ISSN 1471-0498

DEPARTMENT OF ECONOMICS

DISCUSSION PAPER SERIES

DYNAMIC INTERACTIVE EPISTEMOLOGY

Oliver Board

Number 125

November 2002

Manor Road Building, Oxford OX1 3UQ

Dynamic Interactive Epistemology*

Oliver Board
Department of Economics
Manor Road
Oxford, OX1 3UQ
oliver.board@economics.ox.ac.uk

17 September 2002

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 [39] and Battigalli and Siniscalchi [6], and used to analyze rational play in extensive form games. But while the properties of Kripke 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 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

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

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 [37]). 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 [36], 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 [39] 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 (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.* [20] 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 [33], 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

[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.

each player, called her *type* (after Harsanyi [27]). 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 [18], Heifetz and Samet [28], 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* 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 [29] showed how *Kripke structures* (Kripke [31]) 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]). These axioms are reproduced in Appendix A. 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 [23], reproduced in appendix A 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 [25] *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; $\pi : W \times \Phi \to \{\mathbf{true}, \mathbf{false}\}$ 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. $[\phi]_M = \{w \mid (M, w) \models \phi\}$. 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 $\min_i^w$

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

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 $x \in \min_i^w\{[\phi]_M \cap W_i^w\}$.

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 $\mathcal{M}$ of structures if every formula in $\mathcal{L}$ that is provable in AX is valid with respect to $\mathcal{M}$. The system AX is said to be *complete* for $\mathcal{L}$ with respect to $\mathcal{M}$ if every formula in $\mathcal{L}$ that is valid with respect to $\mathcal{M}$ is provable in AX. We can think of AX as characterizing the class $\mathcal{M}$ 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 $\mathcal{M} \models \phi$. 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 B.

[4] Well-foundedness of $\preccurlyeq_i^w$ is stronger than the limit assumption assumption proposed by Lewis [32]: 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).

[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 $B_i\phi$.

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.

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 [33] 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 recovers 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

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

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

[6] A similar restriction is imposed by Friedman and Halpern [21] on their logic of belief change. See section 5 for a more detailed discussion of this work.

Proposition 1 *$(M^c, w_S) \vDash \phi$ 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 B. 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 $\vDash$ 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 [18], Heifetz and Samet [28], 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 $B_i^\phi\psi \Rightarrow B_i^\chi B_i^\phi\psi$
TNI $\neg B_i^\phi\psi \Rightarrow B_i^\chi \neg B_i^\phi\psi$

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 [24], 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 [35]. 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 [21] 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 [25] 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 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 $BRSIC$:

[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.

[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].

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 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:

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

$(M, w) \vDash B_i^\phi\psi$ iff $(M, x) \vDash \psi$ 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 [39]. 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 $Eb^0\phi$ 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 [39] 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 [34]).

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]:

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

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

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

(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′ $\quad K_i\phi \Rightarrow \Big(\phi \wedge \Big(\psi \Rightarrow B_i^{\psi}\phi\Big)\Big)$

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 express 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 [38].

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 [32], 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 different from the axioms of belief revision.

[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.

Stalnaker [39] 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 [21] 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 [22], 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 [33] 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 [39], and to shed light on the forward and backward induction procedures [40]. 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] Fagin, R. (1994), "A Quantitative Analysis of Modal Logic", *Journal of Symbolic Logic* **59**, 209–252.

[19] Fagin, R., J. Y. Halpern, Y. Moses, and M. Y. Vardi (1995), *Reasoning About Knowledge*. The MIT Press, Cambridge, MA.

[20] 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.

[21] 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.

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

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

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

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

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

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

[28] Heifetz, A. and D. Samet (1998), "Knowledge spaces with arbitrarily high rank", *Games and Economic Behavior* **22**, 260–273.

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

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

[31] 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.

[32] Lewis, D. (1973), *Counterfactuals*. Basil Blackwell, Oxford.

[33] Mertens, J.-F. and S. Zamir (1985), "Formulation of Bayesian Analysis for Games of Incomplete Information", *International Journal of Game Theory* **14**, 1–29

[34] Pappas, G. and M. Swain (eds.) (1978), *Essays on Knowledge and Justification*, Cornell University Press, Ithaca, NY.

[35] Piccione, M. and A. Rubinstein, "On the Interpretation of Decision Problems with Imperfect Recall", *Games and Economic Behavior* **20**, 3–24.

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

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

[38] 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.

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

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

A The AGM Axioms for Belief Revision

Appendix A gives a short description of the AGM belief revision theory. A more complete account is provided by Gärdenfors [23].

The AGM theory (so-called after Alchourrón, Gärdenfors, and Makinson [1]) provides a set of axioms which, they argue, any reasonable belief revision system should satisfy. The axioms do not determine a unique belief revision function. 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"). It is assumed that belief sets are closed under logical consequence.

The AGM axioms impose restrictions on the form K^*_ϕ, the agent's revised belief set after information ϕ is learned. Although we normally assume that belief sets are logically consistent, it is convenient to define the *absurd* belief set K_{false} in which everything is believed (i.e. K_{false} contains every formula of propositional calculus). Finally, K^+_ϕ is used to denote the *expansion of* K *by* ϕ, formed by adding ϕ to K and closing under logical consequence. We can now state the axioms:

(K*1) K^*_ϕ is a belief set;

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

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

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

(K*5) $\mathrm{K}^*_\phi = \mathrm{K}_{false}$ if and only if ϕ is logically inconsistent.

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

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

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

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

IE(a): we must show that $\mathcal{M} \vDash B_i^{\phi}\psi \Rightarrow \left(B_i^{\phi\wedge\psi}\chi \Leftrightarrow B_i^{\phi}\chi\right)$. Suppose $(M,w) \vDash B_i^{\phi}\psi$; then $(M,x) \vDash \psi$ 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 $\vDash$, $[\phi]_M \cap [\psi]_M = [\phi \wedge \psi]_M$; 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 $(M,w) \vDash B_i^{\phi\wedge\psi}\chi$ iff $(M,w) \vDash B_i^{\phi}\chi$, as required.

LE: we must show that if $M \vDash \phi \Leftrightarrow \psi$, then $M \vDash B_i^{\phi}\chi \Leftrightarrow B_i^{\psi}\chi$. Suppose that $M \vDash \phi \Leftrightarrow \psi$. Then by the definition of $\vDash$, $[\phi]_M = [\psi]_M$, and so $\min_i^w \{[\phi]_M \cap W_i^w\} = \min_i^w \{[\psi]_M \cap W_i^w\}$. It follows immediately that $(M,w) \vDash B_i^{\phi}\chi$ iff $(M,w) \vDash B_i^{\psi}\chi$, 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[11] 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:

[11] See e.g. Fagin *et al.* [19] pp. 52, 53.

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

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

$(**)$ $(M_\phi, w_S) \vDash \psi$ 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 follows that $(M_\phi, w_S) \vDash \phi$, and so ϕ is satisfiable with respect to $\mathcal{M}$ as required.

For the construction of M_ϕ 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. We now define M_ϕ. Let $M_\phi = \langle W, \pi, \preccurlyeq\rangle$, where

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

$\pi(w_S)(\psi) = \begin{cases} \textbf{true} \text{ if } \psi \in S \\ \textbf{false} \text{ if } \psi \notin S \end{cases}$, 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 $(M_\phi, w_S) \vDash B_i^\chi\zeta$.

So suppose 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. $BRS \vdash \neg B_i^{\xi}\neg\chi \Rightarrow \left(B_i^{\xi\wedge\chi}\eta \Leftrightarrow \left(B_i^{\xi}\eta \vee B_i^{\xi}\left(\chi \Rightarrow \eta\right)\right)\right)$ — **IE(b)**
7. $BRS \vdash \neg B_i^{\xi}\neg\chi \Rightarrow \left(B_i^{\chi\wedge\xi}\eta \Leftrightarrow \left(B_i^{\xi}\eta \vee B_i^{\xi}\left(\chi \Rightarrow \eta\right)\right)\right)$ — 6, **LE**, **Taut**, **MP**
8. $BRS \vdash (\chi \Rightarrow \eta) \Rightarrow \neg\chi$ — 1, **Taut**, **MP**
9. $BRS \vdash B_i^{\xi}\left(\chi \Rightarrow \eta\right) \Rightarrow B_i^{\xi}\neg\chi$ — 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. $BRS \vdash \neg B_i^{\chi}\neg\rho \Rightarrow \left(B_i^{\chi\wedge\rho}\zeta \Leftrightarrow \left(B_i^{\chi}\zeta \vee B_i^{\chi}\left(\rho \Rightarrow \zeta\right)\right)\right)$ — **IE(b)**
13. $BRS \vdash \left(\neg B_i^{\chi}\neg\rho \wedge B_i^{\chi}\zeta\right) \Rightarrow B_i^{\chi\wedge\rho}\zeta$ — 15, **Taut**, **MP**
14. $BRS \vdash \left(\neg B_i^{\chi}\neg\rho \wedge B_i^{\chi}\zeta\right) \Rightarrow B_i^{\rho\wedge\chi}\zeta$ — 16, **LE**, **Taut**, **MP**
15. $BRS \vdash \neg B_i^{\rho}\neg\chi \Rightarrow \left(B_i^{\rho\wedge\chi}\zeta \Leftrightarrow \left(B_i^{\rho}\zeta \vee B_i^{\rho}\left(\chi \Rightarrow \zeta\right)\right)\right)$ — **IE(b)**
16. $BRS \vdash \left(\neg B_i^{\chi}\neg\rho \wedge B_i^{\chi}\zeta \wedge \neg B_i^{\rho}\neg\chi\right) \Rightarrow \left(B_i^{\rho}\zeta \vee B_i^{\rho}\left(\chi \Rightarrow \zeta\right)\right)$ — 17, 18, **Taut**, **MP**

Since $\chi, \zeta, \rho \in Sub^{+}(\phi)$, $B_i^{\rho}\zeta$ and $B_i^{\rho}\left(\chi \Rightarrow \zeta\right)$ are both in $Sub^{++}(\phi)$. So either $B_i^{\rho}\zeta \in S$ or $B_i^{\rho}\left(\chi \Rightarrow \zeta\right) \in S$: if $\neg B_i^{\rho}\zeta \in S$ and $\neg B_i^{\rho}\left(\chi \Rightarrow \zeta\right) \in S$, 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 $(M_\phi, w_T) \vDash \zeta$ by the hypothesis of induction. It follows from the definition of $\vDash$ that $(M_\phi, w_S) \vDash B_i^{\chi}\zeta$.

For the "only if" direction, assume $(M_\phi, w_S) \vDash B_i^{\chi}\zeta$. 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 $(M_\phi, w_T) \vDash \neg\zeta$, since $\neg\zeta \in T$, and it follows immediately from the definition of $\vDash$ that $(M_\phi, w_S) \vDash \neg B_i^{\chi}\zeta$, 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.* [19], 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. $BRS \vdash B_i^{(\psi\vee\chi)\wedge\chi}\zeta \Leftrightarrow B_i^\chi\zeta$ 18, **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 $\preccurlyeq_i^{w_S}$ 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 $|Sub(\phi)|^2$ 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**); $B^{\zeta}_i\psi \in S^{++}_{neg}$ if and only if $B^{\zeta}_i\chi \in S^{++}_{neg}$ (**RE** and **Dist**); and $B^{\psi}_i\zeta \in S^{++}_{neg}$ if and only if $B^{\chi}_i\zeta \in S^{++}_{neg}$. 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. $\mathcal{A} \vDash B^{\phi}_i\psi \Rightarrow B^{\chi}_i B^{\phi}_i\psi$ and $\mathcal{A} \vDash \neg B^{\phi}_i\psi \Rightarrow B^{\chi}_i\neg B^{\phi}_i\psi$. (**TPI**) Suppose that $M \in \mathcal{A}$, and $(M,w) \vDash B^{\phi}_i\psi$. Then for every $x \in \min^{w}_{i}\{[\phi]_M \cap W^{w}_{i}\}$ we have $(M,x) \vDash \psi$. Now for every $y \in \min^{w}_{i}\{[\chi]_M \cap W^{w}_{i}\}$, $z \preccurlyeq^{y}_{i} u$ if and only if $z \preccurlyeq^{w}_{i} u$, so $\min^{w}_{i}\{[\phi]_M \cap W^{w}_{i}\} = \min^{y}_{i}\{[\phi]_M \cap W^{y}_{i}\}$. It follows that $(M,y) \vDash B^{\phi}_i\psi$, and therefore $(M,w) \vDash B^{\chi}_i B^{\phi}_i\psi$. (**TNI**) Suppose that $M \in \mathcal{A}$, and $(M,w) \vDash \neg B^{\phi}_i\psi$. Then there is some $x \in \min^{w}_{i}\{[\phi]_M \cap W^{w}_{i}\}$ such that $(M,x) \nvDash \psi$. Now for every $y \in \min^{w}_{i}\{[\chi]_M \cap W^{w}_{i}\}$, $z \preccurlyeq^{y}_{i} u$ if and only if $z \preccurlyeq^{w}_{i} u$, so $\min^{w}_{i}\{[\phi]_M \cap W^{w}_{i}\} = \min^{y}_{i}\{[\phi]_M \cap W^{y}_{i}\}$. It follows that $(M,y) \vDash \neg B^{\phi}_i\psi$, and therefore $(M,w) \vDash B^{\chi}_i\neg B^{\phi}_i\psi$.

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: $\psi \in Sub^{++}(\phi)$ is the smallest set of formulas such that (a) if $\psi, \chi \in Sub^{+}(\phi)$ then $\psi, B_i\psi, B^{\chi}_i\psi \in Sub^{++}(\phi)$; (b) if $\xi \in Sub^{+}(\phi)$ and $B^{\chi}_i\psi \in Sub^{++}(\phi)$, then $B^{\xi}_i B^{\chi}_i\psi \in Sub^{++}(\phi)$ and $B^{\xi}_i\neg B^{\chi}_i\psi \in Sub^{++}(\phi)$. The proof that $(M_\phi, w_S) \vDash \psi$ if and only if $\psi \in S$ is unaffected, as is the proof that $\preccurlyeq^{w_S}_{i}$ is complete and transitive on $W^{w_S}_{i}$ for all S. Finiteness of W still holds as well, since for every formula of the form $B^{\zeta}_i B^{\xi}_i \ldots B^{\chi}_i\psi \in S$ if and only if $B^{\xi}_i \ldots B^{\chi}_i\psi \in S$, given **TPI** and **TNI**. It remains to show that M is nested.

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

Theorem 3

Soundness

We must check that **WCon** is valid with respect to $\mathcal{I}$, i.e. $\mathcal{I} \vDash \phi \Rightarrow \neg B_i^{\phi} false$. Suppose that $M \in \mathcal{I}$, and $(M, w) \vDash \phi$, 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 $(M, x) \vDash false$, it is not the case that $(M, w) \vDash B_i^{\phi} false$, and it follows from the definition of $\vDash$ that $(M, w) \vDash \neg B_i^{\phi} false$.

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 $x \in \min_{\trianglelefteq_i}(X \cap W_i^w)$. 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 $x \in \min_{\trianglelefteq_i}(X \cap W_i^w)$.■

Theorem 5

The proof of Theorem 5 follows the same steps as the proof of Theorem 3.3.1 in Fagin *et al.* [19], and is omitted.