backward Induction", *Mathematical Social Sciences* **36**, 31–56.

[48] TAN, T. C. C. AND S. R. C. WERLANG (1988): "The Bayesian Foundations of Solution Concepts of Games", *Journal of Economic Theory* **45**, 370–391.

A Epistemic logic

We present here a brief review of epistemic logic. For a very comprehensive account, refer to Fagin *et al.* [21].

The language of epistemic logic is the set of formulas $\mathcal{L}^{EL}$ that can be built up from a nonempty set Φ of primitive formulas, using negation, conjunction, and the modal operators $B_1, \ldots, B_n$. Formally, $\mathcal{L}^{EL}$ is the smallest set such that: $\phi \in \mathcal{L}^{EL}$ if $\phi \in \Phi$; and if $\phi, \psi \in \mathcal{L}^{EL}$, then $\neg\phi, \phi \wedge \psi, B_i\phi \in \mathcal{L}^{EL}$ for $i = 1, \ldots, n$. We shall also use implication, defined in the usual way, i.e. $\phi \Rightarrow \psi$ is shorthand for $\neg(\phi \wedge \neg\psi)$.

Consider the following axioms and rules:

Taut	*true*
K	$(B_i\phi \wedge B_i(\phi \Rightarrow \psi)) \Rightarrow B_i\psi$
T	$B_i\phi \Rightarrow \phi$
4	$B_i\phi \Rightarrow B_iB_i\phi$
5	$\neg B_i\phi \Rightarrow B_i\neg B_i\phi$
D	$\neg B_i false$
MP	from ϕ and $\phi \Rightarrow \psi$ infer ψ
Gen	from ϕ infer $B_i\phi$

Note that *true* stands for any tautology of propositional calculus, and any formulas in $\mathcal{L}^{EL}$ may be substituted for ϕ and ψ. **Taut** and **MP** (*modus ponens*) comprise the standard axiom system of propositional calculus. **K** is the *distribution axiom* for belief, and says that our agents' beliefs are closed under *modus ponens*; and **Gen**, the *rule of epistemization* or *belief generalization rule*, says that they believe every theorem. Taken together, these two properties, often referred to as *logical omniscience*, imply that our agents are very powerful reasoners indeed. Theorem 6 below shows that logical omniscience is an inescapable consequence of the Kripke structure semantics. **T** is the *knowledge axiom*, which says that whatever is believed is true: it is more appropriate if we are modeling knowledge rather than belief. **4** and **5** are the *positive introspection axiom* and *negative introspection axioms*, and jointly imply the agent has complete introspective access to her own beliefs. **D** is the consistency axiom, which says that the agent's beliefs are logically consistent. Various combinations of the axioms and rules make up different axiom systems. Among the more common are K, which consists of **Taut**, **K**, **MP** and **Gen**; $T = K + \mathbf{T}$; $S5 = T + \mathbf{4} + \mathbf{5}$; and $KD45 = K + \mathbf{4} + \mathbf{5} + \mathbf{D}$.

We now describe the semantics for $\mathcal{L}^{EL}$. Truth conditions are assigned to every formula by means of *Kripke structures*. A Kripke structure M for n agents is a tuple $\langle W, \pi, \mathcal{B}_1, \ldots, \mathcal{B}_n \rangle$, where W is a set of *possible worlds* or *states*[11], π is an *interpretation* that associates with each possible world in W a truth assignment to the primitive formulas of Φ i.e. $\pi : \Phi \times W \rightarrow \{\mathbf{true}, \mathbf{false}\}$[12], and each $\mathcal{B}_i$ is a binary relation on W, the *accessibility relation* of agent i. The basic idea behind the set W is that in addition to the ways things actually are, there are number of other *ways things might have been*. A possible world $w \in W$ is simply a complete description of one of these ways.

[11] The terminology *possible world* is generally used in the computer science and philosophy literature, while economists talk about *states*.

[12] Note that these are not the same as the logical constants *true* and *false* of the language .

The interpretation π gives us a truth assignment for each of the primitive formulas at each possible world, so that if ϕ denoted the sentence "it is wet", then $\pi(w)(\phi) = \mathbf{true}$ means that (in model M) it is wet in world w. $\mathcal{B}_i$ tells us, for each possible world, which worlds are accessible to the agent from that world, i.e. which worlds she considers might be the actual world. So if $(w, x) \in \mathcal{B}_i$, then world x is accessible to agent i given her information at world w. This relation is used to give truth conditions for belief statements. Intuitively, an agent believes at world w that it is wet if and only if it is wet at every world accessible to her from w.

We shall now give a complete definition of what it is for a formula to be true at a given world in a Kripke structure. Note that truth needs to be defined at every world in a structure, as a formula may be true in one world yet false in another. We use the notation $(M, w) \models \phi$, read as "ϕ is true at world w of model M" to capture this.

The $\models$ relation is defined by induction on the form of ϕ. For primitive formulas, the interpretation π gives us all the information we need. For all $\phi \in \Phi$,

$(M, w) \models p$ iff $\pi(w)(p) = \mathbf{true}$.

Conjunctions and negations are then dealt with in the standard way from propositional logic:

$(M, w) \models \phi \wedge \psi$ iff $(M, w) \models \phi$ and $(M, w) \models \psi$;

$(M, w) \models \neg\phi$ iff not $(M, w) \models \phi$.

That is, a conjunction $\phi \wedge \psi$ is true exactly if both of the conjuncts are true, and a negation $\neg\phi$ is true exactly if ϕ is not true. Note that this last definition implies that our logic is two-valued: for every formula ϕ, either $(M, w) \models \phi$ or $(M, w) \models \neg\phi$, but not both.

For formulas of the form $B_i\phi$, recall that we say an agent believes ϕ just if ϕ is true at every world accessible to her. Hence,

$(M, w) \models B_i\phi$ iff $(M, x) \models \phi$ for all x such that $(w, x) \in \mathcal{B}_i$.

This completes the definition of $\models$, and gives us the truth conditions of every formula at every world of a structure.

Let $\mathcal{M}$ be the class of all Kripke structures. We are also interested in several subsets of $\mathcal{M}$: let $\mathcal{M}^r$ (respectively $\mathcal{M}^{re}$; $\mathcal{M}^{est}$) be the class of Kripke structures in which each $\mathcal{B}_i$ relation is

reflexive (respectively reflexive and Euclidean; Euclidean, serial, and transitive). Then the following soundness and completeness results[13] link the syntax and semantics:

Theorem 6 *For formulas in the language $\mathcal{L}$:*

(a) K is a sound and complete axiomatization with respect to $\mathcal{M}$;.

(b) T is a sound and complete axiomatization with respect to $\mathcal{M}^r$;

(c) $S5$ is a sound and complete axiomatization with respect to $\mathcal{M}^{re}$;

(d) $KD45$ is a sound and complete axiomatization with respect to $\mathcal{M}^{est}$.

B The AGM Axioms for Belief Revision

Appendix B gives a short description of the AGM belief revision theory. A more complete account is provided by Gärdenfors [25].

As we discussed in the introduction, in order to analyze rational play in extensive form games, it is crucial to have a precise model not only of the players' beliefs at any one point in time but also of the way these beliefs are revised as the game proceeds. The tool traditionally employed by economists to analyze how beliefs change when new information is learned is Bayes' rule. While this theory gives us a unique answer when the new information is compatible with one's existing beliefs, it is not well-defined if we learn something which is logically inconsistent with what we previously believed (and has thus been previously assigned zero probability). But these are precisely the kind of belief changes we must consider if we are to model counterfactual reasoning in games. We need to analyze what the players will or would believe at every information set in the game, and not just at those information sets which (they believe) may actually be reached. That is, we need to consider what happens if the players are surprised by finding themselves at information sets they believed would not transpire, and some of their existing beliefs must be given up. The problem of belief revision is that there is a multitude of ways to select just how this should be done. As Stalnaker [45] puts it:

> A belief change in response to conflicting information will always force one to choose between alternative revisions, none of which can be seen, on logical grounds alone, to

[13] See section 2.4 for an explanation of *soundness* and *completeness*.

> be preferable to the others...we will need to impose additional structure on our notion of a belief state before we can say very much about the way beliefs change or ought to change in response to new information.

The AGM theory of belief revision (so-called after Alchourrón, Gärdenfors, and Makinson [1]) provides a set of axioms which, they argue, any reasonable belief revision system should satisfy. The "additional structure" to which Stalnaker refers is then provided by a belief revision function satisfying the axioms. The AGM theory represents an agent's epistemic state at any one point in time by a set of formulas of propositional calculus, the agent's *belief set* K. This is interpreted as the set of all formulas the agent believes. Belief statements are thus expressed in a meta-language ("inclusion in K"). Note that since $\mathcal{L}^{PC}$ contains no belief operators, we cannot express beliefs about beliefs (hence we adopt the richer language $\mathcal{L}$ in the multi-agent belief revision model developed above). It is assumed that belief sets are closed under logical consequence, according to the axiom system PC consisting of **Taut** and **MP**. Using $Cn(\mathrm{K})$ to denote the set of all logical consequences of K, it follows that if K is a belief set, then $\mathrm{K} = Cn(\mathrm{K})$. Let $\mathbf{K}$ denote the set of all belief sets. Although we normally assume that belief sets are PC-consistent (i.e. $false \notin \mathrm{K}$), it is convenient to define the *absurd* belief set $\mathrm{K}_{false} = \mathcal{L}^{PC}$, in which everything is believed. If the agent's belief set is PC-consistent, then for any formula ϕ only three different epistemic attitudes can be expressed:

(i) $\phi \in \mathrm{K} : \phi$ is believed or *accepted*;

(ii) $\neg\phi \in \mathrm{K} : \phi$ is *rejected*;

(iii) $\phi \notin \mathrm{K}$ and $\neg\phi \notin \mathrm{K} : \phi$ is *indetermined.*

A belief change concerning ϕ is simply a change from one of these epistemic attitudes into another. Thus six types of belief change are possible, but as long as our beliefs sets are PC-consistent, we need only consider three, since "ϕ is accepted" implies that "$\neg\phi$ is rejected". The first type occurs when "ϕ is indetermined" is changed into "ϕ is accepted". This kind of change is called an *expansion*, because new beliefs are added to the belief set but none of the old beliefs are retracted (for reasons we shall see). A *revision* occurs when "ϕ is rejected" is changed into "ϕ is accepted", and the final type of change is called a *contraction*, which occurs when either "ϕ is accepted" or "ϕ is rejected" is changed into "ϕ is indetermined".

As we have said, there is in general no unique way of defining the belief change process, and the most we can hope to do is lay down a set of rules that any such process should satisfy if it is to be rational. The guiding principle of rationality is that when we change our beliefs in response to new evidence, the change should be no greater than is necessary to incorporate that new evidence. Our beliefs are not in general gratuitous, and so on the one hand any unnecessary loss of information should be avoided (the *conservativity principle*), and on the other we should not add any beliefs that have no connection with the new evidence (the *relevancy principle*). The two principles are known jointly as the *criterion of informational economy.* The theorist's task is formally to capture the notion of minimal change expressed by this principle.

In the case of expansions, the criterion of informational economy does in fact imply a unique belief change function (that is, a function that tells us how epistemic states change given any new information). When information is learned that is logically consistent with the current belief set, the new belief set should be formed simply by adding the new sentence to the old belief set, and closing under logical implication. None of the old beliefs are retracted (hence the term *expansion*), in accordance with the conservativity principle, nor are any added except the new information itself and those required to ensure that closure under logical implication is still satisfied. Formally, $\mathrm{K}_{\phi}^{+} = Cn\left(\mathrm{K} \cup \{\phi\}\right)$, where K_{ϕ}^{+} denotes the expansion of K by ϕ. Note that for technical convenience, the expansion function is defined for *all* new information, not only information that is consistent with current beliefs. If ϕ contradicts K (i.e. if $\neg\phi \in \mathrm{K}$), then $\mathrm{K}_{\phi}^{+} = \mathrm{K}_{false}$.

The criterion of informational economy is also applied to revisions, and although it does not yield a unique belief change function in this case, it does enable us to build a set of postulates, the *AGM axioms,* that any revision function should satisfy. The first two ensure simply that the function does indeed give us a new belief state, and that the new information is accepted in the revised state. Letting K_{ϕ}^{*} denote the revised belief set after information ϕ is learned for a given revision function $* : \mathbf{K} \times \mathcal{L}_{PC} \rightarrow \mathbf{K}$, they are respectively

(K*1) K_{ϕ}^{*} is a belief set;

(K*2) $\phi \in \mathrm{K}_{\phi}^{*}$.

In order that K_{ϕ}^{*} is defined for all sentences ϕ, we include the case where $\neg\phi \notin \mathrm{K}$, so the new information does *not* contradict the original belief state, and the belief change is actually an expansion rather than a revision. The next two postulates ensure that, in these cases, K_{ϕ}^{*} is the same as K_{ϕ}^{+}:

(K*3) $K^*_\phi \subseteq K^+_\phi$;

(K*4) if $\neg\phi \notin K$, then $K^+_\phi \subseteq K^*_\phi$.

The first of these postulates is motivated by the relevancy principle, and the second by the conservativity principle. Note that we do not need the proviso that ϕ be consistent with original belief state in (K*3), since if it is not, K^+_ϕ is the absurd belief state, and hence (K*3) is satisfied trivially. For revisions, however, we want to avoid inconsistent beliefs as much as possible. The next postulate says that, unless the new information itself is logically inconsistent, the revised belief state is consistent.

(K*5) $K^*_\phi = K_{false}$ iff ϕ is logically inconsistent.

The sixth postulate ensures that it is the content of the new information rather than the way that information is expressed that determines the belief change.

(K*6) If $\phi \Leftrightarrow \psi$, then $K^*_\phi = K^*_\psi$.

The final two postulates apply to iterated belief changes. The idea is that the minimal change of K to include both ϕ and ψ (i.e. $K^*_{\phi\wedge\psi}$) ought to be the same as the expansion of K^*_ϕ by ψ, as long as ψ does not contradict the beliefs expressed by K^*_ϕ. Any other change would be more than is necessary to incorporate all of the new evidence, and hence go against the criterion of informational economy.

(K*7) $K^*_{\phi\wedge\psi} \subseteq \left(K^*_\phi\right)^+_\psi$;

(K*8) if $\neg\psi \notin K^*_\phi$, then $\left(K^*_\phi\right)^+_\psi \subseteq K^*_{\phi\wedge\psi}$.

As mentioned above, (K*1)–(K*8) do not determine a unique belief revision function. Rather, they are intended to serve as formalization of the criterion of informational economy and provide a logical framework for analyzing belief change.

Finally, contractions can be defined in terms of revisions by means of the *Harper identity*: $K^-_\phi = K \cap K^*_{\neg\phi}$, where K^-_ϕ denotes the contraction of K by ϕ.

C Proofs

THEOREM 1

SOUNDNESS

The proof of soundness is straightforward, and proceeds by induction on the length of a proof of ϕ. Every element of a proof is either an axiom or follows from previous elements by the application of a rule, so we must show that every axiom is valid with respect to $\mathcal{M}$ and that each rule is truth preserving. We consider the cases of **Triv**, **IE(a)** and **LE**, and leave the rest as an exercise.

Triv: we must show that $\mathcal{M} \models B_i\phi \Leftrightarrow B_i^{true}\phi$. Propositional reasoning and the definition of $\models$ imply that $(M, w) \models true$, for all M and w. Thus $[true]_M = W$, and $W_i^w = [true]_M \cap W_i^w$. From the definition of $\models$ again, it follows that $(M, w) \models B_i\phi$ iff $(M, w) \models B_i^{true}\phi$.

IE(a): we must show that $\mathcal{M} \models B_i^\phi\psi \Rightarrow \left(B_i^{\phi\wedge\psi}\chi \Leftrightarrow B_i^\phi\chi\right)$. Suppose $(M, w) \models B_i^\phi\psi$; then $(M, x) \models \psi$ for all $x \in \min_i^w \{[\phi]_M \cap W_i^w\}$, and $\min_i^w \{[\phi]_M \cap W_i^w\} = \min_i^w \{[\phi]_M \cap [\psi]_M \cap W_i^w\}$. From the definition of $\models$, we have $[\phi]_M \cap [\psi]_M = [\phi \wedge \psi]_M$; therefore $\min_i^w \{[\phi]_M \cap [\psi]_M \cap W_i^w\} = \min_i^w \{[\phi \wedge \psi]_M \cap W_i^w\}$. It follows immediately that $(M, w) \models B_i^{\phi\wedge\psi}\chi$ iff $(M, w) \models B_i^\phi\chi$, as required.

LE: we must show that if $M \models \phi \Leftrightarrow \psi$, then $M \models B_i^\phi\chi \Leftrightarrow B_i^\psi\chi$. Suppose that $M \models \phi \Leftrightarrow \psi$. Then by the definition of $\models$, $[\phi]_M = [\psi]_M$, and so $\min_i^w \{[\phi]_M \cap W_i^w\} = \min_i^w \{[\psi]_M \cap W_i^w\}$. It follows immediately that $(M, w) \models B_i^\phi\chi$ iff $(M, w) \models B_i^\psi\chi$, as required.

COMPLETENESS

We start with some definitions. For a given axiom system AX, we say that a formula ϕ is *AX-consistent* if $\neg\phi$ is not provable in AX. A finite set of formulas $\{\phi_1, \ldots, \phi_k\}$ is AX-consistent exactly if $\phi_1 \wedge \ldots \wedge \phi_k$ is $AX-$consistent, and an infinite set of formulas is AX-consistent exactly if all its finite subsets are AX-consistent. Finally, given two sets of formulas S, T with $S \subseteq T \subseteq \mathcal{L}$, we say that S is a *maximal AX-consistent subset of T* if (a) it is AX-consistent, and (b) for all ϕ in T but not in S, the set $S \cup \{\phi\}$ is not AX-consistent.

Now, to prove completeness, we must show that every formula in $\mathcal{L}$ that is valid with respect to $\mathcal{M}$ is provable in BRS. It is sufficient to prove that

$(*)$ Every BRS-consistent formula in $\mathcal{L}$ is satisfiable with respect to $\mathcal{M}$.

For assume that we can prove $(*)$, and that ϕ is a valid formula in $\mathcal{L}$. If ϕ is not provable in BRS, then neither is $\neg\neg\phi$, so, by definition, $\neg\phi$ is BRS-consistent. It follows from $(*)$ that $\neg\phi$ is satisfiable with respect to $\mathcal{M}$, contradicting the validity of ϕ with respect to $\mathcal{M}$.

Before proceeding, we need another round of definitions. Let $Sub(\phi)$ be the set of all subformulas of ϕ; formally, $\psi \in Sub(\phi)$ if either (a) $\psi = \phi$, or (b) ϕ is of the form $\neg\phi', \phi' \wedge \phi'', B_i\phi'$, or $B_i^{\phi'}\phi''$, and $\psi \in Sub(\phi')$ or $\psi \in Sub(\phi'')$; and let $Sub^{+}(\phi)$ consist of all the formulas in $Sub(\phi)$ and their negations and conjunctions, i.e. $Sub^{+}(\phi)$ is the smallest set such that (a) if $\psi \in Sub(\phi)$ then $\psi \in Sub^{+}(\phi)$; and (b) if $\psi, \chi \in Sub^{+}(\phi)$, then $\neg\psi, \psi \wedge \chi \in Sub^{+}(\phi)$. Let $Sub^{++}(\phi)$ consist of all formulas of $Sub^{+}(\phi)$ together with all formulas of the form $B_i\psi$ and $B_i^{\chi}\psi$, where $\psi, \chi \in Sub^{+}(\phi)$; and let $Sub^{++}_{neg}(\phi)$ consist of all the formulas in $Sub^{++}(\phi)$ and their negations. Finally, let $Con(\phi)$ be the set of maximal BRS-consistent subsets of $Sub^{++}_{neg}(\phi)$. It is easy to show[14] that every BRS-consistent subset of $Sub^{++}_{neg}(\phi)$ can be extended to an element of $Con(\phi)$ by addition of formulas; and if S is a member of $Con(\phi)$, it must satisfy the following properties:

- for every $\psi \in Sub^{++}(\phi)$, exactly one of ψ and $\neg\psi$ is in S;
- if $\psi \wedge \chi \in S$ then $\psi \in S$ and $\chi \in S$;
- if $\psi \vee \chi \in S$ then $\psi \in S$ or $\chi \in S$;
- if $\psi \in S$ and $\psi \Rightarrow \chi \in S$, then $\chi \in S$;
- if $\psi \Leftrightarrow \chi$ then $\psi \in S$ if and only if $\chi \in S$;
- if $\psi \in Sub^{++}_{neg}(\phi)$ and $BRS \vdash \psi$, then $\psi \in S$.

To prove $(*)$, we construct a special structure $M_\phi \in \mathcal{M}$ for each ϕ. M_ϕ has a world w_S corresponding to every $S \in Con(\phi)$; we show that for all $\psi \in Sub(\phi)$, we have

$(**)$ $(M_\phi, w_S) \models \psi$ if and only if $\psi \in S$,

i.e. a formula in $Sub(\phi)$ is true at world w_S exactly if it is one of the formulas in S. This is sufficient to prove $(*)$, since if ϕ is BRS-consistent, it is contained in some set $S \in Con(\phi)$; it then follows from $(**)$ that $(M_\phi, w_S) \models \phi$, and so ϕ is satisfiable with respect to $\mathcal{M}$ as required.

[14] See e.g. Fagin *et al.* [21] pp. 52, 53.

For the construction of M_ϕ we introduce some new notation: let $S/B_i^\phi = \left\{\psi \mid B_i^\varphi \psi \in S\right\}$, i.e. S/B_i^ϕ is the set of formulas believed by i when she learns that ϕ. We now define M_ϕ. Let $M_\phi = \langle W, \pi, \preccurlyeq \rangle$, where

$$W = \{w_S \mid S \in Con(\phi)\}$$

$$\pi(w_S)(\psi) = \begin{cases} \textbf{true} \text{ if } \psi \in S \\ \textbf{false} \text{ if } \psi \notin S \end{cases}, \text{ for all } \psi \in \Phi$$

$w_T \preccurlyeq_i^{w_S} w_U$ if there is some $\psi \in Sub^+(\phi) \cap T \cap U$ such that $S/B_i^\psi \subseteq T$

We prove $(**)$ by induction on the structure of formulas: supposing that it holds for all subformulas of $\psi \in Sub(\phi)$, we show it holds for ψ. The cases where ψ is a primitive formula, a conjunction or a negation are straightforward. Suppose ψ is of the form $B_i^\chi \zeta$. We prove the "if" direction first, and assume that $\psi \in S$. This implies that $\zeta \in S/B_i^\chi$. Consider the set $\min_i^{w_S}\left\{[\chi]_{M_\phi} \cap W_i^{w_S}\right\}$. If this set is empty, then it follows immediately from the definition of $\vDash$ that $(M_\phi, w_S) \vDash B_i^\chi \zeta$.

Suppose then that there is some $w_T \in \min_i^{w_S}\left\{[\chi]_{M_\phi} \cap W_i^{w_S}\right\}$, i.e. $w_T \preccurlyeq_i^{w_S} w_U$ for all $w_U \in \left\{[\chi]_{M_\phi} \cap W_i^{w_S}\right\}$. Then there is some $\xi \in Sub^+(\phi) \cap T$ such that $S/B_i^\xi \subseteq T$. We must show that $\zeta \in T$. Since $S/B_i^\xi \subseteq T$, S/B_i^ξ must be a BRS-consistent set. It follows that S/B_i^χ is a BRS-consistent set too. Suppose not: then there is some finite set of formulas $F = \{\phi_1, \phi_2, \ldots, \phi_k\} \subseteq S/B_i^\chi$ such that $BRS \vdash \neg(\phi_1 \wedge \phi_2 \wedge \ldots \wedge \phi_k)$. Letting η denote $(\phi_1 \wedge \phi_2 \wedge \ldots \wedge \phi_k)$, we have:

1.	$BRS \vdash \neg\eta$	assumption
2.	$BRS \vdash \eta \Rightarrow \xi$	1, **Taut**, **MP**
3.	$BRS \vdash B_i^\chi \eta \Rightarrow B_i^\chi \xi$	2, **RE**, **Dist**, **Taut**, **MP**
4.	$BRS \vdash B_i^\chi \xi \Rightarrow \left(B_i^{\chi\wedge\xi}\eta \Leftrightarrow B_i^\chi \eta\right)$	**IE(a)**
5.	$BRS \vdash B_i^\chi \eta \Rightarrow B_i^{\chi\wedge\xi}\eta$	3, 4, **Taut**, **MP**
6.	$BRS \vdash \neg B_i^\xi \neg\chi \Rightarrow \left(B_i^{\xi\wedge\chi}\eta \Leftrightarrow \left(B_i^\xi \eta \vee B_i^\xi(\chi \Rightarrow \eta)\right)\right)$	**IE(b)**
7.	$BRS \vdash \neg B_i^\xi \neg\chi \Rightarrow \left(B_i^{\chi\wedge\xi}\eta \Leftrightarrow \left(B_i^\xi \eta \vee B_i^\xi(\chi \Rightarrow \eta)\right)\right)$	6, **LE**, **Taut**, **MP**
8.	$BRS \vdash (\chi \Rightarrow \eta) \Rightarrow \neg\chi$	1, **Taut**, **MP**
9.	$BRS \vdash B_i^\xi(\chi \Rightarrow \eta) \Rightarrow B_i^\xi \neg\chi$	8, **RE**, **Dist**, **Taut**, **MP**
10.	$BRS \vdash \left(B_i^\chi \eta \wedge \neg B_i^\xi \neg\chi\right) \Rightarrow B_i^\xi \eta$	5, 7, 9, **Taut**, **MP**
11.	$BRS \vdash \left(B_i^\chi \phi_1 \wedge \ldots \wedge B_i^\chi \phi_k \wedge \neg B_i^\xi \neg\chi\right) \Rightarrow \left(B_i^\xi \phi_1 \wedge \ldots \wedge B_i^\xi \phi_k\right)$	10, **Dist**, **Taut**, **MP**

By the hypothesis of induction, we know that $\chi \in T$, since $w_T \in [\chi]_{M_\phi}$. It follows that $\neg B_i^\xi \neg\chi \in S$ since $B_i^\xi \neg\chi \in Sub^{++}(\phi)$. Since $B_i^\chi \phi_1, \ldots, B_i^\chi \phi_k \in S$, we have $\neg B_i^\xi \phi_1, \ldots, \neg B_i^\xi \phi_k \notin S$, or else

S would be inconsistent according to line 11 above. Since $B_i^{\xi}\phi_1, \ldots, B_i^{\xi}\phi_k \in Sub^{++}(\phi)$, it follows that $B_i^{\xi}\phi_1, \ldots, B_i^{\xi}\phi_k \in S$. Thus $F \subseteq S/B_i^{\xi}$, i.e. S/B_i^{ξ} is not a BRS-consistent set, contradicting our original assumption.

So S/B_i^{χ} is a BRS-consistent set, and it therefore has a maximal BRS-consistent extension, U. And since $B_i^{\chi}\chi \in S$ (single instance of **Succ**), we have $\chi \in (S/B_i^{\chi}) \subseteq U$. Thus $w_U \preccurlyeq_i^{w_S} w_T$, by construction; and since $w_T \in \min_i^{w_S}\left\{[\chi]_{M_\phi} \cap W_i^{w_S}\right\}$, $w_T \preccurlyeq_i^{w_S} w_U$. So there is some $\rho \in Sub^{+}(\phi) \cap T \cap U$ such that $S/B_i^{\rho} \subseteq T$.

We started off by assuming that $B_i^{\chi}\zeta \in S$; we also know that $\neg B_i^{\chi}\neg\rho \in S$, since $\rho \in U$ and $S/B_i^{\chi} \subseteq U$; and that $\neg B_i^{\rho}\neg\chi \in S$, since $\chi \in T$ and $S/B_i^{\rho} \subseteq T$. Furthermore,

12.	$BRS \vdash \neg B_i^{\chi}\neg\rho \Rightarrow \left(B_i^{\chi\wedge\rho}\zeta \Leftrightarrow (B_i^{\chi}\zeta \vee B_i^{\chi}(\rho \Rightarrow \zeta))\right)$	**IE(b)**
13.	$BRS \vdash (\neg B_i^{\chi}\neg\rho \wedge B_i^{\chi}\zeta) \Rightarrow B_i^{\chi\wedge\rho}\zeta$	12, **Taut**, **MP**
14.	$BRS \vdash (\neg B_i^{\chi}\neg\rho \wedge B_i^{\chi}\zeta) \Rightarrow B_i^{\rho\wedge\chi}\zeta$	13, **LE**, **Taut**, **MP**
15.	$BRS \vdash \neg B_i^{\rho}\neg\chi \Rightarrow \left(B_i^{\rho\wedge\chi}\zeta \Leftrightarrow (B_i^{\rho}\zeta \vee B_i^{\rho}(\chi \Rightarrow \zeta))\right)$	**IE(b)**
16.	$BRS \vdash (\neg B_i^{\chi}\neg\rho \wedge B_i^{\chi}\zeta \wedge \neg B_i^{\rho}\neg\chi) \Rightarrow (B_i^{\rho}\zeta \vee B_i^{\rho}(\chi \Rightarrow \zeta))$	14, 15, **Taut**, **MP**

Since $\chi, \zeta, \rho \in Sub^{+}(\phi)$, $B_i^{\rho}\zeta$ and $B_i^{\rho}(\chi \Rightarrow \zeta)$ are both in $Sub^{++}(\phi)$. So either $B_i^{\rho}\zeta \in S$ or $B_i^{\rho}(\chi \Rightarrow \zeta) \in S$: if $\neg B_i^{\rho}\zeta \in S$ and $\neg B_i^{\rho}(\chi \Rightarrow \zeta) \in S$, S would be inconsistent according to line 16 above. But $S/B_i^{\rho} \subseteq T$, so either $\zeta \in T$, or $\chi \Rightarrow \zeta \in T$, in which case $\zeta \in T$ again because $\chi \in T$.

Thus for every $w_T \in \min_i^{w_S}\left\{[\chi]_{M_\phi} \cap W_i^{w_S}\right\}$, $\zeta \in T$, and so $(M_\phi, w_T) \models \zeta$ by the hypothesis of induction. It follows from the definition of $\models$ that $(M_\phi, w_S) \models B_i^{\chi}\zeta$.

For the "only if" direction, assume $(M_\phi, w_S) \models B_i^{\chi}\zeta$. It follows that the set $(S/B_i^{\chi}) \cup \{\neg\zeta\}$ is not BRS-consistent. If it were, it would have a maximal BRS-consistent extension T. Now, $B_i^{\chi}\chi \in S$ (single instance of **Succ**), and so $\chi \in (S/B_i^{\chi}) \subseteq T$. Thus by construction, $w_T \preccurlyeq_i^{w_S} w_U$ for all U such that $\chi \in U$, and (given the hypothesis of induction), $w_T \in \min_i^{w_S}\left\{[\chi]_{M_\phi} \cap W_i^{w_S}\right\}$. The hypothesis of induction also tells us that $(M_\phi, w_T) \models \neg\zeta$, since $\neg\zeta \in T$, and it follows immediately from the definition of $\models$ that $(M_\phi, w_S) \models \neg B_i^{\chi}\zeta$, contradicting our original assumption. So $(S/B_i^{\chi}) \cup \{\neg\zeta\}$ is not BRS-consistent. It follows from **Taut**, **Dist**, **MP** and **RE** that $B_i^{\chi}\zeta \in S$, as desired (see Fagin *et al.* [21], p.55).

So we have proved the inductive step for the case where ψ is of the form $B_i^{\chi}\zeta$. The case where ϕ is of the form $B_i\chi$ follows quickly. Pick any instance of $true \in Sub^{+}(\phi)$. Since $BRS \vdash B_i\chi \Leftrightarrow B_i^{true}\chi$ (instance of **Triv**), $B_i\chi \in S$ if and only if $B_i^{true}\chi \in S$. Furthermore, since $true$ is a propositional tautology, it follows from the definition of $\models$ that $(M_\phi, w_T) \models true$ for all w_T, i.e. $[true] = W$. Again

from the definition of $\models$, we have $(M_\phi, w_T) \models B_i\chi$ if and only if $(M_\phi, w_T) \models B_i^{true}\chi$.

We have shown that $(**)$ holds for all formulas $\psi \in Sub(\phi)$. To complete the proof of completeness, we need to show that $M_\phi \in \mathcal{M}$, i.e. that M_ϕ really is a belief revision structure.

It is clear that W is a well-defined set of possible worlds, and π is an interpretation. We need to show that for all S, $\preccurlyeq_i^{w_S}$ is complete and transitive on $W_i^{w_S}$, and is well-founded. For completeness, assume that $w_T, w_U \in W_i^{w_S}$. We need to show that either $w_T \preccurlyeq_i^{w_S} w_U$ or $w_U \preccurlyeq_i^{w_S} w_T$. From the definitions of W_i^w and $\preccurlyeq_i^{w_S}$, there is some $\psi \in Sub^+(\phi) \cap T$ such that $S/B_i^\psi \subseteq T$, and some $\chi \in Sub^+(\phi) \cap U$ such that $S/B_i^\chi \subseteq U$. Since S is a maximal BRS-consistent set, either (a) $B_i^{\psi\vee\chi}\neg\psi \in S$ or (b) $\neg B_i^{\psi\vee\chi}\neg\psi \in S$. We consider each case in turn. First (a) $B_i^{\psi\vee\chi}\neg\psi \in S$. We know that $B_i^{\psi\vee\chi}(\psi \vee \chi) \in S$ (instance of **Succ**) and

17. $BRS \vdash \left(B_i^{\psi\vee\chi}\neg\psi \wedge B_i^{\psi\vee\chi}(\psi \vee \chi)\right) \Rightarrow B_i^{\psi\vee\chi}\chi$ **Dist**
18. $BRS \vdash B_i^{\psi\vee\chi}\chi \Rightarrow \left(B_i^{(\psi\vee\chi)\wedge\chi}\zeta \Leftrightarrow B_i^{\psi\vee\chi}\zeta\right)$ **IE(a)**
19. $BRS \vdash ((\psi \vee \chi) \wedge \chi) \Leftrightarrow \chi$ **Taut**
20. $BRS \vdash B_i^{(\psi\vee\chi)\wedge\chi}\zeta \Leftrightarrow B_i^\chi\zeta$ 19, **LE**

Line 17 implies that $B_i^{\psi\vee\chi}\chi \in S$. So it follows from line 18 that $B_i^{(\psi\vee\chi)\wedge\chi}\zeta \in S$ if and only if $B_i^{\psi\vee\chi}\zeta \in S$, i.e. $S/B_i^{(\psi\vee\chi)\wedge\chi} = S/B_i^{\psi\vee\chi}$. Furthermore, line 20 implies that $S/B_i^{(\psi\vee\chi)\wedge\chi} = S/B_i^\chi$. So $S/B_i^\chi = S/B_i^{\psi\vee\chi} \subseteq U$. But $\psi \vee \chi \in T \cap U$, so $w_U \preccurlyeq_i^{w_S} w_T$.

Next (b) $\neg B_i^{\psi\vee\chi}\neg\psi \in S$. Since

21. $BRS \vdash \neg B_i^{\psi\vee\chi}\neg\psi \Rightarrow \left(B_i^{(\psi\vee\chi)\wedge\psi}\zeta \Leftrightarrow \left(B_i^{\psi\vee\chi}\zeta \vee B_i^{\psi\vee\chi}(\chi \Rightarrow \zeta)\right)\right)$ **IE(b)**

if $B_i^{\psi\vee\chi}\zeta \in S$ then $B_i^{(\psi\vee\chi)\wedge\psi}\zeta \in S$, so $S/B_i^{\psi\vee\chi} \subseteq S/B_i^{(\psi\vee\chi)\wedge\psi} = S/B_i^\psi \subseteq T$. But $\phi \vee \psi \in T \cap U$, so $w_T \preccurlyeq_i^{w_S} w_U$.

To show transitivity, suppose that for some $w_T, w_U, w_V \in W_i^{w_S}$, $w_T \preccurlyeq_i^{w_S} w_U$ and $w_U \preccurlyeq_i^{w_S} w_V$. Then there is some $\psi \in T \cap U$ such that $S/B_i^\psi \subseteq T$, and there is some $\chi \in U \cap V$ such that $S/B_i^\chi \subseteq U$. Since S is a maximal BRS-consistent set, either (a) $B_i^{\psi\vee\chi}\neg\psi \in S$ or (b) $\neg B_i^{\psi\vee\chi}\neg\psi \in S$. We consider each case in turn. (a) $B_i^{\psi\vee\chi}\neg\psi \in S$. We have just shown in part (a) above that this implies $S/B_i^\chi = S/B_i^{\psi\vee\chi} \subseteq U$. But $\psi \in U$, contradicting the assumption that $B_i^{\psi\vee\chi}\neg\psi \in S$. (b) $\neg B_i^{\psi\vee\chi}\neg\psi \in S$. We have just shown in part (b) above that this implies $S/B_i^{\psi\vee\chi} \subseteq S/B_i^\psi \subseteq T$. But $\psi \vee \chi \in T \cap U \cap V$, so $w_T \preccurlyeq_i^{w_S} w_V$ as required. Note that nothing in this proof makes use of the fact that $w_V \in W_i^w$: it follows that $\preccurlyeq_i^{w_S}$ is in fact transitive on the entire domain W.

Well-foundedness of $\preccurlyeq_i^{ws}$ follows immediately from the finiteness of W. To show that W is finite, we must show that $Con(\phi)$ is finite, since $|W| = |Con(\phi)|$. It is clear that there is a finite number maximal BRS-consistent subsets of $Sub(\phi)$, since $Sub(\phi)$ is itself a finite set. And each maximal BRS-consistent subset of $Sub(\phi)$ has a unique extension to a maximal BRS-consistent subset of $Sub^+(\phi)$: suppose S is a maximal BRS-consistent subset of $Sub(\phi)$, and $S^+ \supseteq S$ is a maximal BRS-consistent subset of $Sub^+(\phi)$; then $\psi \in S^+$ only if (a) $\psi \in S$, or (b) ψ is of the form $\neg\chi$ and $\chi \notin S^+$, or (c) ψ is of the form $\chi \wedge \zeta$ and $\chi, \zeta \in S^+$. And if there is no maximal BRS-consistent subset S of $Sub(\phi)$ such that $S \subseteq S^+$, S^+ cannot be a maximal BRS-consistent subset of $Sub^+(\phi)$. Thus there is a one-to-one correspondence between the maximal BRS-subsets of $Sub(\phi)$ and the maximal BRS-consistent subsets of $Sub^+(\phi)$. Propositional reasoning tells us that there are at most $2^{|Sub(\phi)|}$ logically distinct formulas in $Sub^+(\phi)$, corresponding to the combinations of truth-value assignments to the formulas in $Sub(\phi)$. And if ψ and χ are logically equivalent (i.e. $BRS \vdash \psi \Leftrightarrow \chi$), and S_{neg}^{++} is a maximal BRS-consistent subset of $Sub_{neg}^{++}(\phi)$, then $B_i\psi \in S_{neg}^{++}$ if and only if $B_i\chi \in S_{neg}^{++}$ (**Triv**, **RE** and **Dist**); $B_i^{\zeta}\psi \in S_{neg}^{++}$ if and only if $B_i^{\zeta}\chi \in S_{neg}^{++}$ (**RE** and **Dist**); and $B_i^{\psi}\zeta \in S_{neg}^{++}$ if and only if $B_i^{\chi}\zeta \in S_{neg}^{++}$. So each maximal BRS-consistent subset of $Sub^+(\phi)$ has a finite number of extensions to maximal BRS-consistent subsets of $Sub_{neg}^{++}(\phi)$. Thus $Con(\phi)$ is a finite set as required.■

Proposition 1

The proof of Proposition 1 follows the same steps as the proof of $(**)$ in Theorem 1, and is not repeated here.

Theorem 2

Soundness

We must check that **TPI** and **TNI** are valid with respect to $\mathcal{A}$, i.e. $\mathcal{A} \models B_i^{\phi}\psi \Rightarrow B_i^{\chi}B_i^{\phi}\psi$ and $\mathcal{A} \models \neg B_i^{\phi}\psi \Rightarrow B_i^{\chi}\neg B_i^{\phi}\psi$. (**TPI**) Suppose that $M \in \mathcal{A}$, and $(M, w) \models B_i^{\phi}\psi$. Then for every $x \in \min_i^w \{[\phi]_M \cap W_i^w\}$ we have $(M, x) \models \psi$. Now for every $y \in \min_i^w \{[\chi]_M \cap W_i^w\}$, $z \preccurlyeq_i^y u$ if and only if $z \preccurlyeq_i^w u$, so $\min_i^w \{[\phi]_M \cap W_i^w\} = \min_i^y \{[\phi]_M \cap W_i^y\}$. It follows that $(M, y) \models B_i^{\phi}\psi$, and therefore $(M, w) \models B_i^{\chi}B_i^{\phi}\psi$. (**TNI**) Suppose that $M \in \mathcal{A}$, and $(M, w) \models \neg B_i^{\phi}\psi$. Then there is some $x \in \min_i^w \{[\phi]_M \cap W_i^w\}$ such that $(M, x) \not\models \psi$. Now for every $y \in \min_i^w \{[\chi]_M \cap W_i^w\}$, $z \preccurlyeq_i^y u$ if and

only if $z \preccurlyeq_i^w u$, so $\min_i^w \{[\phi]_M \cap W_i^w\} = \min_i^y \{[\phi]_M \cap W_i^y\}$. It follows that $(M, y) \models \neg B_i^\phi \psi$, and therefore $(M, w) \models B_i^\chi \neg B_i^\phi \psi$.

COMPLETENESS

To prove completeness, we must show that every $BRSI$-consistent formula in $\mathcal{L}$ is satisfiable with respect to $\mathcal{A}$. We proceed in the same way as in the proof of Theorem 1, and construct a structure $M_\phi \in \mathcal{A}$ for every formula $\phi \in \mathcal{L}$. The construction of M_ϕ is exactly the same as before, except that the set $Sub^{++}(\phi)$ is enlarged: $Sub^{++}(\phi)$ is the smallest set of formulas such that (a) if $\psi, \chi \in Sub^{+}(\phi)$ then $\psi, B_i\psi, B_i^\chi \psi \in Sub^{++}(\phi)$; (b) if $\xi \in Sub^{+}(\phi)$ and $B_i^\chi \psi \in Sub^{++}(\phi)$, then $B_i^\xi B_i^\chi \psi \in Sub^{++}(\phi)$ and $B_i^\xi \neg B_i^\chi \psi \in Sub^{++}(\phi)$. The proof that $(M_\phi, w_S) \models \psi$ if and only if $\psi \in S$ is unaffected, as is the proof that $\preccurlyeq_i^{w_S}$ is complete and transitive on $W_i^{w_S}$ for all S. Finiteness of W still holds as well, since for every formula of the form $B_i^\zeta B_i^\xi \ldots B_i^\chi \psi \in S$ if and only if $B_i^\xi \ldots B_i^\chi \psi \in S$, given **TPI** and **TNI**. It remains to show that M is absolute.

Suppose that $w_T \in W_i^{w_S}$. Then there is some $\psi \in Sub^{+}(\phi) \cap T$ such that $S/B_i^\psi \subseteq T$. We must show that $w_U \preccurlyeq_i^{w_S} w_V$ if and only if $w_U \preccurlyeq_i^{w_T} w_V$. If $w_U \preccurlyeq_i^{w_S} w_V$, then there is some $\chi \in Sub^{+}(\phi) \cap U \cap V$ such that $S/B_i^\chi \subseteq U$. Suppose that $B_i^\chi \zeta \notin S$. Then $\neg B_i^\chi \zeta \in S$, and since $BRSI \vdash \neg B_i^\chi \zeta \Rightarrow B_i^\psi \neg B_i^\chi \zeta$ (instance of **TNI**), we also have $B_i^\psi \neg B_i^\chi \zeta \in S$. But $S/B_i^\psi \subseteq T$, so $\neg B_i^\chi \zeta \in T$, and $B_i^\chi \zeta \notin T$. Thus $T/B_i^\chi \subseteq S/B_i^\chi \subseteq U$, and $w_U \preccurlyeq_i^{w_T} w_V$, as required. If $w_U \preccurlyeq_i^{w_T} w_V$, then there is some $\chi \in Sub^{+}(\phi) \cap U \cap V$ such that $T/B_i^\chi \subseteq U$. Suppose that $B_i^\chi \zeta \in S$. Since $BRSI \vdash B_i^\chi \zeta \Rightarrow B_i^\psi B_i^\chi \zeta$ (instance of **TPI**), we also have $B_i^\psi B_i^\chi \zeta \in S$. But $S/B_i^\psi \subseteq T$, so $B_i^\chi \zeta \in T$. Thus $S/B_i^\chi \subseteq T/B_i^\chi \subseteq U$, and $w_U \preccurlyeq_i^{w_S} w_V$, completing the proof.■

THEOREM 3

SOUNDNESS

We must check that **WCon** is valid with respect to $\mathcal{I}$, i.e. $\mathcal{I} \models \phi \Rightarrow \neg B_i^\phi false$. Suppose that $M \in \mathcal{I}$, and $(M, w) \models \phi$, i.e. $w \in [\phi]_M$ By the inclusion assumption, $w \in W_i^w$, so $[\phi]_M \cap W_i^w \neq \emptyset$, and by well-foundedness of $\preccurlyeq_i^w$, $\min_i^w \{[\phi]_M \cap W_i^w\} \neq \emptyset$. So there is some world $x \in \min_i^w \{[\phi]_M \cap W_i^w\}$. Since it is not the case that $(M, x) \models false$, it is not the case that $(M, w) \models B_i^\phi false$, and it follows from the definition of $\models$ that $(M, w) \models \neg B_i^\phi false$.

COMPLETENESS

The proof for completeness is the same as for Theorem 1, except that it must also be shown $M_\phi \in \mathcal{I}$, i.e. that every plausibility ordering $\preccurlyeq_i^{w_S}$ in M_ϕ satisfies the inclusion assumption. Let $\phi_1, \ldots, \phi_n$ be an enumeration of the formulas in $Sub(\phi)$, and for some $S \in Con(\phi)$ let $\phi_1' = \phi_1$ if $\phi_1 \in S$, and $\phi_1' = \neg\phi_1$ if $\phi_1 \notin S$. Propositional reasoning implies that $\phi_1' \wedge \ldots \wedge \phi_n' \in S$, and since $BRSC \vdash (\phi_1' \wedge \ldots \wedge \phi_n') \Rightarrow \neg B_i^{\phi_1' \wedge \ldots \wedge \phi_n'} false$ (instance of **WCon**), we have $\neg B_i^{\phi_1' \wedge \ldots \wedge \phi_n'} false \in S$ for some $false \in Sub^+(\phi)$. It follows that $S/B_i^{\phi_1' \wedge \ldots \wedge \phi_n'}$ is a $BRSC$-consistent set. Furthermore, since $BRS \vdash B_i^{\phi_1' \wedge \ldots \wedge \phi_n'} \phi_1' \wedge \ldots \wedge \phi_n'$ (instance of **Succ**), $\phi_1' \wedge \ldots \wedge \phi_n' \in S/B_i^{\phi_1' \wedge \ldots \wedge \phi_n'}$. Now suppose that $B_i^{\phi_1' \wedge \ldots \wedge \phi_n'} \psi \in S$. Then $\psi \in S/B_i^{\phi_1' \wedge \ldots \wedge \phi_n'}$ and $\psi \in Sub^+(\phi)$. Since $Sub^+(\phi)$ consists only of formulas in $Sub(\phi)$ and their conjunctions and negations, it follows from propositional reasoning that either $BRSC \vdash (\phi_1' \wedge \ldots \wedge \phi_n') \Rightarrow \psi$ or $BRSC \vdash (\phi_1' \wedge \ldots \wedge \phi_n') \Rightarrow \neg\psi$. But $S/B_i^{\phi_1' \wedge \ldots \wedge \phi_n'}$ is a $BRSC$-consistent set, so we must have $BRSC \vdash (\phi_1' \wedge \ldots \wedge \phi_n') \Rightarrow \psi$, and therefore $\psi \in S$. We have shown that $S/B_i^{\phi_1' \wedge \ldots \wedge \phi_n'} \subseteq S$. The definition of $\preccurlyeq_i^{w_S}$ implies that $w_S \preccurlyeq_i^{w_S} w_S$, and so $w_S \in W_i^{w_S}$ as required.■

THEOREM 4

Soundness follows immediately from Theorem 2 and Theorem 3. To prove completeness, we follow the construction of M_ϕ described in the proof of Theorem 2, and the same steps imply that $M_\phi \in \mathcal{A}$. The completeness part of the proof of Theorem 3 can then be followed to show that $M_\phi \in \mathcal{I}$.■

PROPOSITION 2

First suppose that $x \in \{x \mid x \trianglelefteq_i w \text{ or } w \trianglelefteq_i x\}$. Then either (a) $x \trianglelefteq_i w$, in which case it follows immediately from the definition of $\trianglelefteq_i$ that $x \preccurlyeq_i^w w$, as required; or (b) $w \trianglelefteq_i x$; from the definition of $\trianglelefteq_i$ we have $w \preccurlyeq_i^x x$; $x \preccurlyeq_i^x y$ for some y from **R4**, and so $x \preccurlyeq_i^w y$ from **R3**, as required.

Now suppose that $x \in \{x \mid x \preccurlyeq_i^w y \text{ for some } y\}$. We know from **R4** that $w \preccurlyeq_i^w z$ for some z. It follows from **R1** that either $x \preccurlyeq_i^w w$, in which case $x \trianglelefteq_i w$ by definition; or $w \preccurlyeq_i^w x$, in which case **R3** gives us $w \preccurlyeq_i^x x$ as so $w \trianglelefteq_i x$. In both cases $x \in \{x \mid x \trianglelefteq_i w \text{ or } w \trianglelefteq_i x\}$ as required.■

Proposition 3

First suppose $x \in \min_i (X \cap W_i^w)$. Then $x \trianglelefteq_i y$ for all $y \in X \cap W_i^w$, and from the definition of $\trianglelefteq_i$, $x \preccurlyeq_i^y y$ for all $y \in X \cap W_i^w$. It follows from **R3** that $x \preccurlyeq_i^w y$ for all $y \in X \cap W_i^w$ and so $x \in \min_i^w (X \cap W_i^w)$.

Now suppose that $x \in \min_i^w (X \cap W_i^w)$. Then $x \preccurlyeq_i^w y$ for all $y \in X \cap W_i^w$, and from **R3** we have $x \preccurlyeq_i^y y$ for all $y \in X \cap W_i^w$. The definition of $\trianglelefteq_i$ gives us $x \trianglelefteq_i y$ for all $y \in X \cap W_i^w$, and so $x \in \min_i (X \cap W_i^w)$.■

Theorem 5

The proof of Theorem 5 follows the same steps as the proof of Theorem 3.3.1 in Fagin *et al.* [21], and is omitted.

Chapter 2

Algorithmic Characterization of Rationalizability in Extensive Form Games*

Abstract: We construct a dynamic epistemic model for extensive form games, which generates a hierarchy of beliefs for each player over her opponents' strategies and beliefs, and tells us how those beliefs will be revised as the game proceeds. We use the model to analyze the implications of the assumption that the players possess common (true) belief in rationality, thus extending the concept of rationalizability to extensive form games.

1 Introduction

This paper seeks to examine the implications of common belief in rationality in extensive form games. It was once thought that the notorious backward induction argument provided a precise characterization of these implications, at least in games of perfect information. The implausibility of the backward induction outcome in games such as the repeated prisoner's dilemma was attributed to the strength of the assumptions made. And once the assumption of common belief (or knowledge) is relaxed even a little, Kreps *et al.* [18] showed that cooperation until the final rounds can become a rational response. But later work questioned the very validity of the backward induction argument. Binmore [8], Pettit and Sugden [22] and Reny [23] were among the first to take this line, and argued that, even if there is common belief in rationality at the beginning of an extensive form game, there may not be at each of the subsequent information sets. Indeed, backward induction typically implies that certain information sets will not be reached. If the backward induction argument is correct, these information sets are therefore not consistent with common belief in rationality. But the argument assumes that there is common belief in rationality at *every* information set in the game. The lesson to be learned from this resolution of the backward induction paradox is that analysis of rational play in extensive form games requires careful consideration not just of the players' beliefs at the beginning of the game, but also of how these beliefs change as the game progresses.

*This paper is a much revised version of Board [11]. Helpful comments from Michael Bacharach, Paolo Battigalli, Adam Brandenburger, Amanda Friedenberg, Matthias Hild and Bob Stalnaker are gratefully acknowledged.

Unlike the backward induction argument, however, most solution concepts in game theory make no explicit reference to players' rationality or beliefs. Nash equilibrium, for instance, is defined purely is terms of conditions on the player's strategy sets. A notable exception is the notion of rationalizability, developed by Bernheim [7] and Pearce [21]. A strategy is said to be rationalizable if it is consistent with common belief in rationality. The idea that game theoretic solution concepts could be characterized epistemically (that is, by a set of restrictions on the players' beliefs and behavior) was developed further by Aumann [2], who showed that rational players with a common prior over the space of uncertainty will play according to a correlated equilibrium distribution; and that every correlated equilibrium distribution is consistent with rationality of the players and the common prior assumption. Aumann used his *information partition* model (Aumann [1]) to provide a precise description of each players' beliefs about the game (i.e. about which strategies would be played), and about each other. Although Bernheim and Pearce did not employ any such formal model of interactive epistemology, the results of their analysis of strategic form games were later proved by Tan and Werlang [28] and Stalnaker [24] in the context of such a model.

But the information partition model of Aumann and the alternative, hierarchical, model of interactive epistemology used by Tan and Werlang (see e.g. Mertens and Zamir [19] and Brandenburger and Dekel [9]) are static: they tell us what each player believes about her opponents' beliefs, but they cannot tell us what she will believe at future information sets, or what she believes her opponents' will believe. Hence they are not rich enough to analyze rational play in extensive form games. Dynamic models have been developed to serve precisely this purpose, most notably by Battigalli and Siniscalchi [5] and Stalnaker [25] (see also Board [12]). In this paper we use such a model to give a precise characterization of rationalizability in extensive form games.

Section 2 gives a brief discussion of related literature. In section 3 we develop the formal model of beliefs in extensive form games, and in section 4 we use that model to give a precise characterization of the implications of common belief in rationality, in terms of an iterated deletion algorithm. Section 5 concludes.

2 Related literature

Most closely related to the current project is the work of Stalnaker [25] and [26], who uses a model of belief revision which is a special case of that presented here. It is shown in Board [12]

that Stalnaker's model makes rather strong introspection assumptions which we do not require[1]. Furthermore, Stalnaker analyzes only strategic form games. On the other hand, he uses the belief revision component to examine several alternative notions of rationality in addition to the basic notion we consider. These stronger notions pick up elements of extensive form reasoning even in the strategic form of the game.

Battigalli and Siniscalchi [6] use a very different kind of model, built up from infinite hierarchies of conditional probability systems. A *conditional probability system* describes a player's beliefs at each stage of an extensive form game, and each level of the hierarchy describes the player's beliefs about every level beneath it. Thus their hierarchical structures provide an explicit model of beliefs and beliefs about beliefs throughout the game tree. They use the structures to investigate the implications of common belief, and of a stronger concept, common *strong* belief, in rationality. Unlike our paper, they consider incomplete information games, where the players may be uncertain of each other's payoffs. But they restrict their attention to games with *observable actions*, where at each stage everyone observes the actions of the previous stage. This assumption is for the sake of tractability, and not imposed by any limitations of their model.

Brandenburger and Keisler [10] use a similar model to Battigalli and Siniscalchi, with *lexicographic probability systems* playing the role of conditional probability systems. A lexicographic probability system is a (finite) sequence of probability measures. Like Stalnaker, they focus on the strategic form of the game, and derive epistemic conditions for iterated deletion of weakly dominated strategies. But their results also shed light on the extensive form procedures of backward and forward induction.

Feinberg [15] and [16] develops a rich language which can be employed to describe what he calls 'subjective' reasoning in extensive form games, and also to describe the structure of the game itself, including payoffs. An system of axioms is used to prove theorems in the language, and semantic structures provide truth conditions. A player is represented by a different hypothetical identity at every information set at which she is on move. Belief is a property of these identities, and only implicitly of players. And beliefs of a player's future identities are not derived from those of her past identities: there is no belief revision component to the logic. Feinberg uses his framework to analyze backward and forward induction, as well as provide epistemic characterizations of Nash equilibrium and sequential equilibrium and to introduce a new concept, the *reasonable solution* of a game.

[1] We conjecture that they are not required for Stalnaker's results either.

For a more detailed discussion of some of these papers and comprehensive surveys of many earlier results, see Dekel and Gul [14] and Battigalli and Bonanno [4].

3 Beliefs in extensive form games

The analysis of this paper is restricted to finite extensive form games of complete information and perfect recall. The description of such a game specifies the following five elements (see also Osborne and Rubinstein [20]):

- a finite set N of players.

- a finite set H of sequences, which satisfies: (i) $\varnothing \in H$; and (ii) if $\left(a^k\right)_{k=1,...,K} \in H$ and $L < K$, then $\left(a^k\right)_{k=1,...,L} \in H$. Each $h \in H$ is a *history*, and each component of h is an *action* taken by a player. The set of histories defines the game tree, with each element h representing a node of the tree, the node that is reached if that history is played. A history $\left(a^k\right)_{k=1,...,K} \in H$ is *terminal* if there is no a^{K+1} such that $\left(a^k\right)_{k=1,...,K+1} \in H$. The set of actions available after the nonterminal history h is denoted $A(h) = \{a : (h,a) \in H\}$, and the set of terminal histories is denoted Z.

- a function $\imath : H \setminus Z \to N$ that assigns to each nonterminal history the player whose turn to move it is.

- a partition $\mathcal{I}$ of $H \setminus Z$ that divides all the nonterminal histories into *information sets*. The cell $\mathcal{I}(h)$ of $\mathcal{I}$ that contains h identifies the nonterminal histories that the player on move cannot distinguish from h based on the information available to her at h. It is required that for every history in a given cell of the partition, the same player is on move and the same actions are available, i.e. if $h' \in \mathcal{I}(h)$, then $\imath(h) = \imath(h')$ and $A(h) = A(h')$. This is implied by the fact that each player knows when it is her turn to move, and what actions are available to her. Thus for any information set $I \in \mathcal{I}$ we can write $\imath(I)$ for the player on move, and A_I for the actions available to her, and we can partition $\mathcal{I}$ into sets $\mathcal{I}_i = \imath^{-1}(i)$.
  To characterize perfect recall, let $X_i(h)$ denote player i's *experience* at a given history h. $X_i(h)$ is the sequence of information sets that player i encounters in the history h and the actions she takes at them, in the order that these events occur. For each player i, if $h, h' \in I$ for some $I \in \mathcal{I}_i$, then $X_i(h) = X(h')$.

- a utility function $U_i : Z \to \mathbb{R}$ for each player i, which assigns an expected utility value to each terminal history.

The collection $\langle N, H, \imath, \mathcal{I}, (U_i)_{i \in N} \rangle$ defines an extensive form game, Γ.

It will be convenient to use the following additional notation. Let $A_{\mathcal{I}} = \times_{I \in \mathcal{I}} A_I$ be the set of *action profiles*, which specify an action $a_I \in A_I$ for every information set $I \in \mathcal{I}$, and let A_{-I} be the set of action profiles at every information set other than I (so that $A_I \times A_{-I} = A_{\mathcal{I}}$). For a given action profile $a_{\mathcal{I}} \in A_{\mathcal{I}}$, let $h(a_{\mathcal{I}})$ be the history *induced by* $a_{\mathcal{I}}$, i.e. $h(a_{\mathcal{I}})$ is a sequence of actions of the form $(a_{\mathcal{I}}(\mathcal{I}(\varnothing)), a_{\mathcal{I}}(\mathcal{I}(a_{\mathcal{I}}(\varnothing))), \ldots)$ such that $h(a_{\mathcal{I}}) \in Z$. We can write $u_i(a_{\mathcal{I}}) = U_i(h(a_{\mathcal{I}}))$, where u_i is player i's strategic form utility function. Finally, for a given information set I, let $A_{\mathcal{I}}(I)$ be the set of action profiles *consistent with* I, i.e. $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ if there is some sequence of actions $(a_{\mathcal{I}}(\mathcal{I}(\varnothing)), a_{\mathcal{I}}(\mathcal{I}(a_{\mathcal{I}}(\varnothing))), \ldots) \in I$.

As we discussed in the introduction, in order to analyze rational play in extensive form games, it is crucial to have a precise model not only of the players' beliefs but also of the way these beliefs are revised as the game proceeds. Traditional theories of belief revision, such as Bayes' rule, have concentrated on modeling how beliefs change when new information is learned that is compatible with one's existing beliefs. But such a focus is too narrow for our purposes: in order to model counterfactual reasoning in games, we will need to know how beliefs change or would change in the event of surprises, when information is learned that contradicts what is currently believed. In this case, some of these existing beliefs must be given up, and the problem is that there is a multitude of ways to select just how this should be done.

Board [12] develops a multi-agent logic of belief revision; the language of that logic can be used to describe players' beliefs in extensive form games. We start with a set of primitive formulas, $\Phi = \{a_I \mid a_I \in A_I \text{ for some } I \in \mathcal{I}\}$. The primitive formulas describe which actions are taken at each information set, so that a_I denotes the sentence "action a_I is chosen at information set I"[2]. The language $\mathcal{L}$ is the smallest set of formulas such that:

(*a*) if $\phi \in \Phi$, then $\phi \in \mathcal{L}$;

(*b*) if $\phi, \psi \in \mathcal{L}$, then $\neg\phi \in \mathcal{L}$ and $\phi \wedge \psi \in \mathcal{L}$;

(*c*) if $\phi, \psi \in \mathcal{L}$, then $B_i\phi \in \mathcal{L}$, $C\phi \in \mathcal{L}$ and $B_i^{\phi}\psi \in \mathcal{L}$, for $i \in N$.

[2] Or "action a_I *was / will be / would have been / would be* chosen at information set I". There is no notion of time in our logic, so sentences should be interpreted as past, present or future, indicative or subjunctive depending on the viewpoint.

With slight abuse of notation, we shall use a_I to denote the sentence "action profile a_I is chosen", and I to denote the sentence "information set I is reached". Formally, a_I and I are abbreviations for longer sentences containing only primitive formulas, negations and conjunctions. B_i represents player i's beliefs before the start of the game, and B_i^ϕ her beliefs after she learns that ϕ is the case. Finally, C is the common (prior) belief operator.

Truth conditions are assigned to the formulas of $\mathcal{L}$ by means of a *model*. A model M for an extensive form game Γ is a tuple $\langle W, f, \preccurlyeq \rangle$ where

- W is a non-empty set of *possibles worlds*;
- $f : W \to A_I$ is an *action function*;
- $\preccurlyeq$ is a vector of *plausibility orderings*, one for each player at every world.

Models work in the same way as the belief revision structures used in Board [12]. The action function plays the role of the interpretation in a belief revision structure, and specifies, for each world, which action will (or would) be taken at every information set in the game. We shall use $f_I(w)$ to denote the action taken at information set I in world w. The structure of the game implies that $f_I(w) \in A_I$, for all I, w. Note that none of the facts about the structure of the game are included in the model. The implication is that all these facts are true at every world in the model (and hence are common belief among the players). This corresponds to the assumption that the game is one of complete information.

$\preccurlyeq_i^w$ denotes the plausibility ordering of player i at world w, and encodes her beliefs and her belief revision policy. $x \preccurlyeq_i^w y$ means that from the point of view of player i at world w, world x is at least as plausible as world y. Intuitively, the player considers possible only the worlds which are most plausible according to her ordering: we call these worlds *accessible*; the remainder of the ordering is used to construct her revised beliefs, as we shall see. We impose two constraints on the form of the $\preccurlyeq_i^w$ relations. Let $W_i^w = \{x \mid x \preccurlyeq_i^w y \text{ for some } y\}$; W_i^w is the set of worlds which are *conceivable* to i at world w, though not necessarily accessible. Then, we assume that:

R1 for all i, w: $\preccurlyeq_i^w$ is complete and transitive on W_i^w;

R2 for all i, w: $\preccurlyeq_i^w$ is well-founded.

R1 ensures that each plausibility ordering divides all the worlds into ordered equivalence classes; the inconceivable worlds, i.e. those not in W_i^w, are a class unto themselves and are to be considered

least plausible. If $\preccurlyeq_i^w$ is well-founded (**R2**), then there are no infinitely descending sequences of the form $\ldots w_n \prec_i^w \prec w_{n-1} \prec_i^w \ldots \prec_i^w w_0$ (where $x \prec_i^w y$ if and only if $x \preccurlyeq_i^w y$ and not $y \preccurlyeq_i^w x$). This guarantees that for every nonempty set $X \subseteq W_i^w$, $\min_i^w (X \cap W_i^w) \neq \emptyset$, where $\min_i^w$ is defined in the obvious way (i.e. $\min_i^w (X) = \{x \in X \mid \text{for all } y \in X,\ x \preccurlyeq_i^w y\ \}$); intuitively, it says that if there are any conceivable worlds in a certain set, then there is a most plausible world in that set. Well-foundedness is satisfied automatically in the case where W is finite. Henceforth we shall assume that all models satisfy **R1** and **R2**.

The model of the game allows us to assign truth conditions to every formula in the language. Let $[\phi]$ denote the set of worlds at which ϕ is true. Truth is assigned to primitive formulas as follows: $[a_I] = \{w \mid f_I(w) = a\}$. Negations and conjunctions are dealt with in the obvious way: $[\neg\phi] = [W \setminus \phi]$ and $[\phi \wedge \psi] = [\phi] \cap [\psi]$. $B_i\phi$ is true precisely if ϕ is true at every world w accessible to i before she learns anything: $[B_i\phi] = \{w \mid \min_i^w (W_i^w) \subseteq [\phi]\}$; and $B_i^\phi\psi$ is true precisely if ψ is true at every world accessible to her after she learns that ϕ: $\left[B_i^\phi\psi\right] = \{w \mid \min_i^w ([\phi] \cap W_i^w) \subseteq [\psi]\}$. Finally, to define the truth conditions for $C\phi$, let $E\phi$ abbreviate $\bigwedge_{i \in N} B_i\phi$, let $E^0\phi$ abbreviate ϕ, and let $E^k\phi$ abbreviate $EE^{k-1}\phi$ for $k = 1, 2, \ldots$. Then $[C\phi] = \bigcap_{k=1,2,\ldots} \left[E^k\phi\right]$.

There is, however, a problem with this account of belief revision: the method just described calculates each player's beliefs at a given information set by revising her original beliefs (as represented by $\preccurlyeq_i^w$) with the information that the information set has been reached. But a given history may pass through several information sets of the player, and beliefs should be revised at each information set. There is in general no guarantee that the beliefs generated by a sequence of such revisions will be the same as the beliefs generated by revising just once. But in games of perfect recall this may be a reasonable assumption to make. In such games, the information received by a given player as the game progresses has a particular property: each new piece of information implies all of the previous pieces. If a history passes through more than one information set of a given player, these informations sets can be strictly ordered in terms of precedence, and the set of histories consistent with a given information set is always a subset of those consistent with every previous information set. And if ψ logically implies ϕ, it may be reasonable to assume that learning ϕ and then ψ will generate the same beliefs as if one learns ψ at first: in both cases the same information is learned. This simplifying assumption saves us the trouble of dealing with iterated belief revisions. Whether the single-revision process is appropriate for modeling beliefs at information sets in games of imperfect recall is an open question and beyond the scope of this paper.

The results of Board [12] give us a precise understanding of the formal language $\mathcal{L}$: we can provide an axiomatic characterization of the formulas which are true at every world of every model of a particular game. Theorem 5 of Board [12] states that the axiom system BRS^C is sound and complete with respect to the class of all belief revision structures which satisfy **R1** and **R2**, i.e. a formula is true at every world of every such belief revision structure if and only if it is provable in BRS. But the models described above are more restrictive than belief revision structures: unlike the interpretation of a belief revision structure, the action function used here to tell us which actions are played at each information set cannot assign arbitrary truth values to primitive formulas. One and only one action must be chosen at each information set, so that if $w \in [a_I]$ it must be the case that $w \in [\neg a'_I]$ for all $a'_I \neq a_I$. To provide a syntactic counterpart of this semantic restriction we add the axiom **Game**, which tells us which combinations actions are consistent with the rules of the game. For example, for a game with only four possible action profiles, $a^1_{\mathcal{I}}, a^2_{\mathcal{I}}, a^3_{\mathcal{I}}, a^4_{\mathcal{I}}$, **Game** would be $a^1_{\mathcal{I}} \vee a^2_{\mathcal{I}} \vee a^3_{\mathcal{I}} \vee a^4_{\mathcal{I}}$. BRS^C + **Game** is sound and complete with respect to the class of all models satisfying **R1** and **R2**.

4 Rationalizability

To characterize rationality in extensive form games, we must compute the players' beliefs at each information set at which they are on move. The information they learn as the game progresses is given by the information structure of the game, as specified by the information sets $\mathcal{I}$. Specifically, at information set $I \in \mathcal{I}_i$, player i learns that she must be at one of the histories in I, i.e. that one of the action profiles in $A_{\mathcal{I}}(I)$ has been chosen.

But to make sense of the definition of rationality given below, we must also make sure that each player has true belief at a given information set about what action she is choosing at that information set (see Board [13] for a more detailed discussion of this point). There are two ways of doing this: the first is to add an additional constraint to the models: for all i, if $I \in \mathcal{I}_i$ then $\min_i^w([I] \cap W_i^w) \subseteq [f_I(w)]$. This constraint says that at every world player i considers possible when she learns that information set I has been reached, her action at that information set is the same as it is in the actual world. The syntactic counterpart is the axiom schema $a_I \Rightarrow B_i^I a_I$. A problem with this approach is that it imposes restrictions not only on the beliefs at information set I, but also at beliefs prior to that. To see why, suppose that a player is moving at two successive information sets, I_1 and I_2, and that she chooses action a_{I_1} and a_{I_2} respectively, with a_{I_1} leading to

I_2. At I_1 she is assumed to believe (correctly) that she is choosing a_{I_1}. So she learns nothing when I_2 is reached, and hence her beliefs do not change. But at the second information set she assumed to believe (again correctly) that she is choosing a_{I_2}. It follows that she must have already believed this at I_1! More generally, the implication is that players must have true beliefs about their actions at every future information set compatible with their current beliefs. To put it another way, they are not allowed to change their minds unless they are surprised. Of course, this may be a reasonable assumption to make in many (or even most) circumstances, but it is not good modeling practice to hide such an assumption in the formalism.

For this reason, we adopt the second approach, and assume that the a player learns what action she will choose at a given information set when that information set is reached. Of course we are not suggesting that the player is told what to do, but rather that she does not necessarily know what she is going to do until required to make the choice. The B_i^{ϕ} operators represent the player's beliefs after deliberation[3], when the player has figured out what she will do, but we do not want encode the outcome of this deliberation process into the prior beliefs. According to this second approach, player i's beliefs at any information set $I \in \mathcal{I}_i$ at which she is on move are therefore given by $B_i^{I \wedge a_I}$, where a_I is the action she chooses at I. In terms of the model of the game, i learns that the true world must lie in the set $[I] \cap [a_I]$. The set of worlds accessible to her at world w after receiving this information is obtained by taking the $\preccurlyeq_i^w$-minimal worlds in $[I] \cap [f_I(w)] \cap W_i^w$.

To define rationality in the standard way, as expected utility maximization, we must first explain how each agent's probabilistic beliefs are derived. Given a (prior) probability measure p_i on the set of worlds W, define the conditional probability measure $p_{i,I}^w$ as follows: for any $E \subseteq W$,

$$p_{i,I}^w(E) = \frac{p_i(E \cap \min_i^w([I] \cap [f_I(w)] \cap W_i^w))}{p_i(\min_i^w([I] \cap [f_I(w)] \cap W_i^w))}.$$

$p_{i,I}^w$ is obtained from p_i by conditioning player i's information, since $\min_i^w([I] \cap [f_I(w)] \cap W_i^w)$ is the set of worlds which agent i considers possible in world w at information set I. The probability which player i assigns to any formula $\phi \in \mathcal{L}$ in world w at information set $I \in \mathcal{I}_i$ is given by $p_{i,I}^w([\phi])$. Of course, this expression may not be well defined, since there is nothing to guarantee that the denominator is greater than zero. To avoid technical issues that are not relevant for the ongoing discussion, we shall simply assume that in such a case the player is not rational. In effect, we are claiming that rational players should not rule out any information sets *a priori* (though

[3]See Aumann [2] (p. 8) for a detailed discussion of this point.

they may certainly do so once the game is in progress).

An action profile is rationalizable if it is consistent with common knowledge of rationality among the players. We build up the definition of rationalizability in several stages. First, an action is defined as rational if it maximizes the expected utility of the player who takes that action.

Definition 1 *Suppose* $I \in \mathcal{I}_i$. $a_I \in A_I$ *is* rational with respect to p_i at world w *if* $p_{i,I}^w$ *is well defined and*

$$\sum_{a_{-I} \in A_{-I}} p_{i,I}^w \left([a_{-I}] \right) . u_i \left(a_I, a_{-I} \right) \geq \sum_{a_{-I} \in A_{-I}} p_{i,I}^w \left([a_{-I}] \right) . u_i \left(a'_I, a_{-I} \right)$$

for all $a'_I \in A_I$.

There is an important difference between this notion of rationality at an information set and the concept employed elsewhere in the literature. It is usually assumed that strategies[4] rather than actions are the objects of choice, and hence the objects of rationality. A strategy is said to be rational at a given information set if it yields the highest expected utility of all those strategies which are consistent with that information set's being reached. But it is unclear when if ever players will actually make a choice between the various strategies available to them. Although we could think of a hypothetical pre-play stage when such choices are made, it seems more appropriate and more accurate to think of the players as making their choices as and when they are on move. Indeed, this is the approach that majority of the work in this area takes[5]. And at each information set a player chooses only part of her strategy, the part which specifies what she does at that information set. It is these choices that should be assessed as rational. For assessing the entire strategy at a particular information set carries with it the substantive assumption that the player on move has control over her choices at all future information sets. To see why, suppose we say that a particular strategy choice (rather than just the action choice) is rational at some information set. Presumably we mean that, among the strategies that are consistent with that information set's being reached, the strategy chosen maximizes expected utility[6]. All of these strategies specify what actions will be taken at future information sets as well as at the current information set. If the player cannot control what she does at these information sets while on move at the current information set, then

[4] Or sometimes *plans of action*, which specify actions only at nodes not ruled out by the player's previous actions. See e.g. Reny [23].

[5] A notable exception is the work of Stalnaker: he discusses this issue in [27] (p. 315), and shows that, under certain assumptions, the two approches are equivalent.

[6] See e.g. Gul [17] "... rational players choose strategies s_i such that s_i is optimal at [an information set] against some conjecture that reaches [that information set] whenever s_i reaches [that information set]" (p. 15). The majority of papers in the Bayesian tradition adopt a similar definition of rationality.

she cannot choose among these strategies. This assumption of self control does not follow from rationality alone; rationality alone does not even imply that a player *knows* what she will do at future information sets![7] The example in Figure 2 below may shed further light on this issue.

Next, we define what it is for a player to be rational. It is not immediately clear how to do this. In particular, we can distinguish *reached-node rationality*, where a player is rational if each action she actually plays is rational; *own-node rationality*, where a player is rational if each of her actions at nodes not ruled out by her previous behavior is rational; and *all-node rationality*, where a player is rational only if her actions are rational at every information set at which she is on move. Reached-node rationality does not seem strong enough, especially if we are thinking about what it means for one player to know that another is rational. Suppose for instance that the second player does not get a chance to move because of an action taken by the first. This makes the second player (vacuously) reached-node rational. And yet intuitively we would expect the first player to be able to make inferences about what the second would do if given the chance to move. For a similar reason own-node rationality will not suffice either. If a player believes herself to be rational, this ought to impose restrictions on what she believes she will or would do at future nodes. But own-node rationality says nothing about her behavior at those nodes which are ruled out by her past actions[8]. Thus in what follows we shall adopt the concept of all-node rationality, and say that a player is rational if she chooses actions that are rational at every information set at which she is on move:

Definition 2 *Player i is* rational at world w *if there is some p_i such that, for all $I \in \mathcal{I}_i$, $f_I(w)$ is rational with respect to p_i at world w.*

Let Rat_i denote the sentence "player i is rational"; Rat_i is true at world w precisely if player i is rational at world w. Note that we are not introducing any new formulas into the language: whether or not a player is rational is determined completely by her choice of actions and her first-order beliefs, i.e. her beliefs about which actions will be chosen. Thus Rat_i is simply an abbreviation for a long formula of the language. Let $Rat = \bigwedge_{i \in N} Rat_i$, and $CTBR = C(Rat) \wedge Rat$ ($CTBR$ stands for "there is common true belief in rationality").

---

[7] Nor even at past information sets: in games of imperfect recall, players can forget their previous action choices.

[8] The importance of the distinction between own-node rationality and all-node rationality arises only because we take actions as the fundamental objects of choice. If players are assumed to choose between the strategies (or plans of action) available to them at a given node, it is specified what they will do at all relevant future nodes. In this case the two concepts yield path-equivalent results.

A action profile is said to be rationalizable if it is consistent with common true belief in rationality. Formally,

Definition 3 *For any game* Γ, *an action profile* $a_{\mathcal{I}} \in A_{\mathcal{I}}$ *is* rationalizable *if there is some model of* M *with a world* $w \in [CTBR]$ *such that* $f(w) = a_{\mathcal{I}}$.

The following theorem provides a characterization of the set of rationalizable strategies. First we define, for each information set $I \in \mathcal{I}$ a sequence of action sets $D_I^1, D_I^2, \ldots$, where

$$D_I^1 = \left\{a_I \in A_I \mid \text{there is no } \alpha_I' \in \Delta A_I \text{ such that } u_{\imath(I)}(\alpha_I', a_{-I}) > u_{\imath(I)}(a_I, a_{-I}) \text{ for all } a_{-I} \in A_{-I}(I)\right\}$$

$$D_I^{m+1} = \left\{a_I \in A_I \mid \text{there is no } \alpha_I' \in \Delta A_I \text{ such that } u_{\imath(I)}(\alpha_I', a_{-I}) > u_{\imath(I)}(a_I, a_{-I}) \text{ for all } a_{-I} \in D_{-I}^m\right\}$$

for $m = 1, 2, \ldots$ (where ΔA_I is the set of probability measures on A_I, and the definition of u_i is extended in the usual way). D_I is the limit of this sequence: $D_I = \bigcap_{m=1}^{\infty} D_I^m$, and $D_{\mathcal{I}}$ is the set of corresponding action profiles. $D_{\mathcal{I}}$ is the set of actions profiles which survive a certain iterated elimination procedure, the generalization of iterated deletion of strictly dominated strategies to extensive form games: every action that is strictly dominated at any information set for the player on move at that information set is deleted in the first round, and the standard procedure is applied to what is left of the whole game.

Theorem 1 *For any game* Γ, *a action profile* $a_{\mathcal{I}}$ *is rationalizable if and only if* $a_{\mathcal{I}} \in D_{\mathcal{I}}$.

The proof of Theorem 1 is given in the appendix, but the intuition behind the result is straightforward. It follows from the definition of rationality that no rational player will play an action that is strictly dominated at any information set at which she is on move. This accounts for the first round of deletion. But we cannot apply iterated deletion at any of these information sets: unless all the players believed with positive probability at the start of the game that a particular information set *would be* reached, there may no longer be common belief in rationality if that information set *is* reached. And it is common belief that drives iterated deletion. Nevertheless, there is common belief at the start of the game, so we can apply iterated deletion to the game as a whole.

Two examples will illustrate the strength and the weakness of the deletion procedure. Figure 1 shows the familiar entry deterrence game, with an entrant (E) first deciding whether to enter the market (**i**) or stay out (**o**), and then an incumbent (I) deciding whether fight (**f**) or acquiesce (**a**) if entry occurs.

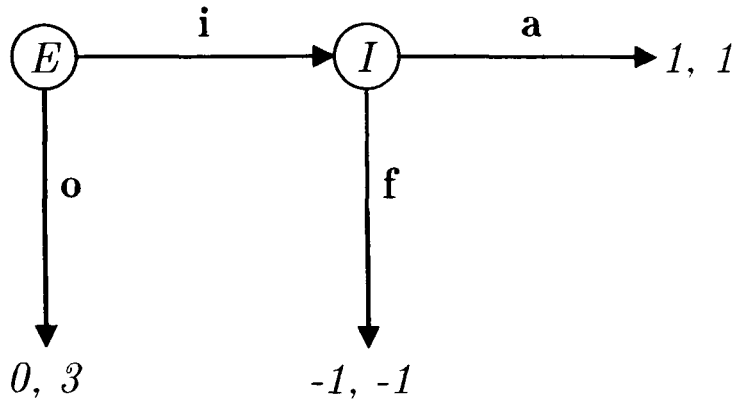


Figure 1: Entry Deterrence

Let 1 and 2 denote the two information sets. Neither of the actions available to E is strictly dominated at information set 1, so $D_1^1 = \{\mathbf{o}, \mathbf{i}\}$. But $\mathbf{f}$ is strictly dominated by $\mathbf{a}$ for I at information set 2, so $D_2^1 = \{\mathbf{a}\}$. Now, given that only actions in D_2^1 are chosen, $\mathbf{o}$ is strictly dominated by $\mathbf{i}$ for E at information set 1. No more actions can be deleted, so $D_I = \{(\mathbf{i}, \mathbf{a})\}$. This simple example shows how the information structure of the game is used to eliminate actions which would survive if the iterated deletion procedure were applied to the strategic form of the game. Furthermore, in this game, rationalizability is stronger than Nash equilibrium.

Figure 2 depicts a single-person decision problem[9]. Self (S) wants to coordinate her actions to maximize her payoff.

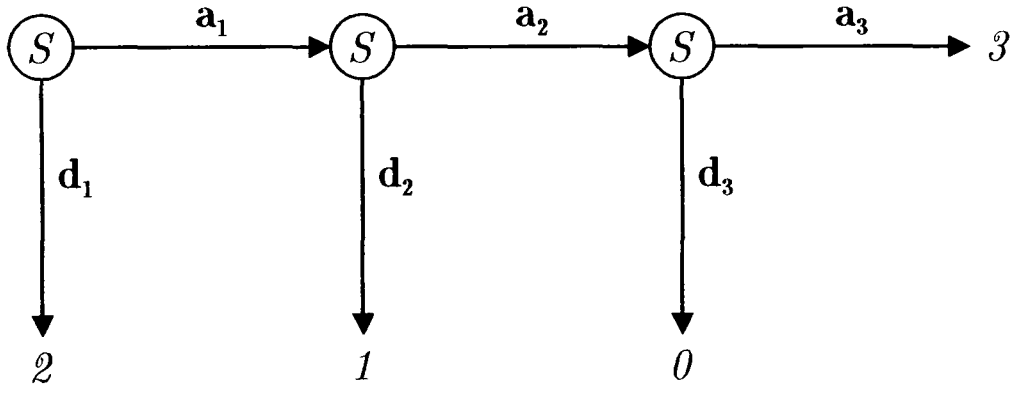


Figure 2: Self Coordination

Label the information sets $1, 2$ and 3 in order. On the first round, $\mathbf{d_3}$ is dominated by $\mathbf{a_3}$ at information set 3, but no other action is dominated at the information set at which it is chosen. But no more actions can be deleted on the second round: $\mathbf{d_2}$ survives because it does just as well as $\mathbf{a_2}$ against $(\mathbf{d_1}, \mathbf{a_3})$. Thus $D_I = \{(\mathbf{a_1}, \mathbf{a_2}, \mathbf{a_3}), (\mathbf{d_1}, \mathbf{a_2}, \mathbf{a_3}), (\mathbf{a_1}, \mathbf{d_2}, \mathbf{a_3}), (\mathbf{d_1}, \mathbf{d_2}, \mathbf{a_3})\}$. This seemingly paradoxical result arises because we do not assume that players can commit to their action choices at future nodes. Consider action profile $(\mathbf{d_1}, \mathbf{d_2}, \mathbf{a_3})$. How can this be consistent with common true belief in rationality? Suppose that if information set 2 were reached, S would no longer believe

[9] This is a one-player version of Figure 2 in Stalnaker [25].

herself to be rational, but rather that she would play $\mathbf{d}_3$ if information set 3 were reached. Then the rational thing to do at information set 2 is to play $\mathbf{d}_2$; and if she believes that she is rational and would have these beliefs at information set 2, the rational thing to do at information set 1 is to play $\mathbf{d}_1$. It follows that if information set 2 is reached, S is certainly right to doubt her own rationality: according to the beliefs just described she has just chosen an irrational action.

This example does not rely on lack of introspection: at information set 1, S has no doubt about what her choices will be throughout the game, and these beliefs are correct. If she plays $\mathbf{a}_1$ she surprises herself and her beliefs must be revised. Rather the issue is one of self control: at a given information set, S can control her action only at that information set, and not at future information sets as well.

5 Conclusions

This paper uses the framework of Board [12] to construct models which describe players' beliefs in extensive form games. Their beliefs about the game and about each other are expressed at the beginning the game and at every information set. These models are used to analyze rational play, and Theorem 1 describes the implications of common (true) belief in rationality.

We believe that our approach has two key strengths. The first is transparency: although the models we use to prove Theorem 1 are based around the rather obscure notion of a possible world, they can be used to provide truth conditions for a formal language of belief revision which has a straightforward interpretation. Furthermore, the properties of this language can be clarified by means of an axiom system: formulas of the language that are true at every world of every model are precisely those that are provable in the axiom system. Thus Theorem 1 can be translated into the formal language. The "only if" part of that theorem tells us that the formula $CTBR \Rightarrow D_I$ is true at every world of every model. It follows that it is provable in the axiom system $BRS^C + \mathbf{Game}$: we have a set of precise conditions which are sufficient to derive our result. The "if" part of the theorem tells us that, for any action profile $a_I \in D_I$, there is some world of some model at which $a_I \wedge CTBR$ is true; thus $a_I \wedge CTBR$ is logically consistent according to $BRS^C + \mathbf{Game}$.

The second strength is flexibility. The axiom system used is minimal in the sense that it imposes a weak set of conditions on the beliefs and belief revision policies of rational players (at least, in relation to most of the related literature). But extra axioms can be added, along with the corresponding restrictions on the models so that the tight link between truth and provability

is retained, and their effects can be examined. For instance we could examine whether the strong introspection assumption (that players are fully aware of all their beliefs, past present and future) implicitly adopted by Stalnaker [25] affects our result. It would also be interesting to analyze the impact of Battigalli's [3] *best rationalization principle*[10]. According to this principle, players should believe each other to be rational as long as those beliefs are consistent with the observed pattern of behavior; subject to that constraint, they should believe that they believe they are all rational as long as it is consistent to do so, and so on. It is not possible to represent this principle in an arbitrary model of a given game: we need to ensure that there *enough* worlds in the model so that if an action is consistent with iterated belief in rationality of a certain depth, then there is a world in the model being used in which that action is played and there is iterated belief in rationality of a certain depth. The *canonical structure* described in Board [12] contains a world for *every* logically consistent set of beliefs, so it provides an ideal tool for investigating the best rationalization principle. We conjecture that adding the principle to the assumption of common belief in rationality would allow us to carry out iterated deletion of actions at each information set, rather than just one round of deletion followed by iterated deletion in the strategic form. In perfect information games, this would generate the backward induction procedure. Finally, we could consider Stalnaker's [25] notion of *perfect* rationality, according to which every action profile of one's opponents is taken into account in the expected utility calculation. We conjecture that common belief in perfect rationality plus the best rationalization principle would give an epistemic characterization of iterated deletion of weakly dominated strategies.

References

[1] AUMANN, R. J. (1976), "Agreeing to disagree", *Annals of Statistics* **4**, 1236–1239.

[2] AUMANN, R. J. (1987), "Correlated Equilibrium as an Expression of Bayesian Rationality", *Econometrica* **55**, 1–18.

[3] BATTIGALLI, P. (1996), "Strategic Rationality Orderings and the Best Rationalization Principle", *Games and Economic Behavior* **13**, 178–200.

[10] This would not be a new project: much of the analysis of Battigalli and Siniscalchi [6] is based around this principle.

[4] BATTIGALLI, P. AND G. BONANNO (1998), "Recent Results on Belief, Knowledge and the Epistemic Foundations of Game Theory," *Research in Economics* **53**, 149–225.

[5] BATTIGALLI, P. AND M. SINISCALCHI (1999), "Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games", *Journal of Economic Theory* **88**, 188–230.

[6] BATTIGALLI, P. AND M. SINISCALCHI (2001): "Strong Belief and Forward Induction Reasoning", forthcoming, *Journal of Economic Theory.*

[7] BERNHEIM, B. D. (1984), "Rationalizable Strategic Behavior", *Econometrica* **52**, 1007–1028.

[8] BINMORE, K. (1987), "Modelling Rational Players I", *Economics and Philosophy* **3**, 179–214.

[9] BRANDENBURGER, A. AND E. DEKEL (1993), "Hierarchies of Beliefs and Common Knowledge", *Journal of Economic Theory* **59**, 189–198.

[10] BRANDENBURGER, A. AND H. J. KEISLER (2002), "Epistemic Conditions for Iterated Admissability", mimeo, Harvard Business School.

[11] BOARD, O. J. (1998), "Belief Revision and Rationalizability", TARK VII, Conference Proceedings, ed. by I. Gilboa.

[12] BOARD, O. J. (2002), "Dynamic Interactive Epistemology", mimeo, Department of Economics, University of Oxford.

[13] BOARD, O. J. (2002), "The Equivalence of Bayes and Causal Rationality in Games", mimeo, Department of Economics, University of Oxford.

[14] DEKEL, E. AND F. GUL (1997), "Rationality and Knowledge in Game Theory", in *Advances in Economics and Econometrics: Theory and Applications: Seventh World Congress, Vol. 1*, ed. by D. M. Kreps and K. W. Wallis. Cambridge University Press, 87–172.

[15] FEINBERG, Y. (2001), "Epistemic Characterizations of Equilibria and the Reasonable Solution", mimeo, Stanford Graduate School of Business.

[16] FEINBERG, Y. (2002), "Subjective Reasoning in Dynamic Games", mimeo, Stanford Graduate School of Business, Stanford.

[17] GUL, F. (1996): "Rationality and Coherent Theories of Strategic Behavior", *Journal of Economic Theory* **70**, 1–31.

[18] KREPS, D., P. MILGROM, J. ROBERTS, AND R. WILSON (1982), "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma", *Journal of Economic Theory* **27**, 245–252.

[19] MERTENS, J. F. AND S. ZAMIR, (1985), "Formalization of Harsanyi's notion of 'type' and 'consistency' in games with incomplete information", *International Journal of Game Theory* **14**, 1–29.

[20] OSBORNE, M. J. AND A. RUBINSTEIN (1994), *A Course in Game Theory.* The MIT Press, Cambridge, MA.

[21] PEARCE, D. G. (1984): "Rationalizable Strategic Behavior and the Problem of Perfection", *Econometrica* **52**, 1029–1050.

[22] PETTIT, P. AND R. SUGDEN, (1989), "The Backward Induction Paradox", *Journal of Philosophy* **86**, 169–182.

[23] RENY, P. (1992), "Rationality in Extensive Form Games", *Journal of Economic Perspectives* **6**, 103–118.

[24] STALNAKER, R. (1994), "On the Evaluation of Solution Concepts", *Theory and Decision* **37**, 49–73.

[25] STALNAKER, R. (1996), "Knowledge, Belief and Counterfactual Reasoning in Games", *Economics and Philosophy* **12**, 133–163.

[26] STALNAKER, R. (1998), "Belief Revision in Games: Forward and Backward Induction", *Mathematical Social Sciences* **36**, 31–56.

[27] STALNAKER, R. (1999), "Extensive and Strategic Form Games: Games and Models for Games", *Research in Economics* **53**, 293–319.

[28] TAN, T, AND S. R. C. WERLANG (1988), "The Bayesian Foundations of Solution Concepts of Games", *Journal of Economic Theory* **45**, 370–391.

A Proof of Theorem 1

First we recall the following lemma (see e.g. Pearce [21]).

Lemma 1 *An action of a player in a finite strategic form game is a best response if and only if it is not strictly dominated.*

We are now in a position to prove the main theorem.

(if) To prove the "if" statement, we must construct, for arbitrary $a_{\mathcal{I}} \in D_{\mathcal{I}}$, a model of Γ in which there a world $w \in [CTBR]$ such that $f(w) = a_{\mathcal{I}}$. Let $W = A_{\mathcal{I}}$[11], and for all $a_{\mathcal{I}}$, let $f(a_{\mathcal{I}}) = a_{\mathcal{I}}$. We show how to construct the plausibility orderings of each player at an arbitrary world $a^*_{\mathcal{I}} \in D$. We do this by constructing a function $k : A_{\mathcal{I}} \to \mathbb{N}$, which assigns each world a numerical ranking according to plausibility.

First we order each information set $I \in \mathcal{I}_i$: if player i is moving from the nth time at information set I, let $order(I) = n$. Given the assumption of perfect recall, this function is well defined. Now take any I such that $order(I) = 1$. We can think of the actions in D_I as player i's strategy set in a strategic form game, and the action profiles in D_{-I} as the strategy profiles of her opponent. i's payoffs are given by $u_i(a_I, a_{-I})$, and her opponent's payoffs are chosen arbitrarily. Since $a^*_I \in D_I$, it is not strictly dominated in this game, and therefore by Lemma 1 there is some probability measure μ_0 over D_{-I} such that a^*_I is a best response to μ_0. Extend the domain of μ_0 to the whole of $A_{\mathcal{I}}$ in the following way: $\mu_0(a_I, a_{-I}) = \mu_0(a_{-I})$ if $a_I = a^*_I$ and $a_{-I} \in D_{-I}$; $\mu_0(a_I, a_{-I}) = 0$ otherwise. Notice that $\mu_0(a_{\mathcal{I}}) > 0$ only if $a_{\mathcal{I}} \in D_{\mathcal{I}}$. Next, construct a probability measure μ_1 over $A_{\mathcal{I}}$ in three steps:

1. if $a_{\mathcal{I}} \notin A_{\mathcal{I}}(I)$ for any $I \in \mathcal{I}_i$ of order 1, let $\mu_1(a_{\mathcal{I}}) = \mu_0(a_{\mathcal{I}})$;

2. if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ for some $I \in \mathcal{I}_i$ of order 1, but $\mu_0(A_{\mathcal{I}}(I)) = 0$, let $\mu_1(a_{\mathcal{I}}) = 0$.

3. if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ for some $I \in \mathcal{I}_i$ of order 1 and $\mu_0(A_{\mathcal{I}}(I)) > 0$, consider the conditional probability $\mu_0(a_{-I} \mid A_{\mathcal{I}}(I)) = \frac{\mu_0(a_{-I} \cap A_{\mathcal{I}}(I))}{\mu_0(A_{\mathcal{I}}(I))}$. There is some $a'_I \in A_I$ which is best response to $\mu_0(\cdot \mid A_{\mathcal{I}}(I))$ (there may be more than one). It must be the case that $a'_I \in D_I$ by Lemma 1, since $\mu_0(\cdot \mid A_{\mathcal{I}}(I))$ places positive weight only on $a_{-I} \in D_{-I}$. $\mu_1(a_{\mathcal{I}})$ is defined as follows: (i) $\mu_1(a_{\mathcal{I}}) = \mu_0(a_{-I})$ if $a_I = a'_I$ (where a_I is the Ith component of $a_{\mathcal{I}}$ and a_{-I}

[11] Note that we are now using action labels for three purposes: the denote actions themselves, to denote formulas of $\mathcal{L}$ describing which actions are chosen, and to denote worlds. Since we do not use the language $\mathcal{L}$ in this proof, there should be no risk of confusion.

is the $-I$th component of $a_{\mathcal{I}}$); (ii) $\mu_1(a_{\mathcal{I}}) = 0$ otherwise. Observe that $\mu_1(a_{-I} \mid A_{\mathcal{I}}(I)) = \mu_0(a_{-I} \mid A_{\mathcal{I}}(I))$, since $\mu_1(a_{-I}) = \mu_1(a'_I, a_{-I}) = \mu_0(a_{-I})$ for all $a_{-I} \in A_{-I}(I)$, and the a_{-I}'s in $A_{-I}(I)$ partition $A_{\mathcal{I}}(I)$. So a'_I is a best response to $\mu_1(\cdot \mid A_{\mathcal{I}}(I))$.

This process is well defined since the $A_{\mathcal{I}}(I)$ sets are disjoint. Notice that $\mu_1(a_{\mathcal{I}}) > 0$ only if $a_{\mathcal{I}} \in D_{\mathcal{I}}$.

Now construct a probability measure μ_2, again in three steps:

1. if $a_{\mathcal{I}} \notin A_{\mathcal{I}}(I)$ for any $I \in \mathcal{I}_i$ of order 2, let $\mu_2(a_{\mathcal{I}}) = \mu_1(a_{\mathcal{I}})$;

2. if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ for some $I \in \mathcal{I}_i$ of order 2, but $\mu_1(A_{\mathcal{I}}(I)) = 0$, let $\mu_2(a_{\mathcal{I}}) = 0$.

3. if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ for some $I \in \mathcal{I}_i$ of order 2 and $\mu_1(A_{\mathcal{I}}(I)) > 0$, consider the conditional probability $\mu_1(a_{-I} \mid A_{\mathcal{I}}(I))$. By the same reasoning as before, there is some $a'_I \in D_I$ which is best response to $\mu_1(\cdot \mid A_{\mathcal{I}}(I))$. Let $\mu_2(a_{\mathcal{I}})$ be defined as follows: (i) $\mu_2(a_{\mathcal{I}}) = \mu_1(a_{-I})$ if $a_I = a'_I$; (ii) $\mu_1(a_I, a_{-I}) = 0$ otherwise. Again by the same reasoning as before, we know that a'_I is a best response to $\mu_2(\cdot \mid A_{\mathcal{I}}(I))$.

Notice that $\mu_2(a_{\mathcal{I}})$ only if $a_{\mathcal{I}} \in D_{\mathcal{I}}$. We have shown that if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ for some $I \in \mathcal{I}_i$ of order 2 and $\mu(a_{\mathcal{I}}) > 0$, then a_I is a best response to $\mu_2(\cdot \mid A_{\mathcal{I}}(I))$. We want to show also that if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ for some $I \in \mathcal{I}_i$ of order 1 and $\mu(a_{\mathcal{I}}) > 0$, then a_I is a best response to $\mu_2(\cdot \mid A_{\mathcal{I}}(I))$. We know that a_I is a best response to $\mu_1(\cdot \mid A_{\mathcal{I}}(I))$, i.e.

$$\sum_{a_{-I}} \mu_1(a_{-I} \mid A_{\mathcal{I}}(I)) . u_i(a_I, a_{-I}) \geq \sum_{a_{-I}} \mu_1(a_{-I} \mid A_{\mathcal{I}}(I)) . u_i(a'_I, a_{-I}) \text{ for all } a'_I \in A_I.$$

Now consider the information sets $I', I'', \ldots \in \mathcal{I}_i$ immediately following I, and corresponding subsets of $A_{-I}(I'), A_{-I}(I''), \ldots$ of A_{-I}. For every a_{-I} not in one of these subsets, $\mu_1(a_{-I}) = \mu_2(a_{-I})$ (by step 1) and therefore $\mu_1(a_{-I} \mid A_{\mathcal{I}}(I)) = \mu_2(a_{-I} \mid A_{\mathcal{I}}(I))$. Next consider every $a_{-I} \in A_{-I}(I')$. If $\mu_1(A_{-I'}(I)) = 0$, then $\mu_1(a_{-I}) = \mu_2(a_{-I})$ (by step2) and therefore $\mu_1(a_{-I} \mid A_{\mathcal{I}}(I)) = \mu_2(a_{-I} \mid A_{\mathcal{I}}(I))$ again. So suppose $\mu_1(A_{-I'}(I)) > 0$. μ_1 and μ_2 generate the same beliefs about actions at every information set except I', but μ_2 assumes that the action chosen at I' is a best response to those beliefs, while according to μ_1 it can be chosen arbitrarily (step 3). Thus, restricting attention to $a_{-I} \in A_{-I}(I')$, we have:

$$\sum_{a_{-I'} \in A_{-I'}(I')} \mu_2(a_{-I} \mid A_{\mathcal{I}}(I)) . u_i(a_I, a_{-I}) \geq \sum_{a_{-I'} \in A_{-I'}(I')} \mu_1(a_{-I} \mid A_{\mathcal{I}}(I)) . u_i(a'_I, a_{-I}) \text{ for all } a'_I \in A_I.$$

On the other hand, if action $a'_I \neq a_I$ is chosen at information set I, information set I' is not reached (given perfect recall) and we have:

$$\sum_{a_{-I'} \in A_{-I'}(I')} \mu_2 \left(a_{-I} \mid A_{\mathcal{I}}(I)\right) . u_i \left(a'_I, a_{-I}\right) = \sum_{a_{-I'} \in A_{-I'}(I')} \mu_1 \left(a_{-I} \mid A_{\mathcal{I}}(I)\right) . u_i \left(a'_I, a_{-I}\right) \text{ for all } a'_I \neq a_I.$$

Aggregating across the subsets $A_{-I}(I'), A_{-I}(I''), \ldots$ of A_{-I} and every a_{-I} not in one of these subsets, we obtain:

$$\sum_{a_{-I}} \mu_2 \left(a_{-I} \mid A_{\mathcal{I}}(I)\right) . u_i \left(a_I, a_{-I}\right) \geq \sum_{a_{-I}} \mu_2 \left(a_{-I} \mid A_{\mathcal{I}}(I)\right) . u_i \left(a'_I, a_{-I}\right) \text{ for all } a'_I \in A_I,$$

as required.

Now construct a probability measure μ_3 by the same procedure, taking each information set $I \in \mathcal{I}_i$ of rank 3. Repeat until every information set in $\mathcal{I}_i$ has been used. We have some μ_k with the property that:

(i) $\mu_k(a_{\mathcal{I}}) > 0$ only if $a_{\mathcal{I}} \in D_{\mathcal{I}}$;

(ii) if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ for some $I \in \mathcal{I}_i$ and $\mu_k(a_{\mathcal{I}}) > 0$, then a_I is a best response to $\mu_k(\cdot \mid A_{\mathcal{I}}(I))$;

(iii) if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ for some $I \in \mathcal{I}_i$ and $\mu_k(a_{\mathcal{I}}) > 0$, then $\mu_k(a_I \mid A_{\mathcal{I}}(I)) = 1$.

For all $a_{\mathcal{I}}$ such that $\mu_k(a_{\mathcal{I}}) > 0$, let $k(a_{\mathcal{I}}) = 0$, and let $p'_i(a_{\mathcal{I}}) = \mu_k(a_{\mathcal{I}})$.

Now consider every information set $I \in \mathcal{I}_i$ such that $\mu_k(A_{\mathcal{I}}(I)) = 0$. These are the information sets that should not be reached according to the beliefs μ_k. For each such set, I, of lowest order, we can use the same technique as for the construction of μ_k to construct a probability measure μ over $A_{\mathcal{I}}(I)$ with analogous properties to μ_k:

(i) $\mu(a_{\mathcal{I}}) > 0$ only if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$;

(ii) if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ for some $I \in \mathcal{I}_i$ and $\mu(a_{\mathcal{I}}) > 0$, then a_I is a best response to $\mu(\cdot \mid A_{\mathcal{I}}(I))$;

(iii) if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ for some $I \in \mathcal{I}_i$ and $\mu(a_{\mathcal{I}}) > 0$, then $\mu(a_{\mathcal{I}} \mid A_{\mathcal{I}}(I)) = 1$.

Note that $\mu(a_{\mathcal{I}}) > 0$ only if $\mu_k(a_{\mathcal{I}}) = 0$. For all $a_{\mathcal{I}}$ such that $\mu(a_{\mathcal{I}}) > 0$, let $k(a_{\mathcal{I}}) = order(I)$ and let $p'_i(a_{\mathcal{I}}) = \mu(a_{\mathcal{I}})$.

Now we take every information set $I \in \mathcal{I}_i$ for which there is no $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ such that $k(a_{\mathcal{I}})$ has been defined, and repeat the process. We continue until there are no information sets left.

The $\preccurlyeq_i^{a_I}$ relation is defined as follows: $a'_I \preccurlyeq_i^{a_I} a''_I$ if and only if $k(a'_I) \leq k(a''_I)$ or $k(a'_I)$ is defined and $k(a''_I)$ is not. $a'_I \in W_i^{a_I}$ if and only if it has been assigned a rank by $k(.)$, and since $\leq$ is complete and transitive on the natural numbers and A_I is finite, $\preccurlyeq_i^{a_I}$ satisfies **R1** and **R2**.

To show that $a_I \in [Rat_i]$ (i.e. that player i is rational at a_I), let p_i be the normalization of p'_i so that $p_i(A_I) = 1$, with $p_i(a_I) = 0$ if $p'_i(a_I)$ is not defined. For arbitrary $I \in \mathcal{I}_i$, we must compute $p_{i,I}^{a_I}([a_{-I}])$. Consider the set $\min_i^{a_I}([I] \cap [f_I(w)] \cap [W_i^w])$. $[I] = A_I(I)$ and $[f_I(w)] = a_I$, from the definition of W. Furthermore, from the construction of $\preccurlyeq_i^{a_I}$, every $\preccurlyeq_i^{a_I}$-minimal element of a given set must have been assigned the same k ranking (and must therefore be in $W_i^{a_I}$). Each of these elements must therefore have been assigned its k rank and its $p'_i(.)$ value (if strictly positive) by the same μ measure (or by μ_k), since by perfect recall, if $a_I \in A_I(I)$ for some $I \in \mathcal{I}_i$ of order n, there is no other $I' \in \mathcal{I}_i$ of order n such that $a_I \in A_I(I')$. So there is some μ or μ_k such that:

$$\begin{aligned}
p_{i,I}^w([a_{-I}]) &= \frac{p_i([a_{-I}] \cap \min_i^w([I] \cap [f_I(w)] \cap W_i^w))}{p_i(\min_i^w([I] \cap [f_I(w)] \cap W_i^w))} \\
&= \frac{p'_i(a_{-I} \cap \min_i^w(A_I(I) \cap a_I))}{p'_i(\min_i^w(A_I(I) \cap a_I))} \\
&= \frac{\mu(a_{-I} \cap A_I(I) \cap a_I)}{\mu(A_I(I) \cap a_I)} \\
&= \mu(a_{-I} \mid A_I(I) \cap a_I) \\
&= \mu(a_{-I} \mid A_I(I))
\end{aligned}$$

The last inequality follows from (iii) above. From (ii) above, a_I is a best response to $\mu(\cdot \mid A_I(I))$, i.e.

$$\begin{aligned}
&\sum_{a_{-I} \in A_{-I}} \mu(a_{-I} \mid A_I(I)).u_i(a_I, a_{-I}) \geq \sum_{a_{-I} \in A_{-I}} \mu(a_{-I} \mid A_I(I)).u_i(a'_I, a_{-I}) \text{ for all } a'_I \in A_I \\
\Rightarrow &\sum_{a_{-I} \in A_{-I}} p_{i,I}^w([a_{-I}]).u_i(a_I, a_{-I}) \geq \sum_{a_{-I} \in A_{-I}} p_{i,I}^w([a_{-I}]).u_i(a'_I, a_{-I}) \text{ for all } a'_I \in A_I
\end{aligned}$$

The same result holds at every $I \in \mathcal{I}_i$, and so player i is rational at a_I as required.

$\preccurlyeq_i^{a_I}$ is defined in the same way for every player i at every world $a_I \in D_I$. If $a_I \notin D_I$, $\preccurlyeq_i^{a_I}$ can be defined in any way that satisfies **R1** and **R2**. We have already seen that $D_I \subseteq [Rat]$. For every player i, notice that if $a_I \in D_I$, $a'_I \in \min_i^{a_I}(W_i^{a_I})$ only if $a'_I \in D_I$. It follows from the definition of $[B_i\phi]$ that $D_I \subseteq [B_iRat]$ for all i. So we have $D_I \subseteq [ERat]$, and repeating the argument we obtain $D_I \subseteq [CTBR]$. So we have shown that, for every $a_I \in D_I$, a_I is rationalizable, as required.

(only if) Take any model of Γ. First, we observe that, for all $I \in \mathcal{I}_i$, if $w \in [Rat_i]$, then $f_I(w) \in D_I^1$. This follows immediately from Lemma 1 and the definition of rationality. Thus, for all $w \in [Rat]$, $f(w) \in D_{\mathcal{I}}^1$. Now suppose that for some $I \in \mathcal{I}_i$, $a_I \notin D_I^2$. By Lemma 1, there is no probability measure over D_{-I}^1 to which a_I is a best response. It follows that $D_{-I}^1 \subseteq A_{-I}(I)$, since if there was some $a_{-I} \in D_{-I}^1$ which did not reach I, a_I would not affect the path through the game if a_{-I} were chosen. Hence a_I would be a best response to a_{-I}. So $[Rat] \subseteq \left[D_{-I}^1\right] \subseteq [A_{-I}(I)]$, and therefore $[B_i Rat] \subseteq [B_i A_{-I}(I)]$. Now suppose $w \in [B_i Rat]$. We must have $\min_i^w(W_i^w) \subseteq [Rat] \subseteq [A_{-I}(I)]$. But $[A_{-I}(I)] = [I]$, so $\min_i^w(W_i^w) = \min_i^w([I] \cap W_i^w)$. It follows from the definition of $p_{i,I}^w(.)$ that $p_{i,I}^w(a_{-I}) > 0$ only if $a_{-I} \in D_{-I}^1$. So from the definition of rationality, if $w \in [B_i Rat] \cap [Rat_i]$, $f_I(w) \in D_I^2$. Aggregating over players and information sets gives us $[ERat] \cap [Rat] \subseteq \left[D^2\right]$, and iteration of the second step yields $[CTBR] \subseteq [D_{\mathcal{I}}]$. Thus if $a_{\mathcal{I}}$ is rationalizable, then $a_{\mathcal{I}} \in D_{\mathcal{I}}$, as required.■

Chapter 3

The Equivalence of Bayes and Causal Rationality in Games*

Abstract: In a seminal paper, Aumann [1] showed how the choices of rational players could be analyzed in a unified state space framework. His innovation was to include the choices of the players in the description of the states, thus avoiding the distinction between acts and consequences made by Savage [17]. But this simplification comes at a price: Aumann's notion of *Bayes rationality* does not allow players to evaluate what would happen were they to deviate from their actual choices. We show how the addition of a causal structure to the framework enables us to analyze such counterfactual statements, and use it to introduce a notion of *causal rationality*. Under a plausible causal independence condition, the two notions are shown to be equivalent. If we are prepared to accept this condition we can dispense with the causal apparatus and retain Aumann's original framework.

1 Introduction

The so-called Bayesian approach to game theory takes the view that games should be analyzed as a number of inter-related single-person decision problems in the sense of Savage [17], with each player maximizing expected utility with respect to some subjective probability distribution over a set of uncertain events, in this case the strategy choices of the other players. This approach, pioneered by Bernheim [6] and Pearce [16], was originally contrasted with the equilibrium approach to games, according to which probabilities can only assigned to events not governed by rational decision makers. In equilibrium, strategies are not the subject of uncertainty.

Aumann [1] provided a reconciliation of these two approaches. He proved that, if the players are Bayesian expected utility maximizers, and possess a common prior over the space of uncertainty (which includes each player's strategy choice), then they will each play their part in some correlated equilibrium. And there is now a large and growing body of literature which seeks to characterize

*I would like to thank Michael Bacharach for many helpful discussions.

game-theoretic solution concepts explicitly in terms of epistemic conditions on expected-utility maximizing agents[1]. We shall refer to this literature as the *Bayesian tradition.*

But although Bayesian decision theory is at the heart of Aumann's paper, the framework that he (and others) adopt is not that of Savage. For in the Savage framework, a distinction is made between acts and consequences, the former being a function from the set of states of the world to the latter. If one's opponents' strategy choices are included among the objects of uncertainty, and hence are part of the description of a state, this distinction implies that we must have a different state space for each agent. Aumann overcame this problem by adopting a unified framework in which acts and consequences are both parts of the description of the state of the world. He describes this as the "chief innovation" in his model (p. 8). In particular, a state describes the strategy choice of *every* agent. Thus Aumann's framework is very much like that of Jeffrey [11], where personal choice is included as a state variable. An act is now a subset of the state space: precisely, the set of states at which that act is carried out.

Within this framework, each player is endowed with a prior (subjective) probability distribution over the entire state space, and is assumed to have certain information about which of the states has occurred. In particular, she knows what strategy she chooses; that is, at every state she considers possible, she carries out the same strategy choice. The player is said to be *Bayes rational* if this strategy choice maximizes her expected utility given her information.

There are, however, dangers in adopting this unified framework, as Jeffrey was aware: in certain circumstances, it gives us the wrong answer. An example is given in section 2. In section 3, we show how the framework can be enriched and a revised definition of rationality given that is not subject to this criticism. In section 4 we present a condition under which the two definitions are equivalent, and in section 5 we extend our analysis to extensive form games. Section 6 gives some concluding remarks.

2 An Example

Consider the following one-person decision problem[2]: President Clinton wants to have an affair with Monica Lewinsky, but fears that doing so may lead to his impeachment. The utility he gains from each of the possible outcomes is shown in figure 1 (a) below, where A is the event that he has

[1] e.g. Aumann and Brandenburger [2]. Dekel and Gul [7] and Battigalli and Bonanno [3] review this literature.

[2] This example is a re-labelled version of Gibbard and Harper's [8] Case 1. Thanks are due to Ehud Kalai for suggesting this modern version of the story of David and Bathsheba.

an affair, and M the event that he is impeached. His prior probability distribution over the state space is given in (b). It is clear from the utilities that having an affair is the unique Bayes rational act. This is true whatever the values of p, q, r, and s[3]. For A is a dominant strategy: whatever the posterior probabilities of M and $\bar{M}$ once he has updated on his private information, it will yield higher expected utility than $\bar{A}$. He reasons as follows: "Either I will be impeached or I won't be. Whether or not I will be impeached, I prefer to have an affair with Monica than not to, so I should go ahead and have one".

	M	$\bar{M}$
A	1	10
$\bar{A}$	0	9

(a)

	M	$\bar{M}$
A	p	q
$\bar{A}$	r	s

(b)

Figure 1

Of course, this reasoning is fallacious: it was, at least indirectly, Clinton's affair with Lewinsky that led to his impeachment, and if he had predicted this, he should have avoided the affair. The problem is that the object of uncertainty, i.e. whether or not the president will be impeached, is not independent of the acts being considered. Note that this cannot happen in the Savage framework, where acts are functions from states to consequences.

At this stage, there are two ways we might proceed. The first is to expand the state space by enriching the description of each state in such a way that we do have the required independence. For example, we could model Clinton as being uncertain between the following four events:

MM : I shall be impeached whatever I do;

$M\bar{M}$: I shall be impeached only if I have the affair;

$\bar{M}M$: I shall be impeached only if I don't have the affair;

$\bar{M}\bar{M}$: I shall not be impeached whatever I do.

These four events hold independently of A and $\bar{A}$. And it is easy to check that, if he places enough weight of probability on the second of these, the expected utility maximizing act is $\bar{A}$. A is no longer a dominant strategy. But even before we tackle the difficulty of how we might assign

[3] As long as $p+q>0$ and $r+s>0$, so Bayesian updating is well defined in all cases. This seems to be a reasonable requirement: a player's prior shouldn't rule out any of her strategy choices.

probabilities to these conditionalized states, there is another problem with this proposal that leads us to reject it for the current purposes. If our aim is to model rational play in games, then one player's uncertainty is another player's strategy choice. And the set of strategy choices is fixed by the structure of the game; we cannot expand it without abandoning Aumann's unified framework altogether, and with it the advantages of expressing every player's decision problem within the same state space.

Instead, we follow the second alternative, which is to adopt a version of *causal decision theory*[4] as our principle of rational choice. According to causal decision theory, the weights we should use for our expected utility calculations are not simple conditional probabilities (where the conditioning event is that we carry out some particular act), but conditional causal chances. In other words, we must consider how likely each uncertain event (opponent's strategy choice) is given what we're actually going to do, and how likely it *would be* if we were to do something different: it is not rational for Clinton to have the affair, because if he were to avoid it, he wouldn't be impeached.

In the next section, we first present Aumann's framework, and then show how it can be enriched to provide a formal statement of causal rationality.

3 Bayes rationality and causal rationality

3.1 Aumann's framework

The starting point of Aumann's analysis is an n-person normal form game G. For each player $i = 1, \ldots, n$, let S_i be player i's set of pure strategies, and $u_i : S \rightarrow \mathbb{R}$ be her utility function, where $S = S_1 \times \ldots \times S_n$. Note the implicit assumption that G is a game of complete information: the only utility-relevant uncertainty faced by the players is what strategies they and their opponents will play. Our formal model of G describes the players' beliefs about these strategy choices (and their beliefs about these beliefs, etc.), and consists of four elements:

- a finite[5] set W, with generic element w;
- for each player i, a binary relation $\mathcal{B}_i$;
- for each player i, a probability measure p_i on W;

[4] See Gibbard and Harper [8], Lewis [14], Skyrms [21], and Sobel [22] for alternative statements of this theory.

[5] The assumption that W is finite is not without loss of generality, even in finite games (see e.g. Battigalli and Siniscalchi [4]), but for the current purposes there is no need to deal with the additional complications raised by the infinite case.

- for each player i, a function $f_i : W \to S_i$.

The set W represents the set of *states of the world*, and w is one particular state. The binary relations $\mathcal{B}_i$, called *accessibility relations*, encode the players' information at each state[6]. At a given state w, the set of states that player i considers possible is given by $\{x : w\mathcal{B}_i x\}$. The propositions that i believes are just those that are true at every state in this set. In order to ensure that i's beliefs are coherent, we assume that this set is nonempty for every $w \in W$ (i.e. we assume that $\mathcal{B}_i$ is serial: $\forall w \exists x \, (w\mathcal{B}_i x)$). The probability measure p_i is i's *prior* over W, from which we obtain her probabilistic beliefs at each state of the world by updating on her information at that state. Thus i's (subjective) probability at w that some proposition ϕ is true, denoted $p_{i,w}([\phi])$, is given by

$$\frac{p_i([\phi] \cap \{x : w\mathcal{B}_i x\})}{p_i(\{x : w\mathcal{B}_i x\})},$$

where $[\phi] \subseteq W$ is the set of states where ϕ is true, i.e. the *event* that ϕ. We denote this probability by $p_{i,w}([\phi])$. To ensure that this ratio is always well defined, we assume that $p_i(w) > 0$ for all w. Finally, f_i is i's *decision function*. It gives the strategy chosen by i at each state of the world. Collectively, the f_i's determine a complete strategy profile for every state, and hence allow us to calculate each player's utility at that state. Uncertainty about states thus translates into uncertainty about strategies and uncertainty about utility. We assume, however, that each player knows her own strategy choice. This is expressed formally by the *own-choice knowledge* condition:

(OK) For all i and for all $w, x \in W$, if $w\mathcal{B}_i x$ then $f_i(w) = f_i(x)$.

Condition (OK) says that, at every state a player considers possible, the strategy she carries out is the same as the one she actually carries out. Note that this rules out the possibility of the player *trembling* (see e.g. Selten [18]), and accidentally playing a strategy other than the one she intended. In order to introducing trembles, we would need to make a distinction between *decisions*, the objects of choice, and *performances*, the strategies actually carried out (see Shin [19]). The player would know her own decision, but not necessarily her performance.

[6] Aumann assumes that these relations are equivalence relations, and hence partition the set of states; but for our purposes there is no need to make this rather restrictive assumption.

3.2 Bayes rationality

Our model of G generates probabilistic beliefs for each player at every state about her opponents' strategy choices, given her information at that state. Following Aumann, we say that a player is Bayes rational at a state w if her strategy choice at w maximizes her expected utility given these beliefs. Before giving the formal definition, we introduce some new notation: for any w, let $f(w) = (f_1(w), \ldots, f_n(w))$, the full strategy profile played at state w; and let $f_{-i}(w) = (f_1(w), \ldots, f_{i-1}(w), f_{i+1}(w), \ldots, f_n(w))$, the strategy profile played by all players other than i. In addition, for any strategy s_i, with a slight abuse of notation we let $[s_i]$ denote the event that s_i is played, i.e. $[s_i] = \{w : f_i(w) = s_i\}$; $[s_{-i}]$ is similarly defined to be the event that strategy profile s_{-i} is played. The definition of Bayes rationality can now be expressed as follows:

Definition 1 *Player i is* Bayes rational at w *if, for all $s_i \in S_i$,*

$$\sum_{s_{-i} \in S_{-i}} p_{i,w}([s_{-i}]) \cdot u_i(f_i(w), s_{-i}) \geq \sum_{s_{-i} \in S_{-i}} p_{i,w}([s_{-i}]) \cdot u_i(s_i, s_{-i}).$$

The left hand side of the inequality is i's expected utility given what she actually plays at state w, and the right hand side is her expected utility if she plays s_i instead.

3.3 Causal rationality

According to the Bayesian decision theory set up above, each player forms a subjective probability assessment over her opponents' strategy profiles by updating her prior with respect to her private information, which includes information about which strategy choice she will carry out. She then evaluates alternative strategy choices according to this probability assessment. Causal decision theory, on the other hand, recognizes that the various actions of each player might be inter-connected: my opponents' choices given that I play s_i might not be the same as they would have been had I chosen to play s_i'. Each player must consider what her opponents will do given her actual choice, and also what they would do if she were to choose something else.

A causal expected utility calculus, then, depends on *counterfactual* sentences such as "if it were the case that player i chose strategy s_i, then it would be the case that her opponents chose strategy profile s_{-i}", which we shall denote by $s_i \mapsto s_{-i}$. The corresponding event is denoted by $[s_i \mapsto s_{-i}]$. Using this shorthand, the definition of causal rationality is as follows:

Definition 2 *Player i is* causally rational at w *if, for all* $s_i \in S_i$,

$$\sum_{s_{-i} \in S_{-i}} p_{i,w}\left([f_i(w) \mapsto s_{-i}]\right) \cdot u_i\left(f_i(w), s_{-i}\right) \geq \sum_{s_{-i} \in S_{-i}} p_{i,w}\left([s_i \mapsto s_{-i}]\right) \cdot u_i\left(s_i, s_{-i}\right).$$

But the framework above gives us no way of evaluating counterfactuals, and so no way of evaluating this definition. To this end, we follow the Stalnaker-Lewis theory of counterfactuals[7], and augment the model with a *closeness* relation, $\preccurlyeq_w$, for each state w. Each $\preccurlyeq_w$ is a binary relation on W which satisfies the following four conditions:

(C1) $\preccurlyeq_w$ is complete;

(C2) $\preccurlyeq_w$ is transitive;

(C3) $\preccurlyeq_w$ is antisymmetric (for all x, y, if $x \preccurlyeq_w y$ and $y \preccurlyeq_w x$, then $x = y$);

(C4) $\preccurlyeq_w$ is centered (for all x, $w \preccurlyeq_w x$).

So, for a given state, w, the closeness relation $\preccurlyeq_w$ gives a total ordering of the states[8], with w at the bottom. The lower a state is in the ordering, the closer it is to w. A counterfactual, $\phi \mapsto \psi$, ("if it were the case that ϕ, it would be the case that ψ") is true just if, at the closest ϕ-world[9], ψ is true. Formally, $w \in [\phi \mapsto \psi]$ if and only if $\min_w [\phi] \in [\psi]$, where $\min_w$ refers to the least element of a subset of W with respect to the relation $\preccurlyeq_w$.

There is an attractive geometric representation of this account of counterfactuals which may clarify matters. Each $\preccurlyeq_w$relation partitions the state space into a "system of rings", with w at the centre, and each successive ring out from w containing the next closest state (see figure 2). $\phi \mapsto \psi$ is true at w just if the intersection of ϕ with the smallest ring for which this intersection is nonempty is wholly contained within ψ. Thus $\phi \mapsto \psi$ is true, but $\phi \mapsto \chi$ is not.

---

[7] See Stalnaker [23] and Lewis [13].

[8] Furthermore, the fact that W is finite implies that it is *well ordered* by $\preccurlyeq_w$, i.e. every nonempty subset has a lowest element. This enables us to dispense with Lewis's *Limit Assumption* (see Lewis [13], pp. 19–21).

[9] That there is a unique such world (implied by the antisymmetry of $\preccurlyeq_w$) is a property that the present account of counterfactuals shares with Stalnaker's theory but not with Lewis's. This property makes valid the law of *Conditional Excluded Middle*:

$$(\phi \mapsto \psi) \vee (\phi \mapsto \neg\psi).$$

See Lewis [13] for a discussion of the appropriateness of this law. For the current purposes, it is analytically very convenient, as it saves us the need to evaluate what Sobel [22] calls *practical chance conditionals*: "if it were the case that ϕ, then it might, with a probability of p, be the case that ψ". Furthermore, since for the causal expected utility calculus it is always *agents' beliefs about* the relevant counterfactuals that we shall be considering, the assumption is without loss of generality: our agents may be unsure about what the closeness relation is.

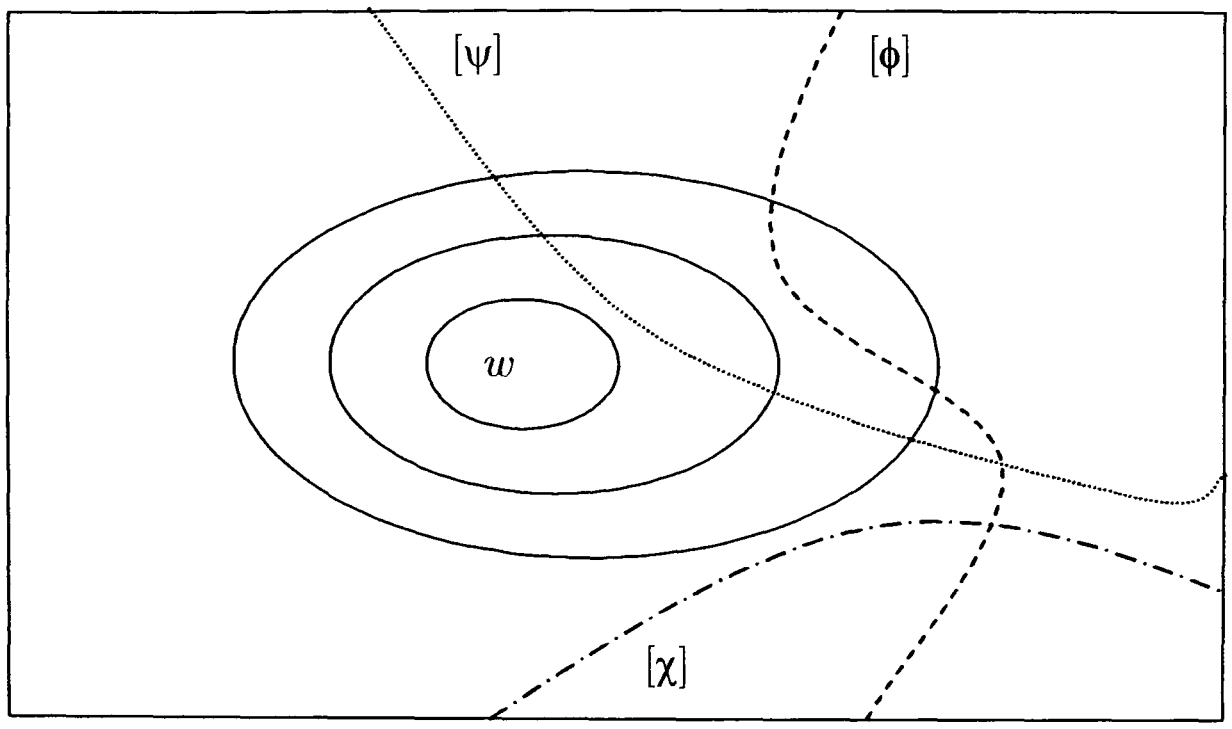


Figure 2

The closeness relation at any given state, then, enables us to evaluate counterfactual statements at that state. But our agents are typically subject to epistemic uncertainty: they are unsure what the actual state is. We assume that they form subjective probabilities for the event that a particular counterfactual is true just as they do for any other event: by conditionalizing on the set of states they consider possible, as given by the appropriate $\mathcal{B}_i$ relation. Thus,

$$p_{i,w}\left([\phi \mapsto \psi]\right) = \frac{p_i\left([\phi \mapsto \psi] \cap \{x : w\mathcal{B}_i x\}\right)}{p_i\left(\{x : w\mathcal{B}_i x\}\right)}.$$

This completes our account of counterfactuals. But an additional condition is required before we can use the augmented models to evaluate our definition of causal rationality. There must be enough states in the model to guarantee that, for each strategy choice of each player, there is a state in which that strategy choice is played. Formally, this *sufficiency* condition can be stated as follows:

(S) For each player i, for every $s_i \in S_i$, there exists a state w such that $f_i(w) = s_i$.

This guarantees that $p_{i,w}\left([s_i \mapsto s_{-i}]\right)$ is well defined for each $s_i \in S_i$. Henceforth, we shall assume that all our models satisfy this condition.

In the next section we compare Bayes rationality with causal rationality. We shall find that the

whether or not the two coincide hinges on a particular independence assumption, and we discuss how appropriate the assumption is.

4 Causal independence in games

In general, Bayes rationality and causal rationality do not coincide. Consider the game of *Odd Coordination* in figure 3 (a) below, where player 1 is choosing row and player 2 is choosing column. Assume that our model of the game takes $W = S$, where the f_i's are defined in the obvious way, so that $f(s) = s$. (b) describes $\mathcal{B}_1$ (which is here an equivalence relation, and can thus be represented by a partition over W) and (c) describes p_1. The causal structures at states (T, L) and (T, R) are given in (d) and (e), with the numbers representing distance, so closer worlds are assigned lower numbers.

	L	R
T	$(1,1)$	$(0,0)$
B	$(0,0)$	$(2,2)$

Odd Coordination

(a)

	L	R
T	•	•
B	•	•

Player 1's partition $(\mathcal{B}_1)$

(b)

	L	R
T	0.4	0.1
B	0.1	0.4

Player 1's prior (p_1)

(c)

	L	R
T	$\bullet^0$	$\bullet^2$
B	$\bullet^3$	$\bullet^1$

$\preccurlyeq_{(T,L)}$

(d)

	L	R
T	$\bullet^2$	$\bullet^0$
B	$\bullet^1$	$\bullet^3$

$\preccurlyeq_{(T,R)}$

(e)

Figure 3

It is easy to verify that player 1 is Bayes rational at world (T, L): her chosen strategy of T yields an expected utility of 0.8 compared with an expected utility of 0.4 from choosing B. But if we calculate causal expected utilities, we find that T again yields 0.8, but B would yield 1.6. Thus she is not causally rational. If she were to play B instead of T, player 2 would also change

his strategy, leading to a better outcome most of the time. Just as with Bill Clinton's dilemma in section 2, the objects of uncertainty faced by the player (in this case her opponent's strategy) are not independent of the various acts available to her, and Bayes rationality gives us the 'wrong' result (that is, it does not coincide with causal rationality).

But there is something odd about the causal structure of this game. If the players are moving simultaneously, or at least in ignorance of each other's choice (as is often considered to be an implicit assumption of the normal form representation of a game), then their strategy choices *should* be independent of each other. Indeed, Harper [10] goes so far as to say "a causal independence assumption is part of the idealization built into the normal form", and Stalnaker [24] writes "... in a strategic form game, the assumption is that the strategies are chosen independently, which means that the choices made by one player cannot influence the beliefs or the actions of the other players". Similarly, appeal is often made to some causal independence condition to reject the symmetry argument for rational co-operation in the prisoner's dilemma: the two players will indeed end up doing the same thing, but if one were to deviate and cooperate, the other would still defect (see Dekel and Gul [7]). This causal independence condition is most easily expressed in the language of counterfactuals: if one player were to do something different, the others players would still do the same. We can state this condition as a formal property of a model:

(CI) for all w and x, for all i, if $x \preccurlyeq_w y$ for all y such that $f_i(y) = f_i(x)$, then $f_{-i}(x) = f_{-i}(w)$.

In other words, if at some world w in the model, player i plays strategy s_i and the other players play s_{-i}, then at the closest possible world in which i plays s_i' instead, the other players still play s_{-i}. It is clear that this gives us the causal independence condition stated above.

In the light of the preceding discussion, our first theorem should come as no surprise. It states that, as long as (CI) holds, Bayes rationality and causal rationality coincide.

Theorem 1 *In any model of G satisfying (CI), player i is Bayes rational at w if and only if player i is causally rational at w.*

Proof:

First we show that $[s_i \mapsto s_{-i}] = [s_{-i}]$ for all s_i, s_{-i}. Suppose $z \in [s_i \mapsto s_{-i}]$. It follows from the definition of $\mapsto$ that if $x \in [s_i]$ and for all $y \in [s_i]$, $x \preccurlyeq_z y$, then $x \in [s_{-i}]$. (S) guarantees that there is such an x. Since $x \preccurlyeq_z y$ for all y such that $f_i(y) = f_i(x)$, (CI) implies that $f_{-i}(x) = f_{-i}(z)$. Therefore $z \in [s_{-i}]$. Now suppose that $z \in [s_{-i}]$. Consider all the worlds $x \in [s_i]$. By (CI), if

$x \preccurlyeq_z y$ for all y such that $f_i(y) = f_i(x) = s_i$, then $f_{-i}(x) = f_{-i}(z) = s_{-i}$. Therefore $z \in [s_i \mapsto s_{-i}]$. So

$$\begin{aligned} & [s_i \mapsto s_{-i}] = [s_{-i}] \\ \Rightarrow \quad & [s_i \mapsto s_{-i}] \cap \{x : w\mathcal{B}_i x\} = [s_{-i}] \cap \{x : w\mathcal{B}_i x\} \\ \Rightarrow \quad & p_{i,w}([s_i \mapsto s_{-i}]) = p_{i,w}([s_{-i}]), \end{aligned}$$

for all s_i and s_{-i}, and in particular for $s_i = f_i(w)$. (The last step follows directly from the definition of $p_{i,w}(\cdot)$.) So the left hand sides of the expressions in the definitions of Bayes and causal rationality are equal to each other, as are the right hand sides.■

The intuition behind the proof is straightforward: the event that you play s_{-i} is just the same as the event that "if I were to play s_i you would play s_{-i}", since under (CI) my action has no causal influence on yours, and you'll carry on doing the same thing *whatever* I do. Thus my probabilistic evaluation of your various strategies is the same whether we hold my strategy fixed (as Bayes rationality does) or whether we vary it (as causal rationality does). Theorem 1 is an extremely convenient result[10]. It allows us to dispense with the causal apparatus developed above and continue using Aumann's simple model to analyze rational play in normal form games, as long as we have the required causal independence.

We must take care to distinguish the type of causal independence discussed above from independence of the probability functions, p_i. There are two possible independence conditions that might be imposed on the p_i functions. First, we might require that the probabilities one player assigns to the strategies of different opponents be independent of each other (Bernheim [6] and Pearce [16], among others, impose this constraint on player's beliefs). But, as Stalnaker [24] points out, our causal independence assumption "has no consequences about the evidential relevance of information about player one's choice for the beliefs that a third party might rationally have about player two". Consider, for example, a third player reasoning about a simple coordination game played by two twins. Even though fully convinced about the causal independence of the twins' strategy choices, she might expect that they end up playing the same strategy, whatever that might be. Hence there seems no reason in general to rule out players holding correlated conjectures about

[10] A similar result has been established by Shin [20], but in a very different framework to that of the current paper. Specifically, Shin constructs a space of *possible worlds* for each player, along with a personal closeness measure, to evaluate the counterfactual beliefs of that player about the unified state space. There is no representation of the (objective) causal reality, and hence no way of expressing causal independence.

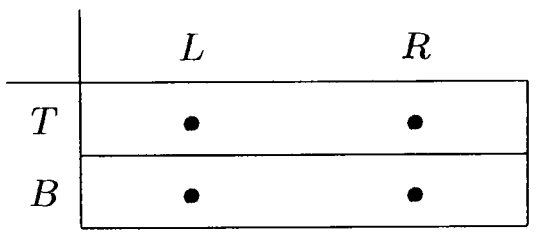
their opponents' strategies.

The other type of independence we might impose is between a player's own strategy choice and those of her opponents. Again, it is quite possible that I consider my own strategy choice evidentially relevant to that of my opponents, as is famously illustrated by Newcomb's problem. This can generate models in which a player is always rational, whatever strategy choice she makes, or indeed is never rational. Consider the game of *Simple Coordination* in figure 4 (a) below, where, as before, player 1 is choosing row and player 2 is choosing column.

	L	R
T	$(1,1)$	$(0,0)$
B	$(0,0)$	$(1,1)$

Simple Coordination

(a)



Player 1's partition ($\mathcal{B}_1$)

(b)

	L	R
T	0.4	0.1
B	0.1	0.4

Player 1's prior

(c)

	L	R
T	0.1	0.4
B	0.4	0.1

Player 1's prior

(d)

Figure 4

(c) represents the prior of the optimist — for her both strategies are Bayes rational; (d) gives the prior of the pessimist — for him neither is. Jeffrey [12] calls these cases "pathological", but we see no reason to exclude them. In any case, in the absence of any theory about the decision-making process itself, our models are perhaps best viewed merely as tools for the theorist to determine whether a given choice of an agent is rational, rather than anything to which the agent herself might appeal. The construction of a model of the latter type, as a means to explicating the decision-making process, seems to be Jeffrey's rather more ambitious goal; but Aumann states explicitly that this is not his aim: "The model describes the point of view of an outside observer" (p. 8). But for us there seems to be no reason to rule out the pessimist (though a psychologist might tell him to reconsider his beliefs). His example demonstrates a kind of non-existence of equilibrium: he will never be *a priori* happy with the choice he has made. Of course, this is not a case of non-existence of a rationalizable outcome. A strategy is rationalizable if there is *some* set of beliefs, satisfying certain conditions, for which it is Bayes rational., i.e. if it is rational for the agent in *some* model satisfying certain requirements.

The next obvious step is to consider how widely applicable the causal independence assumption is. Harper and Stalnaker, quoted above, claim that it is almost axiomatic for normal form games. The suggestion seems to be that the same might not apply for extensive form games (indeed, Harper goes on to make this point explicitly). But the normal form and the extensive form are merely alternative representations of a given situation of strategic interaction. What is really at issue is the move order, and in particular whether the players' moves in the game are simultaneous or sequential (or more precisely, whether each player moves in ignorance or her opponents' moves or not). Only in as much as the normal form is often used to represent simultaneous move games, while the extensive form is used when moves are made sequentially, might causal independence be appropriate for the former and not the latter. The reason is, of course, that the extensive form is more general: it contains all the information of the normal form, and more besides. Among the extra information is a description of the move order. And so it is to the extensive form that we turn in order to consider whether the causal independence assumption remains appropriate in sequential move games. Perhaps surprisingly, the answer shall be that is does, and so the non-causal structure of Aumann's model can be retained.

5 Extensive form games

The description of an extensive form game specifies the following five elements (see also Osborne and Rubinstein [15]):

- a finite set N of players. Note that we have not included *nature* in the set of players. This corresponds to the assumption that the game is one of complete information: there are no chance moves.

- a finite set H of sequences, which satisfies: (i) $\varnothing \in H$; and (ii) if $\left(a^k\right)_{k=1,\ldots,K} \in H$ and $L < K$, then $\left(a^k\right)_{k=1,\ldots,L} \in H$. Each $h \in H$ is a *history*, and each component of h is an *action* taken by a player. The set of histories defines the game tree, with each element h representing a node of the tree, the node that is reached is that history is played. A history $\left(a^k\right)_{k=1,\ldots,K} \in H$ is *terminal* if there is no a^{K+1} such that $\left(a^k\right)_{k=1,\ldots,K+1} \in H$. The set of actions available after the nonterminal history h is denoted $A(h) = \{a : (h,a) \in H\}$, and the set of terminal histories is denoted Z.

- a function $\imath : H \setminus Z \to N$ that assigns to each nonterminal history the player whose turn to

move it is.

- a partition $\mathcal{I}$ of $H \setminus Z$ that divides all the nonterminal histories into *information sets*. The cell $\mathcal{I}(h)$ of $\mathcal{I}$ that contains h identifies the nonterminal histories that the player on move cannot distinguish from h based on the information available to her at h. It is required that for every history in a given cell of the partition, the same player is on move and the same actions are available, i.e. if $h' \in \mathcal{I}(h)$, then $\imath(h) = \imath(h')$ and $A(h) = A(h')$. This is implied by the fact that each player knows when it is her turn to move, and what actions are available to her. Thus for any information set $I \in \mathcal{I}$ we can write $\imath(I)$ for the player on move, and A_I for the actions available to her, and we can partition $\mathcal{I}$ into sets $\mathcal{I}_i = \imath^{-1}(i)$.

- a utility function $U_i : Z \to \mathbb{R}$ for each player i, which assigns an expected utility value to each terminal history.

The collection $\langle N, H, \imath, \mathcal{I}, (U_i)_{i \in N} \rangle$ defines an extensive form game, Γ.

As we can see from the definition of H, the set of histories, each terminal history $h \in Z$ determines a path through the game tree. And h consists of a set of actions, each of which is chosen by the player on move at the history just reached. But actions choices are contingent on the particular history that has been reached: a given action may be available at several histories, but choosing that action at one history is not the same as choosing that action at another history (the two choices may not even be made by the same player); compare, for example, co-operating at the first round of a repeated prisoner's dilemma with co-operating at the last round, given that there has been defection in every previous round.

For this reason, players are usually thought of as choosing between strategies (see e.g. Ben Porath [5], Gul [9], Stalnaker [24]), which specify what action they will or would take at every information set at which they are on move. A strategy s_i for player i, then, is a function from $\mathcal{I}_i$ to A, with the property that $s_i(I) \in A_I$ for all $I \in \mathcal{I}_i$ (note that this definition implies that players must choose the same action at every history in a given information set: if a player cannot distinguish between two histories, it is assumed that she must make the same choice whichever is reached). It is clear that the causal independence assumption is still appropriate, because the causal dependencies between various actions are already built into the definition of a strategy: what would happen if a particular information set were to be reached does not depend on whether or not that information set is actually reached, and therefore the players' strategy choices are independent of each other.

There is, however, a problem with this view of strategies as the objects of choice. It is unclear when if ever players will actually make a choice between the various strategies available to them. Although we could think of a hypothetical pre-play stage when such choices are made, it seems more appropriate and more accurate to think of the players as making their choices as and when they are on move. Indeed, this is the approach that majority of the work in this area takes[11]. And at each information set a player chooses only part of her strategy, the part which specifies what she does at that information set. It is these choices that should be assessed as rational. For assessing the entire strategy at a particular information set carries with it the substantive assumption that the player on move has control over her choices at all future information sets. To see why, suppose we say that a particular strategy choice (rather than just the action choice) is rational at some information set. Presumably we mean that, among the strategies that are consistent with that information set's being reached, the strategy chosen maximizes expected utility[12]. All of these strategies specify what actions will be taken at future information sets as well as at the current information set. If the player cannot control what she does at these information sets while on move at the current information set, then she cannot choose among these strategies. This assumption of self control does not follow from rationality alone; rationality alone does not even imply that a player *knows* what she will do at future information sets![13]

So we shall think of the players as choosing between actions and not strategies when they are on move. To analyze these choices, we construct models of extensive form games, just like the models of normal form games. A model of an extensive form game specifies the following elements:

- a finite set W of states;
- for each information set I, a binary relation $\mathcal{B}_I$;
- for each information set I, a probability measure p_I on W;
- for each information set I, a function $f_I : W \to A_I$;
- for each state w, a binary relation $\preccurlyeq_w$.

[11] A notable exception is the work of Stalnaker: he discusses this issue in [25] (p. 315), and shows that, under certain assumptions, the two approches are equivalent.

[12] See e.g. Gul [9] "... rational players choose strategies s_i such that s_i is optimal at [an information set] against some conjecture that reaches [that information set] whenever s_i reaches [that information set]" (p. 15). The majority of papers in the Bayesian tradition adopt a similar definition of rationality.

[13] Nor even at past information sets: in games of imperfect recall, players can forget their previous action choices.

These models work in exactly the same way as the models of normal form games, except that we must specify probabilistic beliefs (given by $\mathcal{B}_I$ and p_I) and choices (given by f_I) at every information set $I \in \mathcal{I}$ in the game[14]. As before, we assume that the closeness relation $\preccurlyeq_w$ satisfies (C1) – (C4), and that the models satisfy (OK) and (S):

(OK) For all $I \in \mathcal{I}$ and for all $w, x \in W$, if $w \mathcal{B}_I x$ then $f_I(w) = f_I(x)$;

(S) For each information set I, for every $a \in A_I$, there exists a state w such that $f_I(w) = a$.

We have argued that the causal independence assumption is appropriate for strategy choices in extensive form games, and the same applies to action choices:

(CI) for all w and x, for all i, if $x \preccurlyeq_w y$ for all y such that $f_I(y) = f_I(x)$, then $f_{-I}(x) = f_{-I}(w)$,

where $f_{-I} = \{f_{I'}\}_{I' \neq I}$.

Before defining Bayes rationality and causal rationality, it will be useful to introduce some new notation. Let $A_{\mathcal{I}}$ denote the set of actions profiles, which specify an action for every information set in the game, and let A_{-I} denote the set of action profiles at every information set other than I. Clearly $A_{\mathcal{I}} = A_I \times A_{-I}$. Let $\zeta : A_{\mathcal{I}} \to Z$ denote the outcome function, which tell us which terminal history is reached for each action profile. Finally, let $u_i = U_i \circ \zeta$ denote i's utility as a function of $A_{\mathcal{I}}$. We can now define what it means for the choice made at a particular information set to be rational.

Definition 3 *The action chosen at information set I is* Bayes rational at w *if, for all $a \in A_I$,*

$$\sum_{a_{-I} \in A_{-i}} p_{I,w}([a_{-I}]) \cdot u_{\imath(I)}(f_I(w), a_{-I}) \geq \sum_{a_{-I} \in A_{-I}} p_{I,w}([s_{-I}]) \cdot u_{\imath(I)}(a_I, a_{-I}).$$

Definition 4 *The action chosen at information set I is* causally rational at w *if, for all $a \in A_I$,*

$$\sum_{a_{-I} \in A_{-I}} p_{I,w}([f_I(w) \mapsto a_{-I}]) \cdot u_{\imath(I)}(f_I(w), a_{-I}) \geq \sum_{a_{-I} \in A_{-I}} p_{I,w}([a_I \mapsto a_{-I}]) \cdot u_{\imath(I)}(a_I, a_{-I}).$$

Note that $\imath(I)$ is the player on move at information set I, so it is her utility function $u_{\imath(I)}$ that is relevant for these definitions.

[14] A more sophisticated model would derive the beliefs of a given player at every information set at which she was on move from a set of prior beliefs and assumptions about her *belief revision policy* (see Battigalli and Siniscalchi [4] and Stalnaker [24]).

So far we have defined what it is for an action to be rational; but we normally think of rationality as a property of players, and not of the actions they choose. It is not immediately clear what it means for a player to be rational in an extensive form game. In particular, we can distinguish *reached-node rationality*, where a player is rational if each action she actually plays is rational; *own-node rationality*, where a player is rational if each of her actions at nodes not ruled out by her previous behavior is rational; and *all-node rationality*, where a player is rational only if her actions are rational at every information set at which she is on move. For the present purposes, it does not matter which of these definitions we adopt: Theorem 2 holds for all of them.

Theorem 2 *In any model of* Γ *satisfying* (CI), *player* i *is Bayes rational at* w *if and only if player* i *is causally rational at* w.

The proof follows the same steps as the proof of Theorem 1, and is omitted. Theorem 2 extends the result of Theorem 1 to the case of extensive form games. As with normal form games, the causal independence condition allows us to analyze rational play in extensive form games using a non-causal framework.

6 Comments and conclusions

The Bayesian tradition in game theory adopts the view that the choices of rational players are the outcome of a process of expected utility maximization with respect to beliefs about everything that affects their payoffs. In particular, each player is assumed to have beliefs about the strategies played by her opponents. These beliefs are represented by probability distributions over a set of states of the world that is common to all players. A player is *Bayes rational* at a particular state if her strategy choice at that state is expected-utility maximizing given her beliefs about her opponents' strategies. But there is a problem with this notion of rationality: since each state describes what each player does as well as what her opponents do, the player will change the state if she changes her choice. There is no guarantee that her opponents will do the same in the new state as they did in the original state. A player is *causally rational* if her expected utility calculation takes this change into account. In this paper, we show that under a natural causal independence condition, Bayes rationality and causal rationality coincide. We argue that the causal independence condition is appropriate in extensive form games, where choices are made sequentially, as well as in normal form games, where it is usually assumed that they are made simultaneously. Thus the equivalence result justifies the use of Aumann's non-causal framework.

We conclude with an important *caveat.* In the causal expected calculus, counterfactuals enter only as aspects of players' (probabilistic) beliefs. The implication is that is that it is not causal independence itself that is important for the equivalence result, but rather players' beliefs in causal independence. Condition (CI) imposed causal independence at every state in the model, thus guaranteeing that the players also believed in causal independence at every state in the model: a local version of (CI) would not have been sufficient.

References

[1] AUMANN, R. J. (1987), "Correlated Equilibrium as an Expression of Bayesian Rationality", *Econometrica* **55**, 1-18.

[2] AUMANN, R. J. AND A. BRANDENBURGER (1995), "Epistemic Conditions for Nash Equilibrium", *Econometrica* **63**, 1161–1180.

[3] BATTGALLI, P. AND G. BONANNO (1999), "Recent Results on Belief, Knowledge and the Epistemic Foundations of Game Theory", *Research in Economics* **53**, 149–225.

[4] BATTIGALLI, P. AND M. SINISCALCHI (2001), "Strong Belief and Forward Induction Reasoning", *Journal of Economic Theory*, forthcoming.

[5] BEN PORATH, E. (1997), "Rationality, Nash Equilibrium and Backwards Induction in Perfect Information Games", *Review of Economic Studies* **64**, 23–46.

[6] BERNHEIM, B. D. (1984), "Rationalizable Strategic Behavior," *Econometrica* **52**, 1007–1028.

[7] DEKEL, E. AND F. GUL (1997), "Rationality and Knowledge in Game Theory," in D. M. Kreps and K. W. Wallis (eds.) *Advances in Economics and Econometrics: Theory and Applications: Seventh World Congress, Vol. 1.* Cambridge University Press, 87-172.

[8] GIBBARD, A. AND W. L. HARPER (1978), "Counterfactuals and Two Kinds of Expected Utility," in C. A. Hooker, J. J. Leach, and E. F. McClennan (eds.) *Foundations and Applications of Decision Theory, Vol. I.* Dordrecht, Reidel.

[9] GUL, F. (1996), "Rationality and Coherent Theories of Strategic Behavior", *Journal of Economic Theory* **70**, 1–31.

[10] HARPER, W. L. (1988), "Causal Decision Theory and Game Theory: A Classic Argument for Equilibrium Solutions, a Defense of Weak Equilibria, and a New Problem for the Normal Form Representation", in W. L. Harper and B. Skyrms (eds.) *Causation in Decision, Belief Change, and Statistics, II.* Kluwer Academic Publishers, Amsterdam.

[11] JEFFREY, R. C. (1965), *The Logic of Decision.* McGraw-Hill, New York, NY.

[12] JEFFREY, R. C. (1983), *The Logic of Decision*, second edition. The University of Chicago Press, Chicago, IL.

[13] LEWIS, D. (1973), *Counterfactuals.* Basil Blackwell, Oxford.

[14] LEWIS, D. (1981), "Causal Decision Theory", *Australasian Journal of Philosophy* **59**, 5–30.

[15] OSBORNE, M. J. AND A. RUBINSTEIN (1994), *A Course in Game Theory.* The MIT Press, Cambridge, MA.

[16] PEARCE, D. G. (1984), "Rationalizable Strategic Behavior and the Problem of Perfection", *Econometrica* **52**, 1029–1050.

[17] SAVAGE, L. J. (1954), *The Foundations of Statistics.* Wiley, New York.

[18] SELTEN, R. (1975), "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games", *International Journal of Game Theory* **4**, 25–55.

[19] SHIN, H. S. (1989), "Two Notions of Ratifiability and Equilibrium in Games", in M. Bacharach and S. Hurley (eds.) *Foundations of Decision Theory.* Basil Blackwell, Oxford.

[20] SHIN, H. S. (1992), "Counterfactuals and a Theory of Equilibrium in Games", in C. Bicchieri and M. L. Dalla Chiara (eds.) *Knowledge, Belief, and Strategic Interaction.* Cambridge University Press, Cambridge.

[21] SKYRMS, B. (1982), "Causal Decision Theory", *Journal of Philosophy* **79**, 695–711

[22] SOBEL, J. H. (1986), "Notes on Decision Theory: Old Wine in New Bottles Choice", *Australasian Journal of Philosophy* **64**, 407-437.

[23] STALNAKER, R. (1968), "A Theory of Conditionals", in N. Rescher (ed.), *Studies in Logical Theory.* Basil Blackwell, Oxford.

[24] STALNAKER, R. (1996), "Knowledge, Belief and Counterfactual Reasoning in Games", *Economics and Philosophy* **12**, 133-163.

[25] STALNAKER, R. (1999), "Extensive and Strategic Form Games: Games and Models for Games", *Research in Economics* **53**, 293–319.

Chapter 4

The Deception of the Greeks:

Generalizing the Information Structure of Extensive Form Games*

"You are to hear now how the Greeks tricked us. From this one proof of their perfidy you may understand them all" (Aeneas).

Abstract: The standard model of an extensive form game rules out an important phenomenon in situations of strategic interaction: deception. Using examples from the world of ancient Greece and from modern-day Wall Street, we show how the model can be generalized to incorporate this phenomenon. Deception takes place when the action observed by a player is different from the action actually taken. The standard model does allow imperfect information (modeled by non-singleton information sets), but not deception: the actual action taken is never ruled out. Our extension of extensive form games relaxes the assumption that the information sets partition the set of nodes, so that the set of nodes considered possible after a certain action is taken might not include the actual node. We discuss the implications of this relaxation, and show that in certain games deception is inconsistent with common knowledge of rationality even along the backward induction path.

1 Deception in games

Although the Trojan war pre-dates the formal study of games by almost three thousand years, the Greek generals clearly possessed a sound understanding of the basic principles of game theory. Odysseus, in particular, was a master: his dealings with the Sirens, for example, provide an excellent illustration of the value of commitment; returning home at last, he disguises himself as a beggar to collect information about his wife's suitors, only revealing his true identity when the time is

*I am grateful to Michael Bacharach, Alexandru Baltag, Steven Brams, Laurence Emmett, Erik Eyster, Alexander Gümbel, Meg Meyer, Llewelyn Morgan, Rani Spiegler and Daniel Zizzo, as well as participants at the Fifth Conference on Logic and the Foundations of Game and Decision Theory, the Fifth Spanish Game Theory Meeting, Nuffield College and the Computing Laboratory, Oxford for helpful advice.

right, thus trading short-run losses for long-run gains. It would be possible to write an entire game theory text book using the Greek myths as a basis. The current aim is more modest: to construct a game-theoretic model of a single incident at the end of the Trojan war, when the Greeks tricked the Trojans by abandoning their camp and sailing behind the island of Tenedos. We claim that the standard definition of an extensive form game is too restrictive to capture an important feature of this story, namely *deception.*

Deception takes place when one player tricks another into believing that she has done something other than what she actually did. In this case, the Greeks remain in the vicinity of Troy but out of sight behind Tenedos so that the Trojans believe they have sailed home. This phenomenon is ruled out by the way information is modeled in extensive form games. The standard structure of an extensive form game does allow actions to be uninformative (whenever information sets are non-singleton), in that it is not revealed which of several actions has been taken. But they cannot be deceptive: the actual action taken is never ruled out. Relaxing the assumption that the information sets partition the set of nodes allows deception to take place. In particular, the set of nodes considered possible after a certain action is taken might not include the actual node. In section 2 we show how a game with a non-partitional information structure can be used to represent the story above, and consider a more recent example of deception.

Recent work on decision problems of imperfect recall such as the absent-minded driver problem (introduced by Piccione and Rubinstein [14]) has suggested that the interpretation of information sets in extensive form games is not straightforward. If non-partitional information structures are allowed, things become even less clear. Section 3 reviews the standard definition of an extensive form game, and shows how the information structure can be generalized. In section 4 we comment on various issues of interpretation in the generalized model. The notion of equilibrium in these games is also discussed. Section 5 reviews related literature, and some conclusions are offered in section 6.

2 Two examples

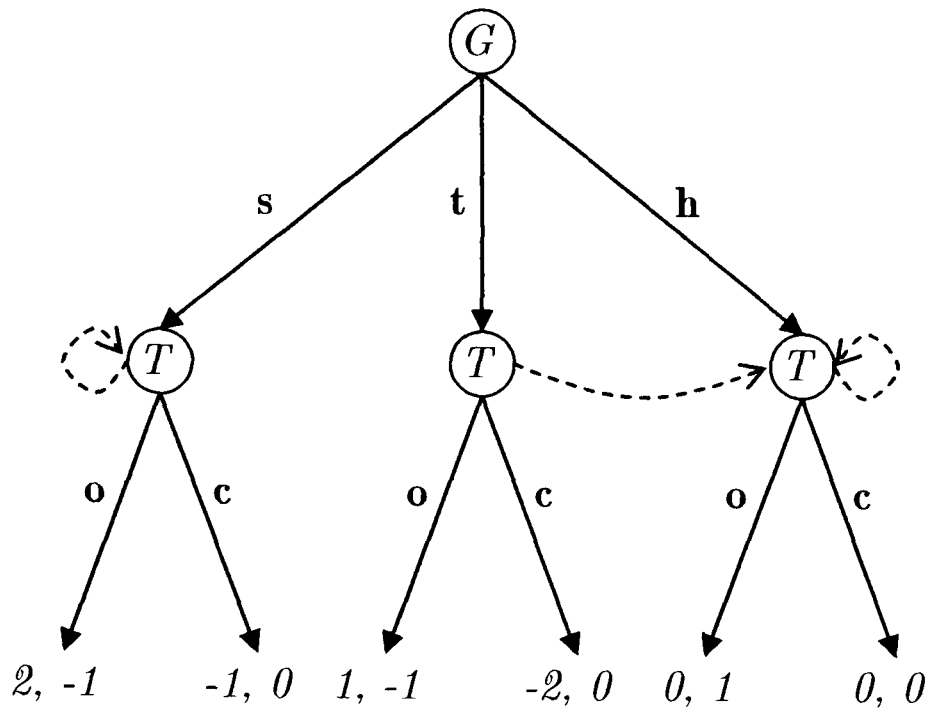
2.1 The Trojan war

> "Within sight of Troy is the island of Tenedos. In the days of Priam's Empire it had wealth and power and was well known and famous, but there is nothing there now, except the curve of the bay affording its treacherous anchorage. The Greeks put to sea

> as far as Tenedos, and hid from sight on its lonely beaches. We thought they had sailed for Mycenae before the wind and gone home. So all the land of Troy relaxed after its years of unhappiness. We flung the gates open and we enjoyed going to look at the unoccupied, deserted space along the shore where the Greek camp had been". (Aeneas, quoted in *The Aeneid* Book II [18].)

In Book II of *The Aeneid*, Vergil tells the story of how the Greeks gained entry to the city of Troy by means of a trick. After ten years waging an unsuccessful war, the Greeks considered their options: to go home and give up the war or to stay and attempt to sack Troy. The latter seemed hopeless until one of their number, Prylis, suggested the following plan: they should sail their ships out of sight behind the island of Tenedos and leave a gigantic wooden horse in front of the city. Believing the Greeks had really gone home, the Trojans accepted the horse as a gift and broke down their walls to wheel it into Troy. The Greeks then leapt out of the horse and successfully sacked the city. Deception was essential for the success of this plan. The Trojans were highly suspicious of the wooden horse and would not have accepted the it into their city unless they really believed the Greeks had gone home.

To model the deception of the Greeks, we allow them *three* action choices at the beginning of the game: to go home (**h**); to sail behind the island of Tenedos (**t**); and to stay put (**s**). In each case, the Trojans can choose to open up their gates and accept the wooden horse (**o**), or to keep them closed and reject it (**c**). We represent the information structure of the game in the usual way, by information sets: if player i is on move at node x, then $\mathcal{I}(x)$ lists the set of nodes she considers possible given the information at that time (i.e. $\mathcal{I}(x)$ is the smallest set of nodes in which she is sure to find the actual node). The Trojans' information sets at their three decision nodes are: $\mathcal{I}(s) = \{s\}$; $\mathcal{I}(h) = \{h\}$; and $\mathcal{I}(t) = \{h\}$ (using the obvious notation). In other words, if the Greeks stay, the Trojans can see that they have stayed; if the Greeks sail away, the Trojans can see that they have sailed away; but if the Greeks sail to Tenedos, it seems to the Trojans as if they have sailed away. It is clear these information sets do not partition the set of decision nodes of the Trojans, and hence this game does not fit the standard definition of an extensive form game. Using dotted arrows to represent the information structure, the game is shown in the diagram below. Payoffs to the Greeks are given first. They prefer to stay if and only if the Trojans open the gates, and suffer minor inconvenience from sailing behind the island. The Trojans prefer to open the gates if and only if the Greeks go home.



The Trojan War

The game can be solved by backward induction. At node s, the Trojans know they are at node s and will keep the gates closed; at nodes t and h, the Trojans believe they are at node h and will open the gates. If the Greeks are know the Trojans are rational, they will be aware of this, and will choose to sail behind the island. And indeed, this is what happened.

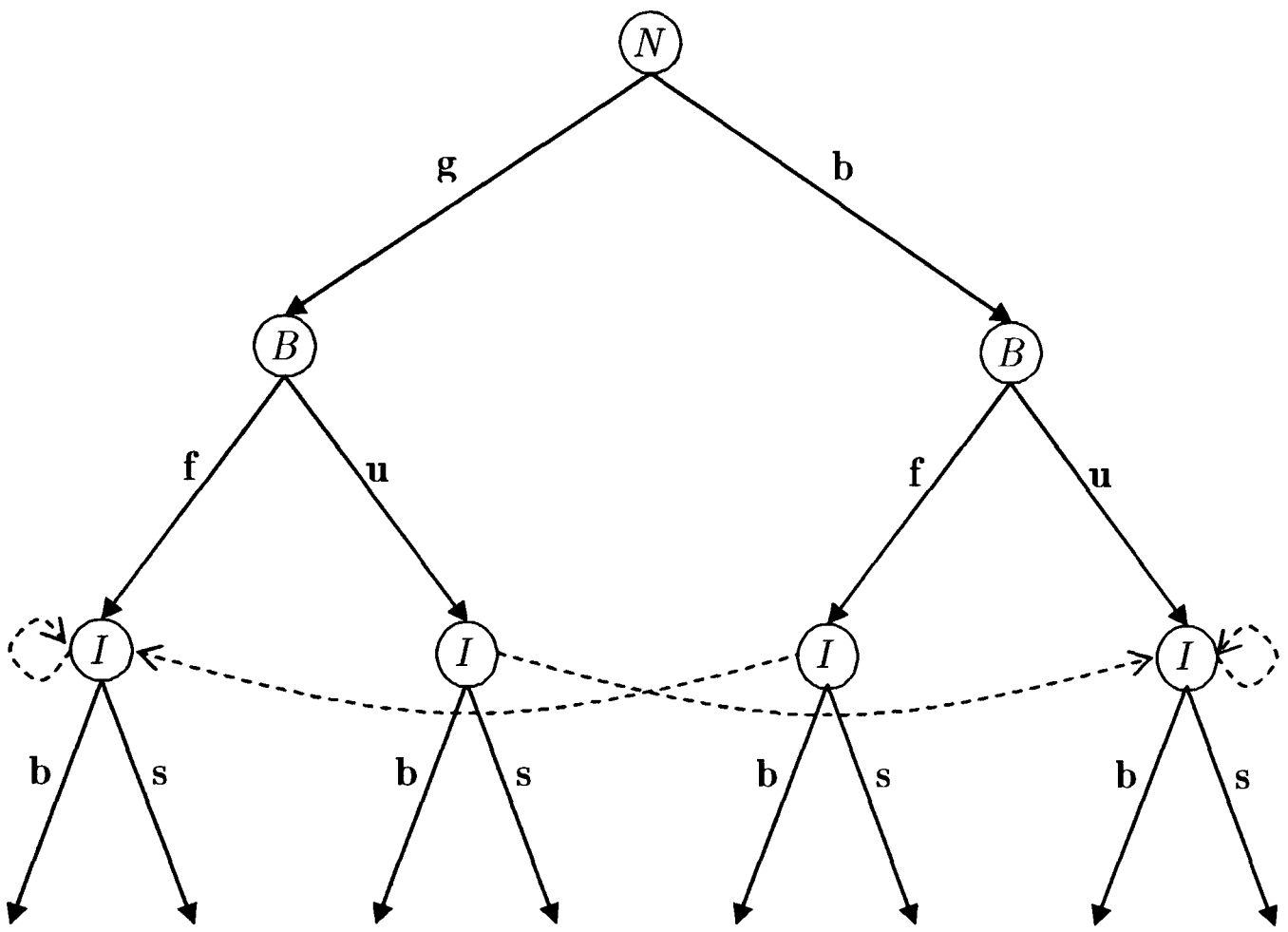
2.2 Investment advice

On 22nd May 2002, *The Times* newspaper ran the following story

> Merrill Lynch agreed today to pay fines of $100 million (£68.7 million) to help settle charges that it was telling clients to buy stocks it secretly believed were "junk"... The deal followed an investigation by New York Attorney General Eliot Spitzer alleging Merrill Lynch gave overly optimistic opinions of companies to win them over as investment banking clients.

The diagram below gives a stylized representation of this situation. Nature (N) moves first and determines whether the stock is good (**g**) or bad (**b**); the bank (B) observes the quality of the stock, and makes a report to the potential investor (I), which can be favorable (**f**) or unfavorable (**u**); finally the investor, having heard the report but in ignorance of the true quality of the stock, decides whether to buy (**b**) or sell (**s**). The payoff structure is very simple: the bank cares only about creating business for itself, so prefers for the investor to buy; the investor wants to buy if and only is the stock is actually good. As before, information is represented by dotted arrows: the investor always believes the bank's report, so that she thinks the stock is good whenever the

bank gives a favorable report, and that the stock is bad whenever the bank gives an unfavorable report. Letting gf, gu, bf, and bu denote the investor's decision nodes, her information sets are $\mathcal{I}(gf) = \mathcal{I}(bf) = \{gf\}$; and $\mathcal{I}(gu) = \mathcal{I}(bu) = \{bu\}$.



Investment Advice

Again we use backward induction to solve the game. At nodes gf and bf, the investor believes she is at node gf, and plays **b**; at nodes gu and bu, the investor believes she is at node bu and plays **s**. The bank will therefore play f whatever the move by nature at the start of the game. Favorable reports will be issued even for stocks which are known to be bad.

2.3 Deceived or simply mistaken?

Is it really the case that standard extensive form games with partitional information sets cannot represent either of the situations described above? It could be argued that the Trojans and the clients of Merrill Lynch were not deceived; rather, they were simply mistaken about which of two possible nodes had been reached. We consider each case in turn.

Adopting this line of argument, the information sets in the Trojan War game would be $\mathcal{I}(s) = \{s\}$ and $\mathcal{I}(h) = \mathcal{I}(t) = \{h, t\}$. These sets partition the Trojans' decision nodes. It is easily verified that there is a unique perfect Bayesian equilibrium in which the Greeks mix between going home

and sailing behind the island, and the Trojans mix between opening the gates and keeping them closed. But the historical outcome of the game cannot be explained as a particular realization of these mixed strategies. For in this mixed strategy equilibrium, Bayesian updating requires that the Trojans assign equal probability to nodes h and t, since the Greeks are mixing 50–50. Yet we are told that they "thought they had sailed for Mycenae before the wind and gone home".

An alternative explanation is that we are observing out-of-equilibrium behavior: the Trojans were simply mistaken about the Greeks' strategy choice and failed to play a best response. In fact, the observed outcome is rationalizable (i.e. consistent with common knowledge of rationality) in the game with standard information sets. The beliefs that rationalize this outcome are as follows:

G The Greeks believe that the Trojans will open their gates, and that the Trojans believe the Greeks have gone home;

T The Trojans believe that the Greeks have gone home, and that the Greeks believe that the Trojans will keep their gates closed.

But this story can give no explanation of why the Trojans came to have these beliefs: many other beliefs are also rationalizable. Indeed they were warned by the priest Laocoön that the Greeks had not gone home: "Do you really believe that your enemies have sailed away?... I still fear Greeks, even when they offer gifts". They ignored his advice because they were deceived: "we gave Sinon [one of the Greeks] our trust, tricked by his blasphemy and cunning". A related point is that rationalizability as a solution concept has little predictive power in the standard game; in the game of deception, on the other hand, there is a unique rationalizabe outcome.

It is even harder to tell a coherent story about what is going on in the investment advice game using standard information sets. The information sets of the investor would be: $\mathcal{I}(gf) = \mathcal{I}(bf) = \{gf, bf\}$; and $\mathcal{I}(bf) = \mathcal{I}(bu) = \{bf, bu\}$. There is a mixed-strategy perfect Bayesian equilibrium in which false advice is given and followed with positive probability, but as before this does not reflect what actually happened (i.e. that the investor believed *all* of the bank's reports). The observed outcome is rationalizable, but only by very odd beliefs on the part of the investor:

B The bank believes that the investor will invest if and only if the report is favorable, and that the investor believes the bank will give a favorable report if and only if the stock is good;

I The investor believes that the bank will give a favorable report if and only if the stock is good, and that the bank believes the investor will ignore all reports.

The reason is that the bank cares only about persuading the investor to buy rather than sell; thus a conditional reporting strategy (one in which the report issued depends on whether the stock is good or bad) can be rational only if the content of the report does not affect the investor's action. Furthermore, once again rationalizability allows a whole range of alternative outcomes in the standard game, in contrast to the unique prediction in the game of deception.

3 A generalization of extensive form games

The games discussed in section 2 differ from standard extensive form games only in their information structure. The assumption that information sets partition the decision nodes is relaxed. The following definition is adapted from Osborne & Rubinstein [13]. Note that here we consider only finite games; the extension to the infinite case is straightforward.

Definition 1 *An* extensive form game with generalized information structure *is a tuple*

$$\left\langle N, H, P, f_c, \mathcal{I}, (u_i)_{i \in N} \right\rangle,$$

where

- N *is a finite set of players*

- H *is a finite set of sequences such that* (a) $\varnothing \in H$; *and* (b) *if* $\left(a^k\right)_{k=1,\ldots,K} \in H$ *and* $L < H$, *then* $\left(a^k\right)_{k=1,\ldots,L} \in H$

  Each member of H *is a* history; *each component of a history is an* action *taken by a player. A history* $\left(a^k\right)_{k=1,\ldots,K} \in H$ *is* terminal *if there is no* a^{K+1} *such that* $\left(a^k\right)_{k=1,\ldots,K+1} \in H$. *The set of actions available after the nonterminal history* h *is denoted* $A(h) = \{a : (h, a) \in H\}$ *and the set of terminal histories is denoted* Z.

- P *is a function which assigns to each nonterminal history a member of* $N \cup \{c\}$. P *is the* player function, *and* $P(h)$ *is the player who takes an action after history* h; *if* $P(h) = c$ *then chance determines which action is taken after history* h.

- f_c *is a function which associates with every history* h *for which* $P(h) = c$ *a probability measure* $f_c(\cdot \mid h)$ *on* $A(h)$. $f_c(a \mid h)$ *is the probability that* a *occurs after history* h. *Each probability measure is independent of every other such measure.*

- $\mathcal{I}$ *is a function which assigns to each nonterminal history a nonempty set of nonterminal histories such that if* $h' \in \mathcal{I}(h)$, *then* (a) $P(h) = P(h')$; *and* (b) $A(h) = A(h')$. $\mathcal{I}$ *is the* information function, *and* $\mathcal{I}(h)$ *is the set of histories that player* $P(h)$ *considers possible if the true history is* h. *Condition* (a) *says that a player always knows when she is on move, and condition* (b) *says that she also knows what actions are available to her.*

- u_i *is a function from* Z *to* $\mathbb{R}$, *the* utility function *of player* i.

An information function is more general than an information partition, in the sense that every information partition can be represented by an information function, but not every information function can be represented by an information partition unless we impose additional constraints on the form of $\mathcal{I}$. More precisely, if we assume that (c) $h \in \mathcal{I}(h)$, and (d) if $h' \in \mathcal{I}(h)$, then $\mathcal{I}(h) = \mathcal{I}(h')$, then we can find a partition which represents $\mathcal{I}$. Condition (c) rules out the possibility of deception (see Definition 2 below); the interpretation of condition (d) is discussed in Section 4.2.

We can use the information function to provide a taxonomy of a player's information whenever she is on move.

Definition 2 *The player* $P(h)$ *on move after history* h *is:*

(a) perfectly informed *if* $\{h\} = \mathcal{I}(h)$;

(b) imperfectly informed *if* $h \in \mathcal{I}(h)$ *and* $|\mathcal{I}(h)| > 1$;

(c) deceived *if* $h \notin \mathcal{I}(h)$.

4 Comments

In this section, we discuss various issues of interpretation concerning games of deception, and address some potential criticisms. First, we question whether the standard assumption that the structure of the game is common knowledge is coherent in these games. Next, we argue that a deceived player need not be irrational, and discuss various forms of bounded rationality. We then ask how it is that deception might arise. Finally, we consider what might be an appropriate solution concept for games of deception.

4.1 Can there be common knowledge of the game?

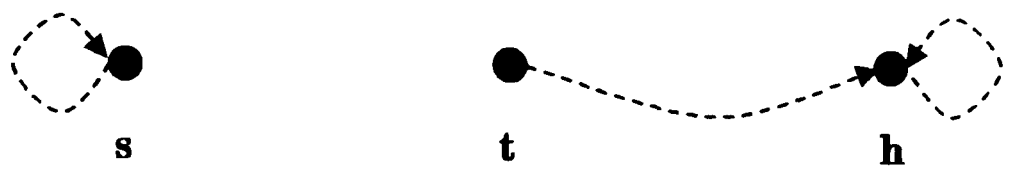
Game theorists standardly assume that the structure of the game, i.e. everything specified in Definition 1 above, is common knowledge among the players. This assumption is crucial is we are to make sense of standard solution concepts. Indeed the very notion of rationality as expected utility maximization presupposes that the rational player knows which options are available to her and what her utility function is. But surely it is not possible for a player to know that she has been deceived? And surely if a player knows that she *will* be deceived, this will undermine the deception? In fact, the first statement is true but the second need not be, and neither rules out common knowledge of the game. A systematic analysis of these issues requires a formal model of the player's knowledge and beliefs.

Let us consider again the first example, *The Trojan War*. There is nothing incoherent in the assumption that both the players have common knowledge of the structure of the game (tree, information and utilities) before any moves are made, although it might seem strange that the Trojans know the Greeks have the option of sailing behind the island and at the same time know that they will be deceived if the Greeks do so. Reinterpreting **t** as "trick" rather than "Tenedos" removes this awkwardness: if the Trojans do not know the exact form of the trick, it is more reasonable to suppose that they will be taken in by it. But what about the Trojans knowledge and beliefs when they come to move? At node t, is it possible for them to be deceived, while retaining knowledge of the structure of the game? We shall construct an *epistemic model* for the Trojans which shows that it is.

An epistemic model for a player tells us what that player knows and believes at a certain point in the game. It consists of a set of *states*, W; a *history function*, $\mathcal{H} : W \rightarrow H$, which tells us which history has been reached at each state; and an *accessibility relation* R, which tells us which states the player considers possible (i.e. if wRw', then if w is the true state, the player considers state w' possible)[1]. A player believes something if it is true at every state she considers possible, and we shall assume (rather crudely) that she knows something if she believes it and it is true. Let B_i and K_i stand for "player i believes that ..." and "player i knows that..." respectively. Consider an epistemic model with three states, $W = \{1, 2, 3\}$, with $\mathcal{H}(1) = \mathbf{s}$, $\mathcal{H}(2) = \mathbf{t}$, $\mathcal{H}(3) = \mathbf{h}$; and $1R1$, $2R3$, $3R3$. A diagrammatic representation is given below (note that here the arrows represent the

[1] Readers familiar with modal logic will recognize that an epistemic model is essentially a Kripke structure, with the history function playing the role of the interpretation; those not are referred to Fagin *et al.* [10] for a very detailed explanation. Stalnaker [17] shows how epistemic models can be used to analyze rational play in games, and to provide a systematic evaluation of game-theoretic solution concepts.

accessibility relation, not the information function of the game as before).



An epistemic model for the Trojans

To see how the epistemic model works, suppose that the true state is 2, i.e. that Greeks have sailed behind the island of Tenedos. Then the only state the Trojans consider possible is state 3, in which the Greeks have gone home. Thus at state 2, the following sentences are true: **t**, $B_T\mathbf{h}$. The information structure of the game can be summarized by the following three sentences: $\mathbf{s} \rightarrow B_T\mathbf{s}$, $\mathbf{t} \rightarrow B_T\mathbf{h}$, and $\mathbf{h} \rightarrow B_T\mathbf{h}$. It is easy to check that these sentences are true at *every* state. In particular, they are true at state 2, the true state, and at state 3, the only state the Trojans consider possible. Thus at state 2, the Trojans know that all three sentences are true: $K_T\left((\mathbf{s} \rightarrow B_T\mathbf{s}) \,\&\, (\mathbf{t} \rightarrow B_T\mathbf{h}) \,\&\, (\mathbf{h} \rightarrow B_T\mathbf{h})\right)$. They have been deceived, and yet retain knowledge of the structure of the game. Indeed, they know that they know it, and know that they know that they know it, and so on. Intuitively, although the Trojans know that *if* the Greeks play **t** they will be deceived, this does not prevent them from being deceived when it actually happens.

A more detailed investigation of what the players can know and believe about each other and about the structure of the game would require a more sophisticated epistemic model, representing the beliefs of all the players at every stage of the game. Such models can be found in Board [3]. But the current aim is merely to convince the reader that deception is not inconsistent with common knowledge of the game. It is hoped that the toy model above is sufficient for this purpose.

4.2 Deception, lack of introspection, and unawareness

We have shown that it is not incoherent to assume that the Trojans knew the structure of the game and yet were still deceived. But is a deceived player necessarily an irrational one? In this section we show that the answer to this question is no, and distinguish between three forms of bounded rationality.

The claim that the Trojans must be irrational could be based on the following argument: if the Trojans know the structure of the game, then they know that whether the Greeks play **t** or whether they play **h**, they will believe that the Greeks have played **h**. So if they find themselves believing that the Greeks have played **h**, they should remain open to the possibility that the Greeks actually

played **t**. But this argument merely denies that the Trojans were deceived, and contradicts the structure of the game. There is nothing irrational is realizing that something could have happened for two reasons, but ruling out the first.

A more subtle argument could be based on the additional premise that the Trojans knew that the Greeks were rational, and that the Greeks knew that the Trojans were rational. If this is so, they should be able to carry out the backward induction argument we used to solve the game in section 2.1, and conclude that the Greeks would play **t**. This is perfectly true, but all it tells us is that the additional premise is inconsistent with node t being reached. At this node common knowledge of rationality breaks down[2]. The idea that common knowledge of rationality may not survive along every path through an extensive form game is not a new one, and is discussed in detail by Reny [15] and many others. In games of deception it is possible that common knowledge of rationality cannot survive along *any* path. This is true of *The Trojan War* but not of *Investment Advice.*

There is a sense, however, in which a player who is deceived must be only boundedly rational. If we think of the set of states in an epistemic model as representing every possible contingency in a particular situation, and the accessibility relation as describing what signal a player receives in each state, then a fully rational player should be able to invert that signal to figure out what state it might have come from. This inversion process will generate a new accessibility relation which partitions the set of states. If beliefs are defined in the same way as before, then everything this fully rational player believes must be true. Of course people in the real world do have false beliefs: it may be that there are simply too many contingencies for us to be able to consider every one of them. This idea is developed further in the next section when we consider how deception might arise.

Information function which satisfy conditions (c) and (d) can be represented by standard information partitions. We have seen that relaxing (c) allows us to model deception. Relaxing (d) gives another form of bounded rationality, *lack of introspection.* A player who lacks introspection does not know all of her own beliefs, i.e. there must be something she believes but does not know

[2] If this feature of *The Trojan War* is thought to be unpalatable, a modification of the game allows to the Trojans to retain their knowledge that the Greeks are rational and know that they are rational, even when deceived. Simply add a move by nature to the beginning of the game, according to which it is determined whether the Greeks have a choice between **s**, **t**, and **h** or just a choice between **s** and **h**. the information structure is such that whenever **t** is played or **h** is played in either game, the Trojans believe that **h** is played *in the smaller subgame.* Common knowledge of rationality can survive at this node. Intuitively, the Trojans are unsure whether the Greeks have a trick they can play or not; when the Greeks actually play the trick, they assume that it was not available.

she believes; or something she does not believe but does not know she does not believe. The link between (d) and introspection follows from a well-known theorem in modal logic (Theorem 3.1.5 in Fagin *et al.* [10]). For the present purposes, it is sufficient to point out that conditions (c) and (d) are logically distinct: a player may be deceived but introspective (as is the case in both of the examples above), or lack introspection even if not deceived.

A final form of bounded rationality is *unawareness*. In many cases, a very plausible explanation for deception is that the deceived was not aware of the possibility of a particular move being made[3]. To model unawareness we would need a richer framework than that discussed above. The assumption of common knowledge of the game must be relaxed and a distinction made between the actual game that is being played and the game as it appears to each player. A more detailed discussion of the difficulties in modelling unawareness can be found in Dekel *et al.* [9], who show that there is no way of representing a plausible notion of unawareness using standard epistemic models and propose an alternative approach.

4.3 How and why does deception occur?

We have so far begged the questions of how and why a player might be deceived. Games of deception encode into the information structure itself the fact that deception occurs. We have argued that these games provide a more satisfactory model of certain situations than games with standard information structures. But the fact that deception occurs is a modelling assumption, and not something derived endogenously. It is therefore important to understand when it might be a reasonable assumption. This is perhaps as much a question for the psychologist as the economist, but we offer two suggestions.

The first is that in the absence of any contradictory evidence, people tend to take their observations at face value. Thus when the Trojans see the Greeks abandon their camp and sail away out of sight, it is natural for them to believe that they have gone home. Similarly, when a bank tells someone that a stock is good, it is natural for them to believe that the stock really is good. Brams [5] adopts this line of argument when he writes "Since Deceived is in possession of no information — in particular, information that would conflict with Deceiver's announcement — there is no reason for him not to believe Deceiver". (This work is discussed in section 5.) Of course in the cases

[3]This does not explain why the Trojans were deceived by the Greeks. The Trojans were certainly aware of the existence of the island of Tenedos, and of the fact that the Greeks might trick them (Laocoön made this quite clear to them).

discussed above it could be argued that there *are* good reasons to disbelieve ones eyes and ears. Deception will fail if these reasons outweigh the natural tendency to believe.

The second suggestion is based on complexity considerations and the boundedness of human minds. There are many possible explanations for the epistemic inputs a particular individual receives, and typically only a small subset can be considered. If this subset excludes certain actions that might be made by others, then the individual will form mistaken beliefs if these actions are taken. In the context of *The Trojan War*, it is plausible that although the Trojans knew the Greeks might play a trick on them, they did not know the exact form the trick might take: they were not able to consider every possible trick, and in particular, they did not think about the Greeks sailing behind the island. This idea can also explain why people are not usually taken in by the same trick twice, and can shed light on the nature of recent corporate deceipts such as the WorldCom and Enron scandals, in which complex networks of companies were set up to hide costs and losses. If this line of argument is taken, our games of deception could be thought of as simplifications of a more complex games in which some players are unaware of some moves. This simplification allows us to model the essential features of a given situation while retaining the standard game-theoretic assumption of common knowledge of the game.

4.4 Solution concepts for games of deception

The two examples we considered in section 2 were simple enough for backward induction to yield unique solutions. But in more complex games this will not be the case, and it is important to discuss what solution concepts might be appropriate in the general case. An obvious proposal is to use sequential equilibrium, with strategies defined as functions from information sets to actions in the normal way. But there is a problem with the interpretation of the consistency requirement of sequential equilibrium in games of deception. In *The Trojan War*, for example, the Trojans have a dominant strategy of (**c** if $\{s\}$; **o** if $\{h\}$), and the Greeks' best response to this strategy is to play **t**. Thus these strategies must be played in any equilibrium. Yet the information structure of the game dictates that the Trojans must believe that they are at node h after the Greeks have played **t**. Although this belief is consistent in the formal sense (since given *any* strictly mixed strategy for the Greeks, Bayesian updating will assign a probability of one to node h, the only node in the Trojans' information set), it is not in the spirit of equilibrium analysis, which assumes that everyone knows what everyone else is doing. Of course there can be no equilibrium in this sense, since the very nature of deception is that the deceived player does not know what the deceiver is doing! We do not

however advocate rejecting the use of sequential equilibrium in games of deception; we merely wish to point out that careful analysis is required to clarify its implications. The framework developed by Board [3] provides an ideal tool for this purpose.

5 Related literature

In the economics literature, discussions of deception can be divided into three main strands. In the first we find *cheap talk* games, introduced by Crawford and Sobel [8]. In these games players communicate by means of costless messages, and can lie in order to gain strategic advantage. But deception can never succeed in equilibrium, where the sender's strategy is assumed to be known, and any potentially harmful message will be ignored.

The second strand is represented by Sobel [16] and Benabou and Laroque [2], among others. Again messages are costless to send, but here deception can succeed because of the existence of *honest* types (who never lie) alongside the standard *opportunistic* types (who can lie). If the proportion of honest types is high enough, it may be worthwhile to believe messages even though there is a chance they could be deceptive.

Crawford [7] is an example of the final strand. Here the messages are no longer cheap talk (and so could be interpreted as actions of any form rather than just statements) and no-one is inherently honest, but successful deception can take place because there are *mortal* as well as *sophisticated* players. In equilibrium, sophisticated players as assumed to know each other's strategies as usual, but mortal players can have arbitrary beliefs about what everyone else is doing. The upshot is that even sophisticated players can deceive each other, if each thinks it sufficiently likely that she is facing an mortal opponent.

While there is much to be learned from all of these stories, we think it is an advantage of our approach that deception can be modeled in a parsimonious framework, without recourse to artificial devices such as hypothetical types of player.

More closely related to the current project is work of the political scientist Brams ([5], [6] and elsewhere). Brams models deception in normal form games, and assumes that it takes the form of a misrepresentation of preferences by one agent (Deceiver) to another (Deceived). More precisely, he assumes that "Deceived has no *a priori* information about the preferences of Deceiver... [and] Deceived believes Deceiver's announcement of his preferences (true or misrepresented), and

Deceived knows that he does" (Brams [5]). He goes on to show that 33 of the 78 2×2 games[4] are *deception-vulnerable*, in the sense that Deceiver can obtain a better outcome if he misrepresents his preferences than if he reports them truthfully. A distinction is made between *tacit* deception, when Deceiver's strategy choice does not reveal his deceipt (i.e. his choice is consistent with his stated preferences), and *revealed* deception, when it does. Brams provides a detailed analysis of the Cuban Missile Crisis within this framework.

The generalization of extensive form games described in section 3 provides a way to formalize Brams' assumption that statements about preferences are always believed. Uncertainty about Deceiver's preferences can be represented by a move by Nature at the start of the game, in which the true preferences are chosen. Deceived does not observe this move, but does observe a statement made Deceiver, and the information structure is such that only subgames consistent with Deceiver's statement are considered possible by Deceived. Whenever that statement is false, this will not include the actual subgame, hence the information function will not satisfy condition (c). Note that in Brams' framework deceptive actions are always cheap talk: the original statement by Deceiver does not affect actual payoffs. But it does affect Deceiver's beliefs about these payoffs, and so deception is possible even in equilibrium notwithstanding the results of Crawford and Sobel [8].

6 Conclusions

There can be little doubt that deception is an important feature of strategic interaction. The proliferation of corporate scandals at some of America's highest-profile firms in recent months (including Enron, Global Crossing, Tyco, Qwest and WorldCom) has prompted George W. Bush to accuse executives of "breaching trust and abusing power", and he has pledged to "end the days of cooking the books, shading the truth and breaking our laws". But deceipt is by no means a new phenomenon in the financial world. Benabou and Laroque [2] tell a story of the banker Nathan Rothschild. Rothschild had a network of carrier pigeons which gave him superior information from France, and in 1815 during the battle of Waterloo he walked around the city of London looking dejected, spreading the news that the battle was going badly, and arranging for his agents to make a public display of selling British government securities. At the same time Rothschild was secretly

[4] There are 78 combinations of pairs of strict preferences orderings over the four outcomes of a 2×2 game, modulo permutations of player and strategy labels.

buying much larger quantities of these securities at the depressed price, waiting for the time when news of the victory would finally reach the masses.

In the military world, the Greeks set a trend that has lasted until the present day. Herodotus [11] describes how Zopyrus mutilated himself, cutting off his nose and ears, in order to convince the Babylonians that he was a deserter from the Persian army; the lies he told facilitated the Persian capture of Babylon, and Zopyrus was made Governor as a reward. More recent examples include the Allied invasion of Normandy on D-Day, June 6 1944, after a feint at Calais had convinced the Germans they would land there (see Kemp [12]); and the misrepresentation of American preferences by John F. Kennedy during the Cuban Missile Crisis in 1962 (analyzed by Brams [5]). Political life is a rich source of further examples: George Bush Senior's 1988 campaign promise, "Read my lips: no new taxes" is one of the more obvious.

The aim of this paper has been to argue that the standard information structure of extensive form games is not able to capture the notion of deception, and to show how replacing information partitions with the more general information functions can provide a solution to this problem. We conclude with a brief discussion of several alternative motivations for considering for this generalization.

Bonanno [4] gives the example (which he attributes to van Bentham) of an individual sitting in a bar who correctly believes that if he has a drink it will be unsafe to drive. After drinking, however, he becomes more confident and believes it is safe to drive. His mistaken belief is the result of alcoholic confusion and not active deception by an opponent, but the implications for the information function are the same: condition (c) must be relaxed and therefore no partitional representation is possible. Absent-mindedness can provide a reason to relax condition (d). Aumann *et al.* [1] consider a more complex version of the absent-minded driver problem in which the driver has three junctions to contend with rather than the usual two. In that paper it is assumed that the driver cannot tell at all which of the three junctions he is at, so all three are contained in a single information set. But it not implausible to suppose that he might remember passing at least one junction when he is at the third; and when he is at the first, he might be sure that he has not passed as many as two: he knows where he is on the road to within plus or minus one junction. This generates three distinct and overlapping information sets, which can be represented by an information function which satisfies condition (c) but not condition (d). As discussed in section 4.2, this player displays a lack of introspection, i.e. he does not know all of his own beliefs. If he did, assuming he also knows the structure of the game, he could invert the information function

and figure out exactly where he was. These additional examples confirm the importance of the information function approach proposed in this paper.

References

[1] AUMANN, R. J., S. HART, & M. PERRY (1997), "The Absent-Minded Driver", *Games and Economic Behavior* **20**, 102–116.

[2] BENABOU, R. & G. LAROQUE (1992), "Using Privileged Information to Manipulate Markets: Insiders, Gurus, and Credibility", *Quarterly Journal of Economics* **107**, 921–958.

[3] BOARD, O. J. (1998), "Belief Revision and Rationalizability", *Theoretical Aspects of Rationality and Knowledge*, Proceedings of the Seventh Conference, ed. by I. Gilboa.

[4] BONANNO, G. (2002), "Memory of Past Beliefs and Actions", Working Paper, Department of Economics, University of California, Davis.

[5] BRAMS, S. J. (1977), "Deception in 2×2 Games", *Journal of Peace Science* **2**, 171–203.

[6] BRAMS, S. J. (1994), *Theory of Moves*. Cambridge University Press, Cambridge.

[7] CRAWFORD, V. P. (2001), "Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions". Discussion Paper 2001-16, University of California, San Diego.

[8] CRAWFORD, V. P. & J. SOBEL (1982), "Strategic Information Transmission", *Econometrica* **50**, 1431–1451.

[9] DEKEL, E., B. LIPMAN, AND A. RUSTICHINI (1998), "Standard State-Space Models preclude Unawareness", *Econometrica* **66**, 159–173.

[10] FAGIN, R., J. Y. HALPERN, Y. MOSES, AND M. Y. VARDI (1995), *Reasoning About Knowledge*. The MIT Press, Cambridge, MA.

[11] HERODOTUS (1954), *The Histories*, translated by A. de Sélincourt. Penguin Books, Harmondsworth, Middlesex, England.

[12] KEMP, A. (1994), *D-Day and the Invasion of Normandy*. Harry N. Abrams, New York.

[13] OSBORNE, M. J. & A. RUBINSTEIN (1994), *A Course in Game Theory.* The MIT Press. Cambridge, MA.

[14] PICCIONE, M. AND A. RUBINSTEIN (1997), "On the Interpretation of Decision Problems with Imperfect Recall", *Games and Economic Behavior* **20**, 3–24.

[15] RENY, P. J. (1992), "Rationality in Extensive-Form Games", *Journal of Economics Perspectives* **6** 103–118.

[16] SOBEL, J. (1985), "A Theory of Credibility", *Review of Economic Studies* **52**, 557–573.

[17] STALNAKER, R. (1994), "On the Evaluation of Solution Concepts", *Theory and Decision* **37**, 49–73.

[18] VERGIL (1956), *The Aeneid*, translated by W. F. J. Knight. Penguin Books, London, England.