ELSEVIER

Mathematical Social Sciences 36 (1998) 31–56

mathematical social sciences

Belief revision in games: forward and backward induction[1]

Robert Stalnaker*

Department of Linguistics and Philosophy, E39-245, MIT, Cambridge, MA 02140, USA

Abstract

The rationality of choices in a game depend not only on what players believe, but also on their policies for revising their beliefs in response to surprising information. A general descriptive framework for representing belief revision policies in game situations is sketched, and the consequences of some assumptions about such policies are explored. Assumptions about epistemic independence and a rationalization principle are considered. It is argued that while such assumptions may be appropriate in some contexts, no substantive constraints on belief revision policies can be justified on the basis of the assumption of common knowledge of rationality.

1. Introduction

In any strategic situation, my decision about what to do will normally depend on my expectations about how you would respond to what I might do – on how I think your response to what I *will* do compares with what your response *would* be to other things that I *could* do. Some of the things I could do would probably surprise you, and so in considering the consequences of my possible actions, I will have to consider how you would revise your beliefs in the face of surprising information – information that conflicts with your probability-one beliefs. Even if I think you *know* what I am going to do, I can consider how I think you would react if I did something that you and I both know I will not do, and my answers to such counterfactual questions will be relevant to assessing the rationality of what I *am* going to do.

*Corresponding author. Tel.: +1 617 253 0472; fax: +1 617 253 5017; e-mail: stal@mit.edu

[1]Thanks to the participants in the LOFT2 meeting for helpful discussion, and especially to Pierpaolo Battigalli, who pointed out the need for a qualification in the theorem stated in Section 5. I also want to thank Giacomo Bonanno, Battigalli again, and a referee for very useful written comments and stimulating correspondence on an earlier draft of the paper.

0165-4896/98/$19.00
PII: S0165-4896(98)00007-9

Should game theory tell us how to revise our beliefs? It seems reasonable to suppose that rational players will revise their beliefs by conditionalization in the case where the new information is compatible with the prior information, but what can be said about how rational players should respond to surprising information? Very little, I will argue. That is, assumptions about rationality, and about common belief in rationality, put no substantive constraints on how an agent does or should revise beliefs in response to surprising information. We can, however, say a great deal about the consequences for action of various assumptions about belief revision policies, and of assumptions about agents' beliefs and common beliefs about the belief revision policies of others. Getting a clear and precise descriptive account of the alternative belief revision policies that a rational person might adopt can help to explain the strengths and weaknesses of some patterns of strategic reasoning, reasoning of players in a strategic situation, as well as the reasoning of theorists about how players reason.

My main aim in this paper is to try to separate out assumptions about belief revision from assumptions of other kinds about a strategic situation: from assumptions about the causal structure of the game, and from assumptions about the rationality of the participants of the game. Substantive belief revision policies, I will argue, are sometimes given a fallacious defense when they are confused with other things.

My plan is this: In Section 2 I will describe the formal tools – the kinds of models that I will use to represent issues concerning belief revision in games, and make some general remarks about some of the ideas motivating the framework. In Section 3 I will consider, in isolation, one component of the models, the belief revision component, defining some descriptive concepts that this theory makes available to represent the policies that players of games may adopt. Then, before deploying some of these concepts in the context of the game models, I will make some comments, in Section 4, on the relation between static and dynamic representations of games, arguing that the kind of epistemic model we have defined for a strategic form game has the resources to explain the sequential interactions in a dynamic game. After this brief digression, I will make precise some alternative assumptions that might be made about the belief revision policies in the game context. Section 5 concerns assumptions about epistemic independence that support backward induction reasoning, while Section 6 considers a rationalization principle that supports some forward induction arguments. I will state a couple of results about the consequences of such policies for what players will do, and I will also look critically at what I will argue are some illegitimate defenses of such policies.

2. Game models

A model for a game, as I will understand it, is a representation of one particular playing of a game, where a game is defined in the standard way. We will confine ourselves to static strategic form representations of games, although the phenomena that are being represented are dynamic interactive situations. The intention is to represent in

the models all features of the situation that might be relevant to decision-making and strategic reasoning in the course of the game, and to the evaluation of the decisions and reasoning of the participants. The assumption underlying the strategic form representation is that all the relevant information will be reflected in the initial dispositions of the players – their conditional decisions about what they would do if certain choice situations were to arise, and their dispositions to change their beliefs in response to the various potential evidence that they might receive. All such models are, of course, highly idealized and artificial, but the aim is to get a precise representation, and to make the idealizations and simplifications explicit; the hope is that the representation will be detailed and realistic enough to be able to represent and clarify some patterns of reasoning that occur, or should occur, in actual situations.

The assumptions of our models are thoroughly Bayesian: to behave rationally in an interactive game, as in a simple decision situation that involves just one agent, is to maximize expected utility, where utility is a measure of the agent's desires, values or priorities, and the probabilities used to define the expectation are representations of the agent's degrees of belief. The definition of the game provides the utilities that motivate the players; the models of the game will add representations of the beliefs and degrees of belief of the different players, and representations of what decisions the players make – what strategies they choose. Our models will also contain a representation of belief revision policies – the ways that each agent would be disposed to revise her beliefs in response to any surprising information that might be received in the course of the play of the game.

The game itself is represented in the standard way: $\Gamma = \langle N, \langle C_i, u_i \rangle_{i \in N} \rangle$, where N is a finite set of players, C_i is a set of strategy options for player i, and u_i is a utility function for player i defined on the strategy profiles. The model for the game represents just one particular playing of it, but to represent all the relevant facts about what actually happens – the capacities and epistemic states of the participants – we need a representation of a range of alternative possible states of the world – alternative complete descriptions of different ways that the relevant facts might have been. The points of this state space W are primitive elements of the model. The properties of the situations they represent are given by various relations and functions defined on the space. So for example, the strategy that player i plays in world x will be given by a function S_i from worlds to strategies: For $x \in W$, $S_i(x) \in C_i$. And for each possible world, or point in the state space, and for each player, there will be a set of possible worlds representing the possibilities that are compatible with what that player believes in that world. We will assume that players have introspective access to their own beliefs, but we allow that players may have false beliefs, which means that the world in which a player has certain beliefs need not itself be compatible with those beliefs. The qualitative structure of a player's belief states in various possible worlds can be represented by a binary relation, R_i: xR_iy says that world y is compatible with what i believes in x. The R relations will be serial, transitive and euclidean, but not necessarily reflexive.[2] (These constraints encode the

[2] A relation R is serial iff $(x)(\exists y)xRy$; it is euclidean iff $(x)(y)(z)((xRy \text{ and } xRz) \rightarrow yRz)$.

assumptions of introspective access to one's own beliefs.) It is also assumed that for all x and y, if xR_iy, then $S_i(x) = S_i(y)$. Intuitively, this is the assumption that each player knows what strategy he himself chooses. For each world x and player i, a probability measure $P_{i,x}$ on $\{y{:}xR_iy\}$ will represent player i's partial beliefs in x.

Since players may have some false beliefs, the structure determined by the R relations will differ from the more familiar partition structures, but there is an underlying partition structure determined by another relation definable in terms of R. We can distinguish the possible worlds compatible with what i beliefs in x from the set of possible worlds in which i has exactly those beliefs. Say that worlds x and y are *subjectively indistinguishable* to player i ($x \approx_i y$) iff$(z)(xR_iz \leftrightarrow yR_iz)$. The *type* of a player (in one sense of the term) could be defined by the equivalence classes determined by this relation, and type spaces of the kind familiar to game theorists can be defined in terms of our models.

Since players can be wrong about what they believe, they might, at some later time, be surprised by a discovery of information previously incompatible with their initial beliefs. Our models will represent policies for responding to such surprises – epistemic priorities that determine how a player is disposed to revise beliefs and partial beliefs in response to unexpected information. It is this feature of the model that will be our main concern in this paper, and we will discuss the motivation for the representation in Section 3. Belief revision policies are represented in the model by another binary relation which provides, for each player, a complete weak ordering of all the possible worlds within each subjective indistinguishability equivalence class.

The following definition of a model is designed to provide an economical representations of all this information:

A model for Γ is a structure $\langle W,\mathbf{a},\langle S_i,Q_i,P_i\rangle_{i\in N}\rangle$, where W is a nonempty finite set (the state space, or set of possible worlds); $\mathbf{a} \in W$ (representing the actual state); each S_i is a function from W to C_i (determining player i's strategy choice in each state); each Q_i is a binary relation that is reflexive, transitive and weakly connected.[3] (the Q's define a structure that represents qualitative information about each player's beliefs and belief revision policies, in a way to be discussed below); each P_i is an additive measure function on the subsets of W such that $P_i(\{x\}) > 0$ for all $x \in W$. (These measure functions, together with the Q relations, determine probability functions, $P_{i,x}$, representing prior and posterior beliefs for each player i and each possible world x.)[4]

The R relations defining prior beliefs and the subjective indistinguishability relations are definable in terms of the Q's as follows: First, $x \approx_i y$ iff either xQ_iy or yQ_ix. Second, xR_iy iff for all z such that $z \approx_i x$, zQ_iy. That is, the worlds compatible with what a player initially believes are those that have highest epistemic priority, as determined by the Q relation. The partial belief functions, $P_{i,x}$ for each world x are defined as follows:

$$P_{i,x}(\phi) = P_i(\phi \cap y{:}xR_iy)/P_i(y{:}xR_iy)$$

[3] A binary relation R is weakly connected iff for any x, y and z, if xRy and xRz, then either yRz or zRy.
[4] See [15] for a more thorough account of this model theory.

[5]The particular terminology and formulation that I have used to define game models may not be exactly the same as that in many standard representations of interactive situations, but the general ideas will be familiar. Since types can be identified with subjective indistinguishability equivalence classes, these models will generate type spaces of a familiar kind. And the qualitative belief revision structure, together with the measure functions on the various prior and posterior belief sets delivered by that structure will determine standard lexicographic probability systems. In particular, it can be shown that the notion of type in the epistemic model presented above corresponds closely to the notion of type in Ben Porath's epistemic models for extensive games.[6] The aim in giving the general definition of a model is not to propose an original explanatory hypothesis, or any explanatory hypothesis, for the behavior of players in games, but only to provide a descriptive framework for the representation of considerations that are relevant to such explanations, a framework that is as *general* and as *neutral* as we can make it. While we want to consider the consequences of assumptions about rationality, and various assumptions about what players believe about each others' rationality, and the way they are disposed to revise those beliefs, the definition of models builds in no assumptions about what anyone does, or what anyone believes about what anyone does. We want to be able to clarify the content of the assumptions by distinguishing models in which they hold from those in which they do not.

Along with generality and neutrality, two other methodological virtues we are aiming at in this inquiry into what has come to be called interactive epistemology are *modularity* and *compositionality*. Let me say a little about them to motivate the discussion of belief revision.

---

[5]There is one further general constraint: A structure of this kind is a model only if the state space W satisfies a closure condition that is designed to ensure that the causal structure of the game is reflected in the models. Intuitively, the requirement is that there be enough points in the state space to represent the capacities of the players to choose whatever the definition of the game says they have the capacity to choose, and to represent the causal consequences of those hypothetical choices. Here is the condition:

For any $w \in W$, player i and $s \in C_i$, there is a point $f(w,s) \in W$ meeting the following four conditions:

1. for all $j \neq i$, if $w \approx_j x$, then $f(w,s) \approx_j x$.
2. if $wQ_i x$, then $f(w,s) Q_i f(x,s)$.
3. $S_i(f(w,s)) = s$
4. $P_i(f(w,s)) = P_i(w)$.

If $s \in C_i$ is a strategy different from $S_i(x)$, then intuitively, $f(w,s)$ will represent a world that is exactly like w, except that player i chooses strategy s instead of the strategy he chooses in w. Since strategies are chosen independently, no player could influence the actions or beliefs of another player by making an alternative choice, and this is reflected in the first condition, which says that other players' beliefs are the same in $f(w,s)$ as they are in w. The second and fourth conditions encode the intuitive assumption that if a player had made a different choice, he would have had the same passive beliefs – the same beliefs about the prior beliefs and strategy choices of other players – as those he has in given situation. Just as a player cannot affect another player's prior state, so he can also not affect his own prior beliefs about that state, by what he chooses to do. (This causal independence condition is a slight strengthening of the condition stated in [15].)

[6]See [5].

Everyone will agree that the dynamic interactive decision situations that game theory seeks to model can be enormously complicated. A single decision by a single agent is complicated enough, with beliefs and degrees of belief about the state of the world and about feasible actions, and with priorities and valuative measures on the consequences that the alternative actions might have. But if we have an interactive sequence of actions involving different agents, with beliefs and degrees of belief, including beliefs about the beliefs, degrees of belief and values of others, changing over time, the complications proliferate. Many different factors go into determining how beliefs change, actually and hypothetically, in response to actual and hypothetical actions, and the way one player may predict the actual and hypothetical belief changes of another. The best hope for understanding such situations is to separate out the components that interact in the interactive situation – the different modules that are relevant to determining and explaining action – and to clarify their conceptual structure in isolation, with the hope that the complexity can be explained as something that arises from the interaction of structures that are relatively simple in themselves. The idea is to think of the game situation as a joint application of a number of different interacting theories, theories that can be independently motivated in contexts that are more general than the application that brings them together. Part of the motivation of the Bayesian approach, I think, is that it treats the game situation as one application of individual decision theory. Rationality is defined in the more general context in which an agent's actions are evaluated in terms of beliefs and degrees of belief about states of the world, abstractly characterized. A more specific situation is one in which the relevant states for one agent are defined by alternative patterns of actions by other agents. Explaining behavior in this kind of situation should require, not a new theory, but an application of the general theory to a specific situation.

If we have clear accounts, in the contexts of our models, of the simple concepts, then complex concepts, such as common belief in, or knowledge of, rationality, or common belief or knowledge that players will revise their belief about each other in certain specified ways can be defined straightforwardly as compositions of their simple parts. Rationality is just rationality, as defined in individual decision theory; belief is just belief as defined in an abstract semantic theory of belief. Common belief is a certain complex composition of beliefs, defined in terms of belief in a way that is exactly analogous to the way common knowledge is defined in terms of knowledge. Beliefs about rationality are simply beliefs with a certain subject matter, and the characterization of the subject matter, or the content, that a belief may have should be independent of the characterization of the attitude – belief – that can have such a content.

3. The abstract belief revision theory

With the desiderata of modularity and compositionality in mind, I want to consider the general features of rational belief revision theory in abstraction from any particular application – in abstraction from any assumptions about the subject matter of the prior and posterior beliefs. The theory we import is the standard one in the belief revision

literature – the so called AGM theory.[7] The account is purely qualitative, simple enough so that its structure is transparent and its features can be motivated independently of the game theoretic application. But it is also a theory that combines easily and naturally with the other components of the overall theory, and that connects with developments within game theory. The qualitative structure can be studied in itself, but it combines naturally with a measure function that yields structures that are relatively familiar in the game theoretic context: lexicographic and conditional probability systems – conditional probabilities that are defined when the condition has probability zero.

The abstract AGM theory is a theory of a single agent, and says nothing about the subject matter of the beliefs of that agent. But it is a straightforward matter to combine representations of different agents in a single model, and to take the subject matter of the beliefs of those agents to include facts about the agents themselves, what they believe about each other, how they revise their beliefs, and what they do in a game situation. But before getting to the application in which the theory is incorporated into the game models, let me sketch the abstract structure by itself.

A state of belief for an agent is represented by a set of possible worlds (a state space) – the worlds compatible with the agent's beliefs (which to keep things simple, we assume to be finite). We need both a set B, representing the prior beliefs of the agent, and a superset of B, B', that contains all of the possible worlds compatible with any surprising information that the agent could conceivably receive – all the worlds that are compatible with some potential posterior belief state. A belief revision policy is represented by a function taking any potential piece of new information ϕ, represented by any subset of B', into a subset of B' $B(\phi)$ that represents the posterior belief state induced by the discovery of the information ϕ. The following are the constraints on this function:

1. $B(\phi) \subseteq \phi$
2. If $B \cap \phi$ is nonempty, then $B(\phi) = B \cap \phi$.
3. If ϕ is nonempty, then $B(\phi)$ is nonempty.
4. If $B(\phi) \cap \psi$ is nonempty, then $B(\phi \cap \psi) = B(\phi) \cap \psi$

Condition (1) is simply the requirement that in the posterior belief state induced by the discovery that ϕ, the agent will believe that ϕ. Condition (2) is a conservative condition: if the information received is compatible with prior beliefs, then no prior beliefs should be given up. Condition (3) is the requirement that for every piece of information consistent with B', there is a consistent posterior state induced by it. Condition (4) is a generalization of the conservative condition: if one piece of information is compatible with the posterior state induced by another, then the posterior state induced by their conjunction should be obtainable by simply adding the second piece of information to the state induced by the first.

Any belief revision function meeting these conditions can be represented by a reflexive, transitive, and connected relation on the set B' – a relation that establishes an

[7] See [11] for a survey of the theory. See also [12].

epistemic priority ordering of the possible worlds. Given any function defined for all subsets of B' meeting the conditions stated above, we can define the relation as follows: xQy iff $y \in B(\{x,y\})$. It can be shown, using the four conditions, that Q is transitive, connected and reflexive, and that

(*) for any ϕ, $B(\phi) = \{y \in \phi$: for all $x \in \phi$, $xQy\}$.

That is, $B(\phi)$ is the subset of ϕ that has highest priority with respect to Q. B – the set of worlds defining the prior beliefs – will be the set of worlds to which all worlds are Q related.

Alternatively, if we begin with any weak total ordering relation on B', we can use (*) to define the belief revision function $B(\phi)$, and show that it satisfies the four conditions. So these are just two alternative formulations of the same theory.

Now we can use the resources of this simple abstract theory to define a number of descriptive concepts that will be useful for characterizing the structure of an agent's beliefs and belief revision policies. In a representation of a simple belief state, all full beliefs have the same epistemic status, but with the belief revision structure we can make further distinctions between them. For example, some full beliefs are more robust than others – they may survive belief revisions that the others will not survive. More precisely, we may say proposition ϕ is believed *robustly with respect to* ψ if ϕ is believed (that is, $B \subseteq \phi$), and would still be believed if ψ were learned (that is, $B(\psi) \subseteq \phi$). A belief ϕ is *robust with respect to the truth* if for any actually true proposition ψ, ϕ is believed robustly with respect to ψ.[8] One may say that a proposition is an *absolutely robust* belief if for any proposition ψ that is compatible with ϕ, ϕ is believed robustly with respect to ψ. And we may say that two beliefs are *epistemically independent* if each is believed robustly with respect to the negation of the other. (That is ϕ and ψ are epistemically independent beliefs iff $B \subseteq \phi \cap \psi$, $B(\sim\phi) \subseteq \psi$, and $B(\sim\psi) \subseteq \phi$.) Intuitively, beliefs are epistemically independent if learning that one is false will not affect one's belief in the other. When probabilities are added to the structure, we can see this notion of epistemic independence as a special case of probabilistic independence. More generally, we can say that ϕ and ψ (whether believed or not) are epistemically independent iff $P(\phi/\psi) = P(\phi/\sim\psi)$, and $P(\psi/\phi) = P(\psi/\sim\phi)$.

Within the abstract theory, these descriptive concepts give us no further constraints on belief revision – no way even to state, much less to defend, any methodological requirements to be added to the four formal constraints on belief revision given above. In a theory where propositions are just sets of undifferentiated points of a given state space, there is no way to generalize about what kinds or categories of propositions might be more robust than others, or about what ought to be epistemically independent of what. But when we move to the game theoretic application, where the possible worlds that define the space of players' beliefs have a rich structure, we will have the resources to distinguish between different kinds of propositions, and so to state substantive

[8] In [15] a definition of *knowledge* as belief that is robust with respect to the truth is discussed.

generalizations about belief revision policies, and to consider the consequences of adopting such policies, and of believing that they are adopted by others.

Now to combine the belief revision structure with the other elements of a model of a game, all we need is an epistemic priority relation for each player, and for each possible world: this is what is provided by the Q_i relations in our definition of a model. The set B' of possible worlds that are compatible with what a player could conceivably learn is represented by the set of worlds in which the player has the same prior epistemic state – the equivalence class of worlds that are epistemically indistinguishable from each other. These qualitative epistemic relations deliver the set of possible worlds that are compatible with all of the different possible prior and posterior belief states that the different players may be in. Just as in the simple abstract theory, the prior belief state B and the posterior states $B(\phi)$ were definable in terms of Q, so in the game models, prior and posterior sets $B_{i,x}$ and $B_{i,x}(\phi)$ can be defined for each player and possible world in terms of Q_i. The general measure functions, P_i, provide the information about the degrees of belief for each of those states, yielding, when combined with the Q's, a lexicographic probability system for each player in each possible world.

Before looking at some alternative substantive belief revision policies, we should note that without assuming anything beyond the formal requirements about how players revise their beliefs, the belief revision structure provides the resources to define a strengthened definition of rationality that is appropriate for the strategic form representations of the interactive dynamic situations that are the subject matter of our models. If a player must make conditional decisions about what to do in situations that he believes, with probability one, will not arise, then he should base such a decision on what he *would* believe – on the way he would revise his beliefs – should he be surprised by the information that he is in that situation. This consideration motivates a refinement of expected utility maximization – lexicographic utility maximization, or what I have called ''perfect rationality'': roughly, the idea is that in case of ties in expected utility, one should maximize conditional expected utility on the hypothesis that one will be surprised.[9]

The strengthened definition of rationality is crucial to the relevance of belief revision to decision making, but it is important to note that the assumption that players are perfectly rational – that they are lexicographic expected utility maximizers – by itself assumes nothing about how beliefs are revised, or about what anyone believes about how they are revised. To say that a player is perfectly rational just means that however she is disposed to revise her beliefs, she will take account of this, when appropriate, in deciding what to do.

4. Static models of dynamic games

Strategic reasoning concerns the interaction of actions and beliefs in the course of the playing of a dynamic game, but our models are static models of strategic or normal form

[9]This is a rough approximation of lexicographic utility maximization, or perfect rationality. See [15] for the definition in the context of our models. See also [9] and [8].

games. They represent a single moment at which simultaneous choices of strategies are made by all the players. How can this kind of model throw light on dynamic strategic reasoning? How can a theory of belief revision be relevant to a static situation in which, it seems, no actual belief revision has time to take place? Before going on to apply the descriptive resources of the belief revision theory to the game situation, we need to make some brief remarks about the relation between dynamic games and static representations of them.

The original idea of the early developers of game theory, as I understand it, was that all the strategically relevant features of the dynamic interaction could be represented in the dispositions that the players had at the beginning of the game. At least if players are fully rational, then what they will decide to do when a certain situation arises can be assumed to be the same as what they now will decide that they *would* do if that situation *were* to arise. The argument was that a conditional choice (to do a if I learn ϕ) is rational if and only if doing a would be rational, if I were to learn ϕ. There was one problem with this assumption that emerged and plagued the subsequent discussions: what if I believe, with probability one, that the condition will not be realized? Then if rationality is simply maximizing expected utility, and if ϕ has zero probability, then it may be that it *is* rational to choose (a if ϕ) even though it *would* be irrational, were you to learn ϕ, to choose a. Some were moved by this kind of problem to turn away from the strategic form representation, but others responded by trying to give a richer representation of the initial states of the players – to develop the strategic form representation by adding structure that reflects the players' dispositions to respond to information that conflicts with their prior beliefs. Rationality is refined so that, in some cases, probability zero possibilities must be taken into account in evaluating the rationality of an action. This is the point of introducing the belief revision structure, and the notion of perfect rationality, into the models of strategic form games.

The belief revision theory represents a believer's epistemic priorities at a single moment of time, but also represents how the agent will and would respond in a dynamic situation that might follow that moment.

We can interpret a strategy choice, not as an instantaneous commitment, but as a representation of what the player will and would do in the course of the playing of the game. And the static epistemic model can be interpreted as a representation of how the player will and would revise his beliefs should the choices of the other players take the game in various different possible directions. On this interpretation, the way that one chooses a strategy in a game is to make the choices that it dictates when they arise, and to be disposed to make choices that it dictates should (contrary to fact) the opportunity to make them arise. Similarly, to have a given belief revision policy is simply to be disposed to change one's beliefs in the way that it dictates. One might have a belief revision policy even for responding to information that one never imagined one might receive. Even if you are absolutely certain that P is false, you would have to respond if you were confronted with incontrovertible evidence that it is true, and even if you have never considered that possibility, there might be a determinate and rational way that you would respond.

Suppose we have a game in extensive form, and a model for its normal form of the kind that we have defined. Suppose we say that a player is *sequentially rational* if he is

disposed to respond rationally at every possible information set. Because our models represent both conditional choices and conditional beliefs, they contain enough information to define a determinate proposition or event – the set of worlds or states at which player *i* is sequentially rational.[10] It will be clear from the definitions that in any generic game, every sequentially rational strategy is perfectly rational, and that every perfectly rational strategy is equivalent to a sequentially rational strategy. (Where strategies are equivalent if they have the same outcome against every strategy profile for the other players.) The normal form representation cannot distinguish sequentially rational strategies from equivalent strategies that are not sequentially rational, but the differences between equivalent strategies are strategically irrelevant since choice between them can make no difference to actual outcomes, or to beliefs about outcomes, no matter what the choices of the other players.[11]

5. Epistemic independence and backward induction

To generalize about substantive belief revision policies in the game situation we need some precise general categories for propositions – categories that permit us to distinguish between different kinds of information so that we can say, for example, that information of one kind is or is not relevant to revising one's beliefs about propositions of some other kind. I might, for example, take information about one player to be epistemically independent of, or irrelevant to, beliefs about other players. To make such an epistemic policy precise, we need to say what it is for a proposition to be *about* a certain player. The notion of aboutness is notoriously problematic, but in the limited and idealized context of the game models, there is a natural way to distinguish propositions about different players. In this context, the relevant facts are all facts about the subjective states of different players – their beliefs, partial beliefs, belief revision

[10] Here is the rough idea: Let v be any information set for player i. First, define $[v_{-i}]$ as the proposition (event) at which all players other than i choose strategies compatible with v being reached. Second, for any strategy s for player i, define strategy s_v as the strategy that (1) is compatible with v being reached, and (2) is otherwise like s. Now suppose player i actually chooses strategy s in world x. Since strategies chosen represent what a player would do if a certain possibly counterfactual choice situation were to arise, and belief revision policies represent what a player would believe in possibly counterfactual circumstances, the following counterfactuals seem to be true in world x: first, if v were to have been reached, player i would have chosen s_v, and second, i's beliefs then would have been given by $P_{i,x}(--/[v_{-i}])$. This implies that a strategy s for i is rational in world x at (possibly counterfactual) information set v iff for any $s' \in C_i$, the conditional expected utility in x of s_v is at least as great as the conditional expected utility of s'_v, conditioned on $[v_{-i}]$. And since a player is sequentially rational if, for any information set, he would be rational if that information set were reached, we can say that a strategy s for player i is sequentially rational in world x iff it is rational at all of i's information sets.

[11] See [13] for a discussion of the relation between normal and extensive form games. Their project, as I understand it, is to try to extract the relevant structure of the extensive form from the normal form representation. What I am suggesting is something more modest: to extract from an extensive form representation together with an epistemic model of its normal form the relevant structure of an epistemic model for the extensive form. But I have only sketched the rough idea. What is needed is an explicit representation of the dynamic structure and the epistemic states at the different times.

policies, and decisions. It is natural to assume that any two worlds that are subjectively indistinguishable for player i will be worlds that are indistinguishable with respect to information about player i. To be precise, start with the equivalence classes induced by player i's subjective indistinguishability relation $\approx_i$. Any proposition that is a union of these equivalence classes will be a proposition solely about player i. To say it in a slightly different but equivalent way: a proposition ϕ is solely *about player i* iff for any worlds x and y such that $x \approx_i y$, either both $x,y \in \phi$ or neither is.

A proposition ϕ is about a set J of players, and only about the players in the set, if ϕ is equivalent to a conjunction of propositions each of which is solely about one of the players in J. Equivalently, we could define the transitive closure of the equivalence relations $\approx_i$ for each $i \in J$, and define propositions about players in J as unions of the equivalence classes defined by this relation.

Notice that the proposition *it is common belief among players in J that* ϕ (for any set J of players) is a proposition solely about players in J, since this proposition is equivalent to the conjunction of propositions each of which is solely about just one of the players in J. For example, the proposition that it is common belief between Alice and Bob that they both are rational is equivalent to the conjunction of the proposition that *Alice* believes that it is common belief between Alice and Bob that both are rational with the proposition that *Bob* believes that it is common belief between Alice and Bob that both are rational.

So we have given a mathematically precise definition, using the resources of the game models, of the concept ''proposition solely about player i'' and ''proposition solely about the players in set J''. We have also, in the context of the abstract belief revision theory which has been incorporated into the models, given a mathematically precise definition of the general notion of *epistemic independence*: two propositions ϕ and ψ are epistemically independent for player i in world x iff $P_{i,x}(\phi/\psi) = P_{i,x}(\phi/\sim\psi)$, and $P_{i,x}(\psi/\phi) = P_{i,x}(\psi/\sim\phi)$. (For the special case where ϕ and ψ are both believed, with probability one, this is equivalent to the requirement that $B_{i,x}(\sim\psi) \subseteq \phi$, and $B_{i,x}(\sim\phi) \subseteq \psi$.) Putting these two concepts together, we can state a candidate for a substantive belief revision policy that one might adopt:

Information about different players should be epistemically independent. If Bob is following this policy and learns that Alice behaved differently than he predicted, he will not let this information affect his beliefs about what Chloë believes, or will do.[12]

I want to make two points about this kind of belief revision policy, one constructive and one critical. The constructive point is that such policies, or the assumption that it is common belief that players adopt them, tend to support strong backward induction reasoning. I will state a backward induction theorem for perfect information games in agent form in which the belief revision policy plays a crucial role. But the critical point is this: the *epistemic* independence assumptions cannot be defended on the basis of the

[12]This is just one epistemic independence condition. With other categories of propositions, one can state other such conditions. For example, using a coarser equivalence relation one can distinguish the *passive* beliefs of a player – the beliefs of a player that cannot be influenced by that player's choices, and define the propositions that are about a player's passive beliefs. The passive belief state of a player is equivalent to one thing that has been meant by a player's *type*.

assumptions about *causal* independence that are built into the structure of the game. It is a feature of the strategic form game that the strategy choice of each player is made independently of the strategy choice of all the other players; that is, no player can influence the strategy choice of another. But this fact has no consequences for the reasonableness of a policy that takes information about what one player does to be epistemically relevant to conclusions about what another player did or will do. After making this general point, I will look critically at a backward induction argument by Robert Aumann that, I will argue, conflates epistemic and causal independence, implicitly making a strong epistemic independence assumption which it explicitly rejects.

First, the theorem (a sketch of the proof is given in Appendix A):

Theorem. *Let Γ be a perfect information game in agent form in which for each player different outcomes have different payoffs. Let M be a model for Γ in which it is common belief that all agents are perfectly rational, and that all agents adopt belief revision policies that treat information about different agents as epistemically independent. Then in M, the subgame perfect equilibrium strategy profile is realized.*

That is, in any such model for a perfect information game in which each player has only one possible move, and for which different outcomes have different payoffs for each player, the backward induction solution will obtain. What makes the argument work is that the epistemic independence assumption ensures that no one learns anything in the course of the game that is relevant to what will happen later in the game. It is this fact that is required to make it legitimate to work backwards, as backward induction arguments do, from the potential last moves of the game, assuming that if the game were to reach that point, the epistemic situation for the relevant players would be the same, in relevant respects, as it is at the beginning of the game. But without the epistemic independence assumption, there will be no license to assume this, and the argument will fail. And if the game is not an agent form game – if some player may move more than once – then the argument will also fail, unless one makes even stronger epistemic independence assumptions.

There may be a reason, in some situations, to adopt a policy of epistemic independence for different players, but it is important to recognize that nothing about the structure of the game, or about the concept of rationality, requires that rational players should adopt it. It is easy to imagine situations in which following such a policy would be unreasonable – situations in which it would be reasonable to take information about one player's behavior to be relevant to conclusions about the beliefs and behavior of a different player. Suppose Bob is in a game with Alice and Chloë. He believes both are rational, and thinks he knows what each will do. Alice moves first, and surprises him. Why might this reasonably lead him to change his mind about Chloë? Suppose, first, that while Bob has an opinion about what Chloë will do, he believes Alice knows her better than he does. Second, suppose that while he believes that both Alice and Chloë are rational, his belief in Alice's rationality is more robust. Third, suppose that Alice's surprising action is rational only if *she* believes that Chloë will act irrationally. Bob might then infer from Alice's action that Chloë is probably not rational after all. This would be a violation of the epistemic independence condition, but there would be

nothing unreasonable about this inference, even though Bob knows that Alice's and Chloë's strategies were chosen independently.

Or consider a game in which the members of a large population vote independently. Could it ever be reasonable to let one's beliefs about the population be influenced by information about a random sample – an exit poll, for example (whose results were unavailable to the voters not polled)? Of course it would be reasonable, but this is incompatible with the epistemic independence condition.

Epistemic and causal independence have often been conflated in discussions of games, even though the distinction is familiar in discussions of statistical reasoning. D. Bernheim, for example, in defending a definition of rationalizability that builds in an epistemic independence condition, says this:

> ''*A question arises here as to whether an agent's probabilistic conjectures can allow for correlation between choices of other players. In a purely non-cooperative framework, such correlations would be nonsensical*: *the choices of any two agents are by definition independent events*; *they cannot affect each other. Consequently, I restrict players to have uncorrelated probabilistic assessments of their opponents' choices*.''[13]

But the non-cooperative framework requires *causal* independence, while the probabilistic correlations are purely epistemic. Just as coin flips (of a coin of unknown bias) can be known to be causally independent, but still be epistemically relevant to each other, so strategies of two players can be known to be causally independent even while information about one's choices is epistemically relevant (for a third player) to beliefs about the other.

A. Brandenburger and E. Dekel recognize the possibility that strategies chosen independently may still be epistemically correlated, and distinguish between what they call ''correlated'' and ''independent'' rationalizability, though they use potentially misleading terminology to describe the difference: ''The difference is that the second requires a player to believe that the other players choose their strategies independently, while the first does not.''[14] If this were a correct description of the difference, then it *would* be appropriate to take ''independently'' rationalizable strategies to be the only ones compatible with common belief in rationality, since the structure of the game requires that the players choose their strategies independently. The weaker notion of correlated rationalizability has application only because correlated beliefs are compatible with the knowledge that players choose their strategies independently.

P. Battigalli identifies, but does not endorse, an epistemic independence condition which he notes ''is assumed in many game-theoretic solution concepts: different players choose their strategies independently, and this is reflected in each player's probabilistic beliefs about her opponents, which satisfy a stochastic independence condition. We call this principle *strategic independence*.''[15] Again, this way of stating the principle

[13]Bernheim [7], 1014. This remark is quoted in [14], 147–148, where the same point is made about it.
[14]Brandenburger and Dekel (1987), 1392 [10].
[15]Battigalli (1996), 190 [4].

suggests an inference from causal to epistemic independence that may make the principle seem more plausible than it should.

Robert Aumann made the point, some years ago, that the fact that choices were made simultaneously did not imply stochastic independence of beliefs about them,[16] but I think that a conflation of causal and epistemic independence plays a role in one of Aumann's more recent arguments. To conclude this section on epistemic independence and backward induction I want to look in some detail at some of the ideas that motivate an argument by Aumann that common knowledge of rationality is enough to support backward induction reasoning in perfect information games.[17]

The intuitive defense of Aumann's argument makes explicit reference to counterfactual conditionals. My complaint will be that he equivocates between epistemic and causal 'if's. Before turning to the exposition of the argument, I will try to make the general point that there are two kinds of 'if' to be distinguished.

I want to emphasize that my criticism of Aumann's argument is a criticism of the intuitive ideas that motivate the proof. Given the formal definitions, the argument is unproblematic; the problem, I will argue, is with the interpretation of the result, and with the motivation for the definitions. I should also emphasize that the model that provides the context for Aumann's argument is different on a number of dimensions from the kind of model I have defined.

Aumann assumes common *knowledge* of rationality, and not just common belief, and there is no belief revision theory explicitly represented. Knowledge is assumed to have an S5 partition structure, which in the context of the more general models that we are using is equivalent to assuming that all beliefs are necessarily true. This assumption makes it difficult to make sense of belief revision, but I will argue that assumptions about belief revision are implicit in Aumann's motivation for his assumptions. Aumann's models contain no probabilities; an action is said to be rational if it is not *known* that some alternative action would result in higher utility. Aumann distinguishes a stronger and a weaker concept of rationality, as applied to players rather than actions, which he calls *material* rationality and *substantive* rationality. A player is materially rational in some possible state of the world if every choice *actually made* is rational; the player is substantively rational in some possible world only if in addition, for each *possible* choice, the player *would* have chosen rationally if he had had the opportunity to choose. The intuitive idea is something like this: crazy O'Leary would drive home drunk if I gave him the chance, so since I know this, I take away his car keys. This action of mine does render O'Leary materially rational by taking away his opportunity to exercise his disposition to act irrationally, but it does not eliminate his substantive irrationality, since he still has that disposition. The notion of perfect rationality defined in our models is an attempt to capture the intuitive idea of substantive rationality, since it requires that a person be disposed to act rationally even in situations that he believes, or even knows, will not arise. But whether perfect rationality entails Aumann's concept of substantive rationality will not be clear until we are clear about how to understand the counterfactuals in his definition of that notion.

[16] Aumann (1976) [2].
[17] Aumann (1995) [3].

Aumann argues that if there is common knowledge at the start of the game that all players are substantively rational, then the backward induction outcome will be realized. I will argue that the argument works only because a strong and implausible belief revision policy has been implicitly built into the definition of substantive rationality, and that this is done by equivocating on different senses of the 'if' used to explain substantive rationality – specifically, a causal and an epistemic sense. Before looking at the argument, I will try to clarify the distinction between the two kinds of 'if' that I want to distinguish.

Examples are familiar in the philosophical literature: one may believe that if Shakespeare had not written Hamlet, it would never have been written, while at the same time believing that if Shakespeare didn't write Hamlet, someone else did.[18] The first is a belief in a causal counterfactual, a belief based on beliefs about the cause of the writing of the play. The second is an expression of a belief revision policy – a policy about how to revise one's beliefs upon receiving the surprising information that Shakespeare did not write Hamlet. One believes the counterfactual, but one would give it up if one learned that the antecedent were true.

Just to make the general pattern clear, let me tell two stories: Suppose I initially believe the following three things: first, General Smith is a shrewd judge of character – he knows (better than I) who is brave and who is not. Second, the general sends only brave men into battle. Third, Private Jones is cowardly. It follows from these three propositions that Jones will not be sent into battle, so I also initially believe that. Let us assume that someone is cowardly if he would run away under fire. So I believe that Private Jones would run away if he were to be sent into battle (which, for that reason, he won't be). These are my beliefs, but how would I revise them if I learned that I was mistaken – for example if I learned that the general had sent Jones into battle? Since I think the general is a better judge of character than I, I would revise by giving up my belief that Jones is cowardly. Of the three beliefs mentioned above, the first two are more robust than the third. This, it seems to me, is a perfectly consistent belief revision policy, but it implies that I initially believe both (1) that Jones would run away if he were sent into battle, and (2) that if Jones *is* sent into battle than he won't run away. The first is a causal counterfactual – a belief about how Jones is disposed to behave – how he would behave in circumstances that I believe will not in fact arise. The second is a belief revision policy. Now I might be quite certain about all three of my beliefs, and I might be right about all three. But the belief revision question still makes sense. About any two logically independent beliefs, however certain, I might (in counterfactual circumstances) be forced to choose between them.

A second example: consider a poker game. I know (1) that Alice cheats – she has seen her opponent's cards. I also know (2) that Alice has a losing hand, since I have seen both her hand and her opponent's. In addition I know (3) that Alice is rational, so I conclude that she will not bet. But how would I revise my beliefs if I learned that Alice did bet? I would have to give up one of the three beliefs – which one depends on the relative strength of my evidence. It might be perfectly reasonable for me to be disposed

[18]The particular example is due to Jonathan Bennett. Examples of this kind were first proposed in [1].

to give up (2). In this case, I believe that if Alice *were* to bet, she would lose (since she has a losing hand), but if I were to learn that she *did* bet, I would give up that belief, concluding that she will win.

To locate the equivocation between these two kinds of 'if' in Aumann's argument, let me present a prima facie counterexample to his conclusion, and then see why it is not a counterexample to the claim as he understands it. I will describe a very simple perfect information game, and a model for it.[19] Here is the game:

This is a common interest game – the numbers at the terminal nodes (see Fig. 1) represent payoffs for both players. There are four strategies for Alice (D_1D_2, D_1A_2, A_1D_2, A_1A_2) and two for Bob (d, a). The backward induction solution is (A_1A_2, a). Here is a model for the game in which this solution is nowhere realized: There is just one state x. The strategy pair realized there is (D_1A_2, d). The knowledge partition, of course, is $\{x\}$ for both players, so in this model, there will be common knowledge of everything that is true. Now Alice's D_1 is a best response to Bob's d, and A_2 – the purely counterfactual part of Alice's strategy – would be a best response, if she were to find herself in a position to apply it, so it will be uncontroversial that Alice is substantively rational. But what about Bob's rationality? Bob is *materially* rational, whatever he does, since he does not get a chance to act, but is he *substantively* rational? Would action d be rational for Bob, *if* he had a chance to act? That depends on what Bob would know, or believe, if that were to happen, which depends on how he is disposed, at the start of the game, to revise his beliefs, were he to learn that Alice acted differently than he in fact knows she will act. Everyone will agree that Bob knows, and so believes, that Alice chooses either strategy D_1A_2, or strategy A_1D_2 (since this is a logical consequence of his knowledge that she chooses D_1A_2). Suppose that this belief of his is robust: he is disposed to revise his beliefs, were he to learn that Alice chose A_1, by concluding that since she was irrational once, she will be irrational again, and so would choose D_2. (It would be enough if he concluded only that for all he then would know, she *might* act irrationally again.) Bob cannot be faulted for having this belief revision policy, for as we have seen, Alice would indeed be irrational if she chose A_1, and it is agreed that Bob knows that this action would be irrational. Why should Bob's knowledge of Alice's rationality require him to believe that if she were (contrary to what he knows) to be irrational once, she wouldn't be irrational again? But if Bob does have

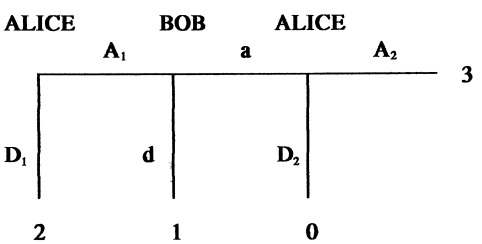

Fig. 1. A simple common interest game.

[19]This same game and a similar model is discussed briefly in [15].

this not unreasonable belief revision policy, then he is substantively rational in the sense that he chooses a strategy such that every action prescribed by it would be rational if it were implemented – a strategy that prescribes an action that Bob would believe (in the situation in which the action would be performed) would lead to a better outcome than the alternative action. If we understand substantive rationality in this way, then this model is a counterexample to Aumann's backward induction conclusion.

Now as mentioned above, Aumann's models have no representation of belief revision, but the intuitive idea of substantive rationality used to motivate the argument makes no sense without some at least implicit assumptions about how beliefs are revised. Bob is rational only if he *would* act rationally *if* his node were reached, but we can assess Bob's rationality in that counterfactual circumstance only if we have some way of determining what he *would* know if that circumstance were realized. Aumann's definitions do provide a counterfactual knowledge state for Bob, but I will argue that it is one that can be justified only by conflating causal and epistemic 'if's.

How do Aumann's definitions avoid the counterexample? According to his definition of rationality, to determine whether Bob is rational, take that part of Alice's strategy that determines what she does on her second move, ignoring her first move, which is no longer relevant. Assume that since Bob knows that Alice chooses D_1A_2, he knows even if she doesn't choose D_1, she will still choose A_2. The idea seems to be that Bob should reason as follows: ''I know that Alice will choose D_1A_2. The only part of her strategy that is relevant to my choice is A_2. So I will assume that if my node is reached, she will choose A_2.'' But why should Bob assume this if he knows that if his node is reached, then Alice did *not* choose D_1A_2? What Aumann's definition of substantive rationality does is implicitly to build an epistemic independence assumption into the belief revision policies of all rational players, an assumption that is considerably stronger than the epistemic independence assumption discussed above. Aumann's assumption is that not only beliefs about different players, but even beliefs about different parts of a single player's strategy are epistemically independent. In effect, no information about earlier moves in a game are permitted to be epistemically relevant to conclusions about any later moves.

The fallacy is this: Bob has the following initial belief: Alice would choose A_2 on her second move *if* she had a second move. This is a causal 'if' – an 'if' used to express Bob's opinion about Alice's disposition to act in a situation that they both know will not arise. Bob knows that since Alice is rational, if she somehow found herself at the second node, she would choose A_2. But to ask what Bob would believe about Alice *if* he learned that he was wrong about her first choice is to ask a completely different question – this 'if' is epistemic; it concerns Bob's belief revision policies, and not Alice's disposition to be rational. No assumption about Alice's substantive rationality, or about Bob's knowledge of her substantive rationality, can imply that Bob should be disposed to maintain his belief that she will act rationally on her second move even were he to learn that she acted irrationally on her first.

An analogy: there are two coins, one two-headed, and the other two-tailed. Take the two-headed one: flip it, and if the result is tails, flip it again. You get to bet on the second flip, conditional on there being a second flip. How do you bet? If you are Aumann, you should bet heads (even at very unfavorable odds) for the following reason: since you

know the coin is two-headed, you know it will land heads the first time, and so won't be flipped a second time. But since it is two headed, if it *were* flipped a second time, it is certain to land heads then too. But your decision about the conditional bet should not be based on your present beliefs about the dispositions of the coin, but on what you would believe if you learned that the condition for the bet were realized. Granted that *if* the two-headed coin were flipped a second time, it would land heads, but if the coin comes up tails the first time, then, it seems reasonable to believe, it must not be the two-headed one. If (contrary to what I know) that were to happen, I would conclude that probably the coin is the two-tailed one, and so I'll bet on tails. This reasoning takes the first flip to be epistemically relevant to the question what would happen if there were a second flip.

In the discussion in his paper, Aumann explicitly addresses the possibility of learning, in the course of a game, from previous play, and he rejects the claim that rationality, as he understands it, imposes an epistemic independence assumption that excludes such learning: ''Rationality permits the players to take account of past play in any way they want when forming estimates as to what will happen in the future. In fact, it places no restrictions at all on what they think others will do.''[20] Aumann's claim is that even where players may learn from past play, the assumption of *common knowledge* of rationality will still suffice to force the backward induction solution. But this cannot be right. In the model defined above, if Bob is permitted to learn from experience, and to draw conclusions about Alice's future play from evidence about her past play, then he will be able to conclude, were he to learn that she acted irrationally once, that she might act irrationally again. And if it is true in the actual state that this is what he *would* conclude, then he is substantively rational if he is disposed to play *d*, in which case there is common knowledge, in the actual state, that all are substantively rational.

Aumann, in his argument, emphasizes that we are talking about *knowledge* of rationality, and not just belief. If I *know* that the coin is two-headed, or that Alice is rational, then how could I possibly lose by betting on heads, or by being disposed to choose *a*? I can't, nor could I possibly lose by betting on tails, or going down. This is not relevant to substantive rationality; the question is, what is likely to happen *if* the coin lands tails the first time, or *if* Alice makes the irrational choice. It is the epistemic 'if' that is relevant to this question.

One might argue that if I *know* that Alice is rational, or that the coin is two headed, then I cannot make sense of the possibility that I am mistaken. Now it may be that if I am absolutely certain of something, and have never even imagined the possibility that it is false, then I will not have an explicit policy in mind for how to revise my beliefs upon discovering that I am wrong. But I think one should think of belief revision policies as dispositions to respond, and not necessarily as consciously articulated policies. Suppose Bob is absolutely certain that his wife Alice is faithful to him – the possibility that she is not never entered his mind. And suppose he is right – she is faithful, and he really *knows* that she is. We can still ask how Bob would revise his beliefs if he walked in one day and found his wife in bed with another man. We need not assume that Bob's absolute certainly implies that he would continue to believe in his wife's faithfulness in

[20] Aumann (1995), 17 [3].

these circumstances, and we need not believe that these circumstances are logically impossible. We can make sense of them, as a logical possibility, even if they never occur to Bob. And even if Bob never thought of this possibility, it might be true that he would react to the situation in a certain way, revising his beliefs in one way rather than another.

6. Robust belief in rationality and forward induction

There is a different argument, based on a quite different assumption about belief revision policies, that might instead be used to reject the outcome that is realized in the game model sketched just above. In that model, Bob would have judged Alice to be irrational had she chosen A_1, as indeed she would have been, given her beliefs. But instead of changing his opinion about Alice's rationality, maintaining his beliefs about her (passive) beliefs, he might have maintained his belief in her rationality, changing his beliefs about her beliefs about him. That is, Bob's belief in Alice's rationality might have been robust; if it had been, and if Alice had believed that it was, then in this game, the subgame perfect outcome would have been realized after all. In this section, I want to consider the consequences of a *rationalization principle*, a belief revision policy that requires the belief that all are rational to be robust, and the kind of strategic reasoning that it supports. I will state a characterization theorem about the consequences of common belief in the robustness of the belief that all are perfectly rational – a result that determines the set of strategies that can be realized in models in which the belief revision condition is satisfied. Then I will consider the extent to which this assumption about belief revision supports forward induction reasoning. The upshot will be that while some so-called forward induction reasoning is supported and explained, there is no unlimited iteration of such reasoning, and the model theory helps to explain why. I will illustrate the point by looking at the notorious money burning game.

Battigalli states the rationalization principle as follows: ''*A player should always try to interpret her information about the behavior of her opponents assuming that they are not implementing 'irrational' strategies.*''[21] In the context of our game models with their belief revision structures, we might state the rationalization principle this way:

A player should believe that all players are perfectly rational, and this belief should be robust relative to any compatible information about the behavior of any player.

That is, if you are surprised by the actions of some player, you should change your beliefs about that player's passive beliefs, rather than about her rationality. If possible, find an alternative hypothesis about her beliefs about other players that will make what she does perfectly rational.

Now suppose we assume that this principle is adopted by all players in a game, and that it is common belief that it is adopted. What will the consequences be? This assumption will constrain the strategy choices in some games, but what is interesting is how limited the effect of this assumption is. In contrast with the independence assumptions, this principle does not support iterative or inductive arguments. The reason

[21] Battigalli, (1996), 179 [4].

for the contrast is this: in general, a player's beliefs about what another player will do are based on an inference from two other kinds of beliefs: beliefs about the passive beliefs of that player, and beliefs about her rationality. If one's prediction based on these beliefs is defeated, one must choose whether to revise one's belief about the other player's beliefs or one's belief that she is rational. The rationalization principle says that one should keep the belief in rationality, which will normally require one to change beliefs about her beliefs. But the assumption that the rationalization principle is common belief is itself an assumption about the passive beliefs of other players, and so it is itself something that (according to the principle) might have to be given up in the face of surprising behavioral information. So the rationalization principle undermines its own stability, in contrast with the independence assumptions, which reinforce their own stability. I should emphasize that this is not a logical problem, or a reason to be suspicious of the rationalization principle. There is nothing problematic or unreasonable about a belief about the belief revision policy of another person that is not robust – that one is disposed to revise in the face of certain evidence.

To be specific, the following algorithm will define, for any game, the strategies that are realizable in any sufficiently rich model in which it is common belief that all players are perfectly rational, and that all players adopt the rationalization principle: Eliminate weakly dominated strategies for *just two* rounds, and then eliminate *strictly* dominated strategies iteratively. It can be proved that all and only strategies that survive this process are realizable in sufficiently rich models in which it is common belief that all players are rational, and that all revise their beliefs in conformity with the rationalization principle stated above.[22] So while this principle supports some so-called forward induction reasoning, the reasoning does not iterate indefinitely. Before giving an example to illustrate the point, I will comment on the qualification made in stating the result: that the models must be ''sufficiently rich.''

We have imposed certain closure conditions on our models in order to ensure that there are enough possible worlds to represent the causal structure of the game, but our definition of a model does not ensure that all of the logically possible combinations of beliefs and actions are represented in some possible world. Our models are not ''universal'' in any sense. In fact, to avoid technical complications that are irrelevant to

[22] I leave the proof of this to the reader, but I will sketch very quickly the main idea. To show that any strategy that survives the elimination process is realizable in such a model, we construct a model in which all such strategies are realized in some world compatible with common belief. The construction will use strategy profiles (more than one copy of some of them) to represent the possible worlds. The set W will be the union of three disjoint sets: W_1 will contain one world for each profile of strategies that survive the elimination process; W_2 will contain one world for each profile of admissible strategies; W_3 will contain one world for each member of C, the set of all profiles of the game. The Q relations and P functions will be defined so that, first, all players are perfectly rational in all W_1 and W_2 worlds. (This is possible since for any admissible strategy, there are beliefs relative to which it is perfectly rational.) Second, the Q relations are defined so that only W_1 worlds are compatible with the beliefs of players in W_1 worlds, so that it will be common belief in any W_1 world that a W_1 world is realized. Third, the Q's are defined so that within W_1 worlds, W_2 worlds have priority over W_3 worlds. Since every admissible strategy profile is played in some W_2 world, and since all players are rational in all W_2 worlds, this will ensure that all players conform to the rationalization principle in all the W_1 worlds. So in any W_1 world, it is common belief that all are perfectly rational, and all conform to the rationalization principle.

the conceptual issues that were the main focus of interest, we have assumed that our models are all finite. So even if there are models in which some strategy could be rationally chosen, there is no assurance that a given model will contain a possible world in which the player has the beliefs that will make that strategy rational. But the rationalization principle would be no constraint at all unless we assume that whenever rationality is possible, there are possible worlds in the model in which players act rationally. So we will say that a model is *sufficiently rich* only if for any players i and j, for any possible world x, and or any admissible strategy s for player i, there is a possible world y such that $x \approx_j y$, $S_i(y) = s$, and i is perfectly rational in y. This is not a substantive constrain on anyone's beliefs; it is only the assumption that if it is logically possible for x to play s rationally, then it is *conceivable* for j that i should have the beliefs that make it rational for i to play s. In a less simplified theory that gave up the finiteness assumption, this assumption would be a consequence of much more general and more natural assumptions that required the models to be universal in some sense, but such assumptions would require much more complicated models, and raise conceptual as well as technical problems. Since these more general assumptions, however they are spelled out, will ensure that the models are sufficiently rich in the sense defined, our results based on this ad hoc assumption can be expected to be sustained in a more natural, if more complex, setting.

Now to illustrate the result; if we apply the algorithm to the money burning game, two strategies will remain for each player, and a model can be constructed in which the rational thing to do is for the first player to burn the money, and for the second player to expect the first to burn it. Here is the game, and the model:[23]

The basic game is a battle of the sexes: Alice chooses U or D, while Bob chooses L or R, with payoffs as follows:

	L	R
U	4,1	0,0
D	0,0	1,4

But prior to the game, Alice has the option of burning two units. Bob knows, before making his choice, whether Alice has taken this option. So the strategic form of the whole game is as follows:

	LL	LR	RL	RR
BU	2,1	2,1	−2,0	−2,0
BD	−2,0	−2,0	−1,4	−1,4
NU	4,1	0,0	4,1	0,0
ND	0,0	1,4	0,0	1,4

Here is the idea of the model: Bob believes (correctly) that Alice will choose BU, because she believes (correctly) that Bob will play strategy LR. But if Bob were

[23] This famous game is discussed in [6] and in [16].

surprised by Alice choosing *N*, he is disposed to infer that she believed instead that he was choosing strategy *RR*, and so would make the rational response to this belief, *ND*. Alice is still rational in the world in which she chooses *ND*, and so Bob's belief revision conforms to the rationalization principle. Alice, whatever her belief revision policies, is perfectly rational, since *BU* is the *only* rational response to her belief about Bob. The final steps in the usual forward induction argument (arguing that the only tenable strategies are *NU* and *LL*) are blocked, since we cannot assume that belief in the rationalization principle itself will be robust.

Robustness is a comparative notion: to assume that a belief is robust inevitably forces us to assume that competing beliefs (including beliefs about the belief revision policies of others) are less robust. So the stability of belief in the rationalization principle is not compatible with the principle itself. One could make belief in this principle *somewhat* robust, adding that one should, if possible, preserve *both* the belief in the rationality of others and also the belief that others adopt the rationalization principle. This further assumption will buy you just one more round of elimination of weakly dominated strategies (in the middle of the process). Only if one assumes a specific infinite hierarchy of belief revision priorities can one be sure that unlimited iteration of forward induction reasoning will work. (One needs a policy something like this: first priority: all are rational; second, all believe that all are rational; third, all believe that all believe that all are rational, etc.). But it seems to me that such detailed assumptions about belief revision policy (requiring, for example, that a ten layered iterated belief in perfect rationality should have priority over an eleven layered iterated belief) have no intuitive plausibility.

Suppose, for example, you and I are playing an iterated prisoners' dilemma with a hundred rounds. I expect you to defect, but you surprise me by cooperating. Being a follower of the rationalization principle, I assume that your beliefs about me make it rational for you to cooperate. You must have thought that by cooperating, you could get some mutual cooperation started, and would benefit. But what do you think now, after one round in which I defected, perhaps surprising you, and what will you do next? Have I blown my chance by defecting, or should I now believe that it would be beneficial to cooperate? The rationalization principle does not say, and I don't think there is any plausible principle of belief revision that can give a general answer to this question. The best I can do is to form an hypothesis about what might best explain your unexpected action (which might include an hypothesis about what kind of hypothesis of this kind I think you expect me to form).

7. Conclusion

Faced with surprising behavior in the course of a game, the players must decide what then to believe. Their strategies will be based on how their beliefs would be revised, which will in turn be based on their epistemic priorities – whether an unexpected action should be regarded as an isolated mistake that is thereby epistemically independent of beliefs about subsequent actions, or whether it reveals, intentionally or inadvertently, something about the player's expectations, and so about the way she is likely to behave in the future. The players must decide, but the theorists should not – at least they should

not try to generalize about epistemic priorities that are meant to apply to any rational agent in all situations. They shouldn't try to generalize even about the restricted class of situations in which there is common belief or knowledge of rationality. What the theorist should offer is some conceptually clean descriptive concepts and a general framework for the representation of the deliberative reasoning of rational agents in strategic situations.

Appendix A

Sketch of the proof of the theorem stated in Section 5

Let Γ be any perfect information game in agent form with no relevant ties in payoffs. (Each player has only one possible move, and distinct outcomes have distinct payoffs for each player.) Let M be any model for Γ in which it is common belief (1) that all players are perfectly rational, and (2) that all players have belief revision policies that treat beliefs about different players as epistemically independent. Then the unique subgame perfect equilibrium profile (the backward induction solution) is realized, and it is common belief that it is realized.

The proof is by induction on the size of the game. Suppose a game has one node. Then trivially, the only rational strategy for that player is the subgame perfect strategy. Now assume that the theorem holds for games with k nodes or fewer, and assume that Γ is a game with $k+1$ nodes. Let M be any model for the game that satisfies the condition, and let $\mathbf{w}$ be either $\mathbf{a}$ or any world that is compatible with what player 1 believes in $\mathbf{a}$ (that is $\mathbf{w}\in\{\mathbf{a}\}\cup\{x\in W\colon \mathbf{a}R_1x\}$). We define a model for each of the immediate subgames (each of the games Γ^s that follow one of player 1's possible moves s) in the following way: for subgame Γ^s, $W^s=W\cap[s]$, $\mathbf{a}^s=f(\mathbf{w},s)$[24], and the S_i^s, Q_i^s, and P_i^s's are simply the restrictions of those functions and relations to W^s. We show that for each $s\in C_1$, M^s is a model that satisfies the condition. It will then follow, by hypothesis of induction, that for each M^s, it is true and common belief in $\mathbf{a}^s$ that the subgame perfect strategy profile is played. This implies that in the original model, players 2 to n all play subgame perfect strategies, and that player 1 believes this. Since by definition, a best response by player 1 to the subgame perfect strategies for all the other players is the subgame perfect strategy, and since player 1 is perfectly rational, it follows that he chooses this strategy.

The substance of the proof is the argument that in each of the submodels, it is

[24]See the causal independence condition stated in note 4. This condition guarantees, for each world x and strategy s, the existence of a world $f(x,s)$ representing the world that would be realized if strategy s were chosen instead of the one chosen in x. (If s is the strategy chosen by the relevant player in x, then $f(x,s)=x$). Since strategy choices are (causally) independent of the prior beliefs of other players, Player 1 could not influence the prior beliefs of other players by choosing s, which is to say, in the counterfactual world in which he does choose s, the prior beliefs of players 2 to n are exactly as they are in the actual world. Causal independence is enough to ensures that the *prior* beliefs are the same, but the additional epistemic independence condition will be required for the different conclusion that certain *posterior* beliefs remain the same.

common belief among the players who play in that subgame that all those players are perfectly rational, and that each player treats beliefs about different players as epistemically independent.

Let proposition X be the following proposition (in the original model M): it is common belief among players 2 to n that players 2 to n are perfectly rational, and that players 2 to n treat beliefs about different players as epistemically independent.

First, note that proposition X is a proposition about players 2 to n, in the technical sense we have defined, while proposition $[s]$ (that player 1 chooses strategy s) is a proposition about player 1. Hence in any world that is a member of X, all players believe X, and have belief revision policies that treat X and $[s]$ as epistemically independent, which means that they would continue to believe X even upon learning $[s]$. That is to say, for each player i and world y such that $y \in X$, $B_{i,y}([s]) \subseteq X$. (Recall that the belief revision function, $B_{i,y}$, is defined in terms of Q as follows: $B_{i,y}(\phi) = \{z \in \phi$: for all $x \in \phi$ such that $y \approx_i x$, $xQ_i z\}$)

Second, we note a fact about the R relations of the submodel: recall the way that, in general, R is defined in terms of Q: $xR_i y$ iff for all z such that $z \approx_i x$, $zQ_i y$. So the submodel R relations will be defined as follows: For all i and all $x,y \in W^s$, $xR_i^s y$ iff for all $z \in [s]$ such that $z \approx_i x$, $zQ_i y$. In other words $xR_i^s y$ iff $y \in B_{i,x}([s])$.

Third, recall that $\mathbf{a}^s$, the actual world of the submodel is $f(\mathbf{w},s)$, a world in which players 2 to n have exactly the same beliefs as they have in $\mathbf{w}$, an arbitrary world that is compatible with player 1's beliefs in the actual world of the original model. So since $\mathbf{w} \in X$, it follows that $\mathbf{a}^s \in X$.

Now from these three facts: (1) for any i and w, if $w \in X$, then $B_{i,w}([s]) \subseteq X$; (2) $wR_i^s y$ iff $y \in B_{i,w}([s])$, and (3) $\mathbf{a}^s \in X$, it follows that $\{w: \mathbf{a}^s R^{s*} w) \subseteq X$. ($R^*$ is the transitive closure of the R relations, and so is the relation that defines the set of worlds compatible with common belief. R^{s*} is the common belief relation for the model M^s for the subgame Γ^s.) That is, it is common belief in the submodel that proposition X is true.

Two thing remain to be shown before we are finished: first, since the players have different beliefs in the submodels than they have in the original model, it cannot simply be assumed that if player i is perfectly rational in world w in the original model, then she will be perfectly rational in the same world in the submodel, nor that if she satisfies the epistemic independence condition in a world of the original model, then she satisfies it in the same world in the submodel. That is, it cannot be assumed without argument that the proposition X means the same thing in the submodel as it means in the original model. But because of the way perfect rationality is defined, and because this is a perfect information game, we can be sure that players 2 to n will be perfectly rational in the submodel if they are perfectly rational in the original model. For suppose a player were to adopt a strategy t that is inferior, in the submodel, to some strategy t'. Then in the original model, t will be inferior to this strategy: (t' if the subgame is reached, t otherwise). So perfect rationality in the original model implies perfect rationality in each submodel. Also, it can be shown that if ϕ and ψ are propositions about players 2 to n, then for any $w \in X$ and player $i > 1$, if ϕ is epistemically independent of ψ for i in w, then $\phi \cap [s]$ is also epistemically independent of $\psi \cap [s]$ for i in w. This implies that ϕ and ψ will be epistemically independent in the submodel if they are epistemically independent in the original model.

So it follows that it is common belief, in each submodel, that players 2 to n (which includes all the players who play in any of the subgames) are perfectly rational, and that each has a belief revision policy that treats propositions about different players as epistemically independent. This completes the proof.

References

[1] E. Adams, Subjunctive and indicative conditionals, Foundations of Language 6 (1970).
[2] R. Aumann, Subjectivity and correlation in randomized strategies, Journal of Mathematical Economics 1 (1976) 67–96.
[3] R. Aumann, Backward induction and common knowledge of rationality, Games and Economic Behavior 8 (1995) 6–19.
[4] P. Battigalli, Strategic rationality orderings and the best rationalization principle, Games and Economic Behavior 13 (1996) 187–200.
[5] E. Ben Porath, Rationality, Nash equilibrium and backward information games, Review of Economic Studies 64 (1997) 23–46.
[6] E. Ben Porath, E. Dekel, Signaling future actions and the potential for self sacrifice, Journal of Economic Theory 57 (1987) 36–51.
[7] B. Bernheim, Rationalizable strategic behavior, Econometrica 52 (1984) 1007–1028.
[8] L. Blume, A. Brandenburger, E. Dekel, Lexicographic probabilities and choice under uncertainty, Econometrica 59 (1991) 61–79.
[9] A. Brandenburger, Lexicographic probabilities and iterated admissibility, in: P. Dasgupta et al. (Eds.), Economic Analysis of Markets and Games, MIT Press, Cambridge, MA, 1992.
[10] A. Brandenburger, E. Dekel, Rationalizability and correlated equilibrium, Econometrica 55 (1987) 1391–1402.
[11] P. Gärdenfors, Knowledge in Flux: Modeling the Dynamics of Epistemic States, MIT Press, Cambridge, MA, 1988.
[12] A. Grove, Two modelings for theory change, Journal of Philosophical Logic 17 (1988) 157–170.
[13] G. Mailath, L. Samuelson, J. Swinkels, Extensive form reasoning in normal form games, Econometrica 61 (1993).
[14] B. Skyrms, The Dynamics of Rational Deliberation, Harvard University Press, Cambridge, MA, 1992.
[15] R. Stalnaker, Knowledge, belief, and counterfactual reasoning in games, Economics and Philosophy 12 (1996) 133–163.
[16] E. Van Damme, Stable equilibria and forward induction, Journal of Economic Theory 48 (1989) 476–496.