Theory and Decision (2006) 61: 1–19
DOI 10.1007/s11238-006-7868-y

OLIVER BOARD

THE EQUIVALENCE OF BAYES AND CAUSAL RATIONALITY IN GAMES

ABSTRACT. In a seminal paper, Aumann (1987, *Econometrica* 55, 1–18) showed how the choices of rational players could be analyzed in a unified state space framework. His innovation was to include the choices of the players in the description of the states, thus abolishing Savage's (1954, *The Foundations of Statistics*. Wiley, New York) distinction between acts and consequences. But this simplification comes at a price: Aumann's notion of Bayes rationality does not allow players to evaluate what would happen were they to deviate from their actual choices. We show how the addition of a causal structure to the framework enables us to analyze such counterfactual statements, and use it to introduce a notion of causal rationality. Under a plausible causal independence condition, the two notions are shown to be equivalent. If we are prepared to accept this condition we can dispense with the causal apparatus and retain Aumann's original framework.

KEY WORDS: causal decision theory, counterfactuals, Game theory, independence, rationality

1. INTRODUCTION

The so-called Bayesian approach to game theory takes the view that games should be analyzed as a number of interrelated single-person decision problems in the sense of Savage (1954), with each player maximizing expected utility with respect to some subjective probability distribution over a set of uncertain events, in this case the strategy choices of the other players. This approach, pioneered by Bernheim (1984) and Pearce (1984), was originally contrasted with the equilibrium approach to games, according to which probabilities can only assigned to events not governed by rational

decision makers. In equilibrium, strategies are not the subject of uncertainty.

Aumann (1987) provided a reconciliation of these two approaches. He proved that, if the players are Bayesian expected utility maximizers, and possess a common prior over the space of uncertainty (which includes each player's strategy choice), then they will each play their part in some correlated equilibrium. And there is now a large and growing body of literature which seeks to characterize game-theoretic solution concepts explicitly in terms of epistemic conditions on expected-utility maximizing agents.[1] We shall refer to this literature as the *Bayesian tradition*.

But although Bayesian decision theory is at the heart of Aumann's paper, the framework that he (and others) adopt is not that of Savage. For in the Savage framework, a distinction is made between acts and consequences, the former being a function from the set of states of the world to the latter. If one's opponents' strategy choices are included among the objects of uncertainty, and hence are part of the description of a state, this distinction implies that we must have a different state space for each agent. Aumann overcame this problem by adopting a unified framework in which acts and consequences are both parts of the description of the state of the world. He describes this as the "chief innovation" in his model (p. 8). In particular, a state describes the strategy choice of *every* agent. Thus Aumann's framework is very much like that of Jeffrey (1965), where personal choice is included as a state variable. An act is now a subset of the state space: precisely, the set of states at which that act is carried out.

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

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

2. AN EXAMPLE

Consider the following one-person decision problem:[2] President Clinton wants to have an affair with Monica Lewinsky, but fears that doing so may lead to his impeachment. The utility he gains from each of the possible outcomes is shown in Figure 1(a) below, where A is the event that he has an affair, and M the event that he is impeached. His prior probability distribution over the state space is given in (b). It is clear from the utilities that having an affair is the unique Bayes rational act. This is true whatever the values of p, q, r, and s.[3] For A is a dominant strategy: whatever the posterior probabilities of M and $\bar{M}$ once he has updated on his private information, it will yield higher expected utility than $\bar{A}$. He reasons as follows: "Either I will be impeached or I would not be. Whether or not I will be impeached, I prefer to have an affair with Monica than not to, so I should go ahead and have one".

Of course, this reasoning is fallacious: it was, at least indirectly, Clinton's affair with Lewinsky that led to his

(a)

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

(b)

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

Figure 1. Clinton and Lewinsky.

impeachment, and if he had predicted this, he should have avoided the affair. The problem is that the object of uncertainty, i.e. whether or not the president will be impeached, is not independent of the acts being considered. Note that this cannot happen in the Savage framework, where acts are functions from states to consequences.

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

$M M$: I shall be impeached whatever I do;
$M \bar{M}$: I shall be impeached if and only if I have the affair;
$\bar{M} M$: I shall be impeached if and only if I do not have the affair;
$\bar{M} \bar{M}$: I shall not be impeached whatever I do.

These four events hold independently of A and $\bar{A}$. And it is easy to check that, if he places enough weight of probability on the second of these, the expected utility maximizing act is $\bar{A}$. A is no longer a dominant strategy. This is effectively the path taken by Aumann. In the context of a game, one player's uncertainty is another player's strategy choice. Suppose for instance that Clinton is actually playing a game against Congress, which decides whether or not to impeach after observing whether or not Clinton has an affair. Then Congress has four strategies, corresponding precisely to the four events described above. In Aumann's framework, each state describes the entire profile of strategies, one for each player, and since the definition of a strategy allows players to condition actions on available information, it may be reasonable to suppose that the necessary independence condition is satisfied. However, it is not clear exactly what this independence condition should be. We shall argue below that it has nothing to do with the independence of the probability distributions, which represent each agent's beliefs. Rather what is required is a casual independence condition that cannot be expressed without a richer model than that employed by Aumann.

Another way to avoid the kind of fallacy illustrated by the original example is to adopt a version of *causal decision theory*[4] as our principle of rational choice. According to causal decision theory, the weights we should use for our expected utility calculations are not simple conditional probabilities (where the conditioning event is that we carry out some particular act), but conditional causal chances. In other words, we must consider how likely each uncertain event (opponent's strategy choice) is given what we are actually going to do, and how likely it *would be* if we were to do something different.[5] According to causal decision theory, it is not rational for Clinton to have the affair, because if he were to avoid it, he would not be impeached. By comparing causal rationality with Aumann's concept of Bayes rationality, we shall show that the latter implicitly makes a causal independence assumption. We shall state this assumption explicitly, and then discuss whether it is reasonable in the context of normal and extensive form games.

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

3. BAYES RATIONALITY AND CAUSAL RATIONALITY

3.1. *Aumann's framework*

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

- a finite[6] set W, with generic element w;
- for each player i, a binary relation $\mathcal{B}_i$;

- for each player i, a probability measure p_i on W;
- for each player i, a function $f_i : W \to S_i$.

The set W represents the set of *states of the world*, and w is one particular state. The binary relations $\mathcal{B}_i$, called *accessibility relations*, encode the players' information at each state.[7] At a given state w, the set of states that player i considers possible is given by $\{x : w\mathcal{B}_i x\}$. The propositions that i believes are just those that are true at every state in this set. In order to ensure that is beliefs are coherent, we assume that this set is nonempty for every $w \in W$ (i.e. we assume that $\mathcal{B}_i$ is serial: $(\forall w)(\exists x) w\mathcal{B}_i x$). The probability measure p_i is is *prior* over W, from which we obtain her probabilistic beliefs at each state of the world by updating on her information at that state. Thus is (subjective) probability at w that some proposition ϕ is true, denoted $p_{i,w}([\phi])$, is given by

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

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

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

Condition (*OK*) says that, at every state a player considers possible, the strategy she carries out is the same as the one she actually carries out. Note that this rules out the possibility of the player *trembling* (see, e.g. Selten (1975)), and accidentally playing a strategy other than the one she intended. In order to introducing trembles, we would need to make

a distinction between *decisions*, the objects of choice, and *performances*, the strategies actually carried out (see Shin (1989)). The player would know her own decision, but not necessarily her performance.

3.2. *Bayes rationality*

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

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

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

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

3.3. *Causal rationality*

According to the Bayesian decision theory set up above, each player forms a subjective probability assessment over her opponents' strategy profiles by updating her prior with respect to her private information, which includes information about which strategy choice she will carry out. She

then evaluates alternative strategy choices according to this probability assessment. Causal decision theory, on the other hand, recognizes that the various actions of each player might be inter-connected: my opponents' choices given that I play s_i might not be the same as they would have been had I chosen to play s_i'. Each player must consider what her opponents will do given her actual choice, and also what they would do if she were to choose something else.

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

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

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

But the framework above gives us no way of evaluating counterfactuals, and so no way of evaluating this definition. To this end, we follow the Stalnaker–Lewis theory of counterfactuals (see Stalnaker (1968) and Lewis (1973)), and augment the model with a *closeness* relation, $\preccurlyeq_w$, for each state w. Each $\preccurlyeq_w$ is a binary relation on W which satisfies the following four conditions:

(C1) $\preccurlyeq_w$ is complete;
(C2) $\preccurlyeq_w$ is transitive;
(C3) $\preccurlyeq_w$ is antisymmetric (for all x, y, if $x \preccurlyeq_w y$ and $y \preccurlyeq_w x$, then $x = y$);
(C4) $\preccurlyeq_w$ is centered (for all x, $w \preccurlyeq_w x$).

So, for a given state, w, the closeness relation $\preccurlyeq_w$ gives a total ordering of the states, with w at the bottom. The lower a state

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

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

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

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

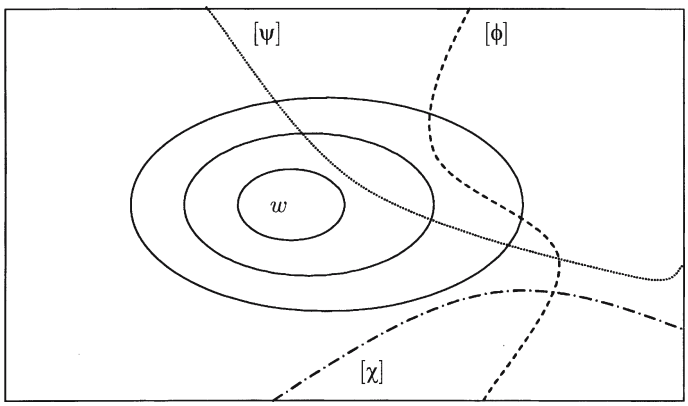

Figure 2. Evaluation of counterfactuals.

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

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

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

In the next section, we compare Bayes rationality with causal rationality. We shall find that the whether or not the two coincide hinges on a particular independence assumption, and we discuss how appropriate the assumption is.

4. CAUSAL INDEPENDENCE IN GAMES

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

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

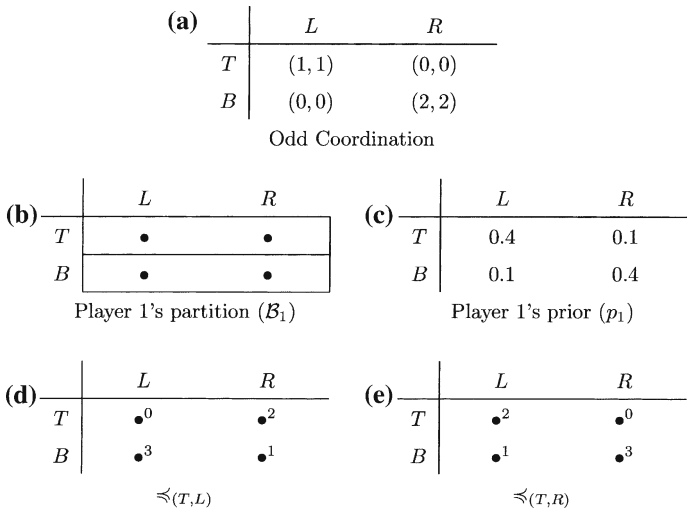

Figure 3. Analysis of odd coordination.

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

But there is something odd about the causal structure of this game. If the players are moving simultaneously, or at least in ignorance of each other's choice (as is often considered to be an implicit assumption of the normal form representation of a game), then their strategy choices *should* be independent of each other. Indeed, Harper (1988) goes so far as to say "a causal independence assumption is part of the idealization built into the normal form", and Stalnaker (1996) writes "... in a strategic form game, the assumption is that the strategies are chosen independently, which means that the choices made by one player cannot influence the beliefs or the actions of the other players". Similarly, appeal is often made to some causal independence condition to reject the symmetry argument for rational co-operation in the prisoner's dilemma: the two players will indeed end up doing the same thing, but if one were to deviate and cooperate, the other would still defect (see, e.g.

Dekel and Gul (1997)). This causal independence condition is most easily expressed in the language of counterfactuals: if one player were to do something different, the others players would still do the same. We can state this condition as a formal property of a model:

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

In other words, if at some world w in the model, player i plays strategy s_i and the other players play s_{-i}, then at the closest possible world in which i plays s_i' instead, the other players still play s_{-i}. It is clear that this gives us the causal independence condition stated above. It should be noted, however, that (CI) is a global condition: causal independence is assumed to hold at every world in the model. This implies not only that the players' strategies are causally independent of each other, but also that they believe this to be the case (and indeed that this is common belief). In fact, since counterfactuals enter the causal expected utility calculus only as aspects of players' (probabilistic) beliefs, it is players' beliefs in causal independence rather than causal independence itself that drives the result we about about to prove.

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

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

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

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

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

The intuition behind the proof is straightforward: the event that you play s_{-i} is just the same as the event that "if I were to play s_i you would play s_{-i}", since under (CI) my action has no causal influence on yours, and you will carry on doing the same thing *whatever* I do. Thus my probabilistic evaluation of your various strategies is the same whether we hold my strategy fixed (as Bayes rationality does) or whether we vary it (as causal rationality does). Theorem 1 is an extremely convenient result.[9] It allows us to dispense with the causal apparatus developed above and continue using Aumann's simple model to analyze rational play in normal form games, as long as we have the required causal independence.

We must take care to distinguish the type of causal independence discussed above from independence of the probability functions, p_i. There are two possible independence conditions that might be imposed on the p_i functions. First, we might require that the probabilities one player assigns to the strategies of different opponents be independent of each other (Bernheim (1984) and Pearce (1984), among others, impose this constraint on player's beliefs). But, as Stalnaker (1996) points out, our causal independence assumption "has no consequences about the evidential relevance of information about player one's choice for the beliefs that a third party might rationally have about player two". Consider, for example, a third player reasoning about a simple coordination game played by two twins. Even though fully convinced about the causal independence of the twins' strategy choices, she might expect that they end up playing the same strategy, whatever that might be. Hence,

there seems no reason in general to rule out players holding correlated conjectures about their opponents' strategies.

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

(c) represents the prior of the optimist—for her both strategies are Bayes rational; (d) gives the prior of the pessimist—for him neither is. Jeffrey (1983) calls these cases "pathological", but we see no reason to exclude them. In any case, in the absence of any theory about the decision-making process itself, our models are perhaps best viewed merely as tools for the theorist to determine whether a given choice of an agent is rational, rather than anything to which the agent herself might appeal. The construction of a model of the latter type, as a means to explicating the decision-making process, seems to be Jeffrey's rather more ambitious goal; but Aumann states explicitly that this is not his aim: "The model describes the point of view of an outside observer" (p. 8). But for us there seems to be no reason to rule out the

(a)

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

Simple Coordination

(b)

	L	R
T	•	•
B	•	•

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

(c)

	L	R
T	0.4	0.1
B	0.1	0.4

Player 1's prior

(d)

	L	R
T	0.1	0.4
B	0.4	0.1

Player 1's prior

Figure 4. Analysis of simple coordination.

pessimist (though a psychologist might tell him to reconsider his beliefs). His example demonstrates a kind of non-existence of equilibrium: he will never be *a priori* happy with the choice he has made. Of course, this is not a case of non-existence of a rationalizable outcome. A strategy is rationalizable if there is *some* set of beliefs, satisfying certain conditions, for which it is Bayes rational., i.e. if it is rational for the agent in *some* model satisfying certain requirements.

The next obvious step is to consider how widely applicable the causal independence assumption is. Harper and Stalnaker, quoted above, claim that it is almost axiomatic for normal form games. The suggestion seems to be that the same might not apply for extensive form games (indeed, Harper goes on to make this point explicitly). But the normal form and the extensive form are merely alternative representations of a given situation of strategic interaction. What is really at issue is the move order, and in particular whether the players' moves in the game are simultaneous or sequential (or more precisely, whether each player moves in ignorance or her opponents' moves or not). Only in as much as the normal form is often used to represent simultaneous move games, while the extensive form is used when moves are made sequentially, might causal independence be appropriate for the former and not the latter.

Consider a sequential version of the famous battle-of-the-sexes game: Alice moves first, and chooses to go the *pub* or the *cafe*. Bob learns of her choice and is then faced with the same choice himself. Of course Bob's action is likely to depend on Alice's choice, but his choice of strategy already takes this dependence into account: a player's strategy is a list of counterfactual statements, which describe what he does or would do in every situation in which he might be called on to make a choice. In this simple game, Bob has four strategies: (*pub* if *pub*, *pub* if *cafe*), (*pub* if *pub*, *cafe* if *cafe*), (*cafe* if *pub*, *pub* if *cafe*), and (*cafe* if *pub*, *cafe* if *cafe*); each strategy describes what he decides to do given what Alice has done. One problem with taking strategies rather than actions as the objects of choice, however, is that it is unclear when if ever the

players will actually make a choice between the various strategies available to them. Although we could think of a hypothetical pre-play stage when such choices are made, it seems more appropriate and more accurate to think of the players as making their choices as and when they are on move, and to evaluate rationality at information sets. Indeed, this is the approach that the majority of the work in this area takes.[10] Thus a player's strategy choice is really a number of different choices, one for each information set at which he is on move. Nevertheless, it seems clear that the causal independence assumption is still appropriate, because the causal dependencies between various actions are already built into the definition of a strategy. What Bob would do if Alice were to go to the cafe cannot depend on whether or not Alice actually does go the cafe, since the counterfactual already assumes that she does.

5. COMMENTS AND CONCLUSIONS

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

and causal rationality coincide. Even though we analyze only normal form games here, where it is usually assumed that the players move simultaneously, we argue that the causal independence condition in appropriate even when players move sequentially, as represented by the extensive form. Thus the equivalence result justifies the use of Aumann's non-causal framework.

NOTES

1. e.g. Aumann and Brandenburger (1995). Dekel and Gul (1997) and Battigalli and Bonanno (1999) review this literature.
2. This example is a re-labeled version of Gibbard and Harper's (1978) Case 1. Thanks are due to Ehud Kalai for suggesting this modern version of the story of David and Bathsheba.
3. As long as $p + q > 0$ and $r + s > 0$, so Bayesian updating is well defined in all cases. This seems to be a reasonable requirement: a player's prior should not rule out any of her strategy choices.
4. See Gibbard and Harper (1978), Lewis (1981), Skyrms (1982), and Sobel (1986) for alternative statements of this theory.
5. A recent paper by Zambrano (2004) offers an interesting variant of causal rationality (which he calls *W-rationality*), in which a player evaluates each action not according to her *actual* beliefs about what her opponents would do if she chose that action, but rather according to what she *would* believe her opponents would do if she took that action. This is similar in flavor to Jeffrey's (1965) evidential decision theory, where one's action provides information about the liklihood of various uncertain events.
6. The assumption that W is finite is not without loss of generality, even in finite games (see, e.g. Battigalli and Siniscalchi (2002)), but for the current purposes there is no need to deal with the additional complications raised by the infinite case.
7. Aumann assumes that these relations are equivalence relations, and hence partition the set of states; but for our purposes there is no need to make this rather restrictive assumption.
8. That there is a unique such world (implied by the antisymmetry of $\preccurlyeq_w$) is a property that the present account of counterfactuals shares with Stalnaker's theory but not with Lewis's. This property makes valid the law of *Conditional Excluded Middle:*

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

   (see Lewis (1973) for a discussion of the appropriateness of this law). For the current purposes, it is analytically very convenient, as it

saves us the need to evaluate what Sobel (1986) calls *practical chance conditionals*: "if it were the case that ϕ, then it might, with a probability of p, be the case that ψ". Furthermore, since for the causal expected utility calculus it is always *agents' beliefs about* the relevant counterfactuals that we shall be considering, the assumption is without loss of generality: our agents may be unsure about what the closeness relation is.

9. A similar result has been established by Shin (1992), but in a very different framework to that of the current paper. Specifically, Shin constructs a space of *possible worlds* for each player, along with a personal closeness measure, to evaluate the counterfactual beliefs of that player about the unified state space. There is no representation of the (objective) causal reality, and hence no way of expressing causal independence.
10. A notable exception is the work of Stalnaker: he discusses this issue in Stalnaker (1999, p. 315), and shows that, under certain assumptions, the two approches are equivalent.

REFERENCES

Aumann, R.J. (1987), Correlated equilibrium as an expression of Bayesian rationality, *Econometrica* 55, 1–18.

Aumann, R.J. and Brandenburger, A. (1995), Epistemic conditions for Nash equilibrium, *Econometrica* 63, 1161–1180.

Battigalli, P. and Bonanno, G. (1999), Recent results on belief, knowledge and the epistemic foundations of game theory, *Research in Economics* 53, 149–225.

Battigalli, P. and Siniscalchi, M. (2002), Strong belief and forward-induction reasoning, *Journal of Economic Theory* 106, 356–391.

Ben Porath, E. (1997), Rationality, Nash equilibrium and backwards induction in perfect information games, *Review of Economic Studies* 64, 23–46.

Bernheim, B.D. (1984), Rationalizable Strategic Behavior, *Econometrica* 52, 1007–1028.

Dekel, E. and Gul, F. (1997), Rationality and knowledge in game theory, in Kreps, D.M. and Wallis, K.W. (eds.), *Advances in Economics and Econometrics: Theory and Applications: Seventh World Congress*, Vol. 1, Cambridge University Press, Cambridge, pp. 87–172.

Gibbard, A. and Harper, W.L. (1978), Counterfactuals and two kinds of expected utility, in Hooker, C.A., Leach, J.J. and McClennan, E.F. (eds.), *Foundations and Applications of Decision Theory*, Vol. I, Reidel, Dordrecht.

Harper, W.L. (1988), Causal decision theory and game theory: a classic argument for equilibrium solutions, a defense of weak equilibria, and a

new problem for the normal form representation, in Harper, W.L. and Skyrms, B. (eds.), *Causation in Decision, Belief Change, and Statistics, II*, Kluwer Academic Publishers, Amsterdam.
Jeffrey, R.C. (1965), *The Logic of Decision*, McGraw-Hill, New York, NY.
Jeffrey, R.C. (1983), *The Logic of Decision*, 2nd ed., The University of Chicago Press, Chicago, IL.
Lewis, D. (1973), *Counterfactuals*. Basil Blackwell, Oxford.
Lewis, D. (1981), Causal decision theory, *Australasian Journal of Philosophy* 59, 5–30.
Osborne, M.J. and Rubinstein, A. (1994), *A Course in Game Theory*, The MIT Press, Cambridge, MA.
Pearce, D.G. (1984), Rationalizable strategic behavior and the problem of perfection, *Econometrica* 52, 1029–1050.
Savage, L.J. (1954), *The Foundations of Statistics*, Wiley, New York.
Selten, R. (1975), Reexamination of the perfectness concept for equilibrium points in extensive games, *International Journal of Game Theory* 4, 25–55.
Shin, H.S. (1989), Two notions of ratifiability and equilibrium in games, in Bacharach, M. and Hurley, S. (eds.), *Foundations of Decision Theory*, Basil Blackwell, Oxford.
Shin, H.S. (1992), Counterfactuals and a theory of equilibrium in games, in Bicchieri, C. and Dalla Chiara, M.L. (eds.), *Knowledge, Belief, and Strategic Interaction*, Cambridge University Press, Cambridge.
Skyrms, B. (1982), Causal decision theory, *Journal of Philosophy* 79, 695–711.
Sobel, J.H. (1986), Notes on decision theory: old wine in new bottles choice, *Australasian Journal of Philosophy* 64, 407–437.
Stalnaker, R. (1968), A theory of conditionals, in Rescher, N. (ed.), *Studies in Logical Theory*, Basil Blackwell, Oxford.
Stalnaker, R. (1996), Knowledge, belief and counterfactual reasoning in games, *Economics and Philosophy* 12, 133–163.
Stalnaker, R. (1999), Extensive and strategic form games: games and models for games, *Research in Economics* 53, 293–319.
Zambrano, E. (2004), Counterfactual reasoning and common knowledge of rationality in normal form games, *Topics in Theoretical Economics* 4, article 8.

Addresses for correspondence: Oliver Board, Department of Economics, University of Pittsburgh, Pittsburgh, PA 15260, USA. Phone: +1-412-648-1748; E-mail: ojboard@pitt.edu