CRISTINA BICCHIERI AND MITCHELL S. GREEN

SYMMETRY ARGUMENTS FOR COOPERATION IN THE PRISONER'S DILEMMA

I. INTRODUCTION

A variety of philosophers, decision theorists, and game theorists have advanced arguments according to which rational agents who play a one-shot prisoner's dilemma (PD) should choose to cooperate if they know either that they are in identical circumstances, or are in some sense identical twins. The reasoning in favor of the coöperative solution raises the question whether the PD with the appropriate "Identicality" assumption is inconsistent, since there is a separate argument employing dominance reasoning that favors non cooperation. We argue here that the question can only be answered relative to a clarification of the Identicality assumption. There turns out to be only one interpretation of the Identicality assumption that justifies cooperation, but this interpretation may be controversial. On another interpretation of this assumption the description of the PD game is indeed inconsistent, while on the remaining interpretations of that assumption the argument for the cooperative solution is unsound. Seeing the inconsistency teaches us a lesson concerning the kinds of postulates that can be added to the description of a game. Seeing the unsoundness will motivate a more careful treatment of the relation of modal concepts to the notion of rational choice, and will help us to illuminate the relevance to game theory of technical tools developed by logicians.

II. MOTIVATING EXAMPLE

Consider the classical "prisoner's dilemma" game:

		Player 2	
		C	*D*
Player 1	*C*	3, 3	0, 4
	D	4, 0	1, 1

Figure 1

Player 1 chooses a row and Player 2 chooses a column. Furthermore, it is assumed that the players' choices are causally independent of one another, in

G. Holmström-Hintikka and R. Tuomela (eds.), Contemporary Action Theory. Vol. II, 229–249.

that one player's choice does not influence the other player's choice or beliefs. The payoff matrix above gives, for each pair of choices of the two players, the utility of Player 1 and of Player 2, respectively, where 4 is the highest utility and 0 the lowest.[1] We assume players' utilities to encompass all the relevant psychological and social characteristics that might influence their choices, such as altruism, envy, the pursuit of some moral or political goal, and so forth. We also assume that the game in Figure 1 is played exactly once by the players and that they believe this to be the case.

The informal characterization of the game given above can be expressed more rigorously with explicit postulates concerning players' payoffs and actions as well as their rationality and beliefs.[2] One such postulate is that each player's act is causally independent of the other player's act, as expressed below:

> *Independence*: For each player i and each strategy s_i available to i, if i plays strategy s_i, then for each player i' and each strategy $s_{i'}$ available to i', it is causally possible for i' to play $s_{i'}$.[3]

Independence implies that if Player 1 chooses *C*, it is causally possible that Player 2 plays *C* and causally possible that Player 2 plays *D*. The same holds for Player 1's choice of *D*. The situation is symmetric as between the two players.[4]

We also employ a Rationality assumption according to which a player will choose a strictly dominant strategy if one exists, provided that the Independence postulate holds.[5] As formulated, only in the presence of a dominant strategy does the Rationality assumption predict a choice under uncertainty.[6] If a player does not have a dominant strategy, the Rationality assumption does not predict any choice. We could strengthen our Rationality assumption by saying that whenever a player is in a situation of uncertainty, she will choose the action that maximizes her expected utility. This further requirement is unnecessary, however, in treating the PD, since in this game both players have dominant strategies. Finally, we postulate that the full description of the game, as well as the assumptions of Rationality and Independence, are common belief among the players.[7] These postulates jointly imply that the players' choice will be (*D*,*D*). This conclusion has, nevertheless, been a source of puzzlement since (*C*,*C*) Pareto-dominates (*D*,*D*), i.e. both players would have been better off had they decided to play *C*.[8]

On the other hand we now also see how one might formulate an argument in favor of joint cooperation. The proponent of the cooperative solution may add to our description of the PD game a postulate according to which the players are qualitatively identical in some sense. Although the exact meaning of such an "Identicality" assumption is far from clear, for present purposes we shall take it as saying that, necessarily, Player 1 chooses *C* (*D*) iff Player 2 chooses *C* (*D*), leaving open for the moment the interpretation of 'necessarily'. As we shall see later, a great deal turns on whether we construe this modality in a causal or doxastic sense.

There would appear to be no immediate reason for game theorists to refuse to allow the Identicality assumption to be added as a postulate governing the PD. After all, the game has a unique dominant strategy – indeed, the same one – for both players, and given that Identicality appears to be consistent with Independence, the Rationality assumption will still imply that both players will play *D*. However, with these same assumptions one may also suggest the following reasoning:

> Suppose that I am Player 1. Since Player 2 is identical to me, she will end up making the same choice as I do. That is, if I play *C*, so will she, and the same applies to *D*. In particular, if I choose *C* the outcome is (*C*,*C*) and my payoff is 3. If I choose *D*, the outcome is (*D*,*D*) and I get 1. Thus *C* dominates *D*, and since I am rational I will play *C*.

Thus, so the argument goes, if identical rational players, who are aware of their identicality, play the PD, they will choose to cooperate.

There seems, then, to be a paradox. If both players believe in Identicality, then each of them believes that "My playing *C* implies that the other player plays *C*" is true, and the Rationality assumption appears to entail that each player will choose *C*. On the other hand, we also argued with the help of the Rationality assumption that each player will choose *D*. In other words, barring our finding a flaw in one of these two lines of reasoning, we should conclude that our description of the game is after all inconsistent. This is the case even though the Identicality assumption itself does not contradict the implication of the other assumptions (namely, the conclusion that the players will play *D*).

III. THE IDENTICALITY ASSUMPTION: DOXASTIC INTERPRETATIONS

Many game theorists would reject the reasoning favoring the cooperative solution on the ground that such reasoning ignores the causal independence of different players' choices.[9] Such game theorists would contend that the choice situation presented in the argument for cooperation is best represented by the following one-player game (where $i=1,2$), to which we will refer as the "Mirror's Choice":

Player i	*C*	3
	D	1

Figure 2

This matrix incorporates the identicality of the two players by not leaving room for the possibility of their doing different things. According to this representation each player is facing a single-agent decision problem in which he has two possible choices, *C* or *D*, the outcome of each of which is certain. The game theorist would argue that Figure 2 should not be used to represent the original game, even if the players in it end up making identical choices. The reason is that even though the players in the originally described scenario make identical choices, and believe that they will do so, it does not follow that they could not behave differently from one another. Yet the Mirror's Choice representation of their situation presupposes exactly this.

Even adhering to the original representation of the PD, however, it is not clear that the argument in favor of the cooperative solution is sound. For unless it can be proved that the postulates defining the PD together with the Identicality assumption imply that the Mirror's Choice correctly characterizes the PD, each player is in a position to reason as follows: "The other player will in fact choose as I do and she is going through the same reasoning and will reach the same conclusions that I will. In these circumstances, playing *C* seems to be an acceptable choice, but what would happen were I to choose *D* instead?" For all we have said so far, even with the Identicality assumption each player might find that the answer to this question leads him to opt for *D*. We therefore do well to treat the question with some care.

Consideration of subjunctive conditionals such as the foregoing is central to deliberation.[10] In deliberation, one typically considers several scenarios that differ from one another only with respect to the action chosen, and evaluates the consequences of the alternative feasible actions. Deliberation of this sort may also involve the provisional settling upon one course of action and exploration of what would happen were one to deviate from that course. It may involve, for example, deliberation about what would happen were one to choose irrationally. Having arrived at the provisional conclusion that her best choice is, say, to play strategy s_1, the player may ask what would happen were she to choose an alternative strategy s_2 instead. If the expected outcome of s_2 is preferable to that of s_1, then a rational player will give up her provisional commitment to perform s_1.

In order to represent alternative possibilities, as well as players' deliberation processes, it will be helpful to provide a model-theoretic framework within which we can perspicuously represent what a player believes and how possibilities are entertained. The usual description of a strategic-form game *G* is a triple $<N, S_{i \in N}, u_{i \in N}>$, where *N* is the set of players and, for each player $i \in N$, S_i is the set of pure strategies available to *i*, and u_i is player *i*'s utility function.[11]

The description of *G* is, however, only a partial specification of the decision problem faced by the players, since it gives players' utilities and feasible actions but does not specify players' beliefs about one another or about the game, nor what the players are actually going to do. A model of a game G then represents a completion of the partial specification of the decision problem given by the definition of G.[12] A *model M of game G* is a quadruple $<W, w^*, C, <s_i,R_i>_{i \in N}>$, where

(i) W is a nonempty set (the set of "possible worlds", each of which is a realization of exactly one complete play of the game G),
(ii) $w^* \in W$ (w^* is the "actual world"),
(iii) $C \subseteq W^2$ is a binary relation of nomic accessibility such that wCw' iff w' is consistent with the laws of nature that hold at w.
(iv) (a) for each $i \in N$, s_i is a function from W to S_i;
(b) R_i is a binary relation defined on W that gives, for each world $w \in W$, a subset W' of W each member of which is consistent with what i believes at w.

A few comments on the model M and how it applies to our game are now in order. The set W represents all the things that could happen in a given game. When a game G is represented in the strategic form, for each cell in the matrix there will be at least one member of W.[13] Moreover, clause (iv)(a) gives, for each $w \in W$ and each player i, a pure strategy played by that player in that world. Each model of a game G thus represents a complete play of that game.

R_i is a relation of doxastic possibility, giving the set of possible worlds that are consistent with what i believes at w^*. To refer to this set we write $\{w: w^*R_iw\}$. This is the set of states of affairs that, for all that player i believes at w^*, could be actual. In order to impose certain intuitive conditions on players' beliefs, we must further specify the properties of the accessibility relation R_i. For example, R_i might not be reflexive, because our concern is beliefs, which might be false. To ensure that i has coherent beliefs in each world, R_i must be serial, that is, for any world w there must exist at least one world w' such that wR_iw'. Also, to ensure that players know their own minds, we require that R be transitive and Euclidean.[14]

In a model of the sort just defined, it is natural to construe a proposition as a subset of the set of possible worlds.[15] For example, the proposition that Player 1 plays D is identified, in a given model M, with the set of possible worlds w' of M for which $s_i(w')=D$. Because the set W is constrained by the definition of G, what a player can believe in a given model is limited by the description of the game she is playing. We say that a proposition p is *doxastically necessary* at world w for agent i in model M if and only if p is true in all worlds w' that are doxastically accessible to i at w in M (i.e., iff p is true in all w' such that wR_iw' in M). We also say that a proposition p is *doxastically possible* at world w for agent i in model M if and only if its negation is not doxastically necessary for i at w in M.[16]

We have interpreted the accessibility relation R_i as a relation of doxastic possibility, but we have also defined a model for a game as containing a relation of causal possibility. C is a binary relation defined in W that gives, for each world $w \in W$, a set of possible worlds that are consistent with the laws of nature that hold at w. Thus we say that a world w' is causally possible relative to w just in case everything that occurs at w' is consistent with the laws of nature that hold at w. Since a proper description of a game must list all the possible outcomes that may result from all the possible combinations of players' strategies, and we are modeling games in which players have common beliefs about the structure of the game (i.e., all the available strategies and payoffs), we must assume that every outcome in a game is causally possible, even if some of them are ruled out by rationality considerations. Thus, for the games we are considering, it is the case that $\cup_{i \in N} R_i \subseteq C$, that is, for all w,w' in W, if wR_iw', then wCw'.[17]

The primary aim of incorporating a relation of causal possibility into our models is to ensure the satisfaction of the Independence assumption. We may do so by stipulating of the C relation that:

> for each player i, and each $w \in W$, if i plays strategy s at w, then for each $i' \in N$ (where $i \neq i'$) and each $s' \in S_{i'}$, there is a world $w' \in W$ such that wCw' and $s_{i'}(w') = s'$.

Observe that an action may be doxastically impossible for a player at a world w and yet be causally possible relative to w.[18] For example, given that John has made up his mind what he will do next week, his gambling away all his money in Las Vegas will be doxastically impossible for him, but doing so would not violate any law of nature. If he wanted to, he could go to Las Vegas and fritter away his savings.

With causal independence thus guaranteed, we may stipulate the satisfaction of the Rationality assumption by requiring that if M is a model of game G then the actual world, w^*, of M is one in which no player plays a dominated strategy. In conjunction with the requirement of causal independence, this stipulation has the consequence that although in each model of a game no player acts irrationally in the actual world of that model in the sense of playing a dominated strategy, in each such model it is causally possible that some player plays such a strategy.

Following Stalnaker (1993), common belief is here defined as the transitive closure R^* of the set of all the R_i relations. That is, wR^*w' iff there exists a finite sequence of worlds $w_1,\ldots,w_n$ such that $w = w_1$ and $w' = w_n$, and for each k from 1 to $n-1$, there is some j such that $w_kR_jw_{k+1}$. For any proposition p, p is common belief at w just in case $\{w' : wR^*w'\} \subseteq p$. We require that in any model of the PD game the structure of the game and the Rationality and Independence assumptions are common belief among the players. As we are about to explain, one can give several interpretations of the Identicality assumption, but for each such interpretation we shall take it that the assumption is common belief among the players.

An advantage of providing a class of models for a game G is that it is only with reference to a model that we can meaningfully represent a player's deliberation process. Each model for G represents a particular play of G and completes the partial specification of the decision problem defined by G in a way that is compatible with the conditions imposed by G's definition. Each model of a game will contain many possible worlds, since we need to represent not only what happens in a particular play of the modeled game, but also what might happen in alternative situations compatible with the description of the game. To ensure that a model M represents all the options that the definition of G says are open to the players, we have required that for every possible world $w \in W$, every player i and every strategy choice s open to i, there is a causally accessible world w' in which i has the same beliefs about the game as she has in w and in which she plays s' (where $s \neq s'$).[19] For example, when a player is considering what would happen were she to choose, say, a dominated strategy, she is not thinking of an alternative game in which she has different beliefs from those she has in the actual world of the model. What the player is contemplating is a situation in which she makes what is by her lights an irrational choice, and we want to represent a situation such as this with a possible world in which such an action is taken. The possible world in which a rational player chooses a dominated action may not be doxastically accessible to the player in the actual world, because that player may believe that she will not choose a dominated strategy. Such a world is nevertheless causally possible. Finally, we want to stress that because a player's beliefs can differ in different models of a given game, strictly speaking it is only relative to a model that we can ask what it is rational for a player to do. In what follows we shall always raise questions of rationality with respect to a given model of a game.[20]

Consider now what it is rational for players to do in models of the PD game in which both players believe that they will necessarily do the same thing. This question admits different answers depending upon what kind of necessity is being invoked. Suppose first that the Identicality assumption is understood to mean that in the actual world of the model players believe that they will act alike (and it is common belief that they will). Thus for each player, "The players make the same choice" is true in all worlds that are doxastically accessible to that player from the actual world.[21] On such an interpretation, Player 1 will believe that if she plays C, Player 2 will play C, and if she plays D, Player 2 will play D. Analogously for Player 2. A model of the game incorporating the Identicality assumption so construed is one in which in the actual world of that model it is doxastically necessary for both players that they do the same thing. In such a model it can still be the case that off-diagonal outcomes (i.e., those outcomes that result from a play of (C,D) or (D,C)) are causally possible, and that players are aware of this.

Now players may not believe that an off-diagonal outcome is causally possible. If they do not, then as we shall explain in Section IV it may be rational for them to cooperate. Our question is instead whether just by virtue of being in a situation in which it is doxastically impossible for both players that an off-diagonal outcome is played, the players can reason to a cooperative solution. One might argue that since Player 1, for instance, believes that either (C,C) or (D,D) will be played, then given that the payoff of the former outcome is greater than that of the latter, the cooperative outcome is the rational choice. Player 2 may reason in the same way and conclude in favor of playing C. Moreover, the players' conclusions will be common belief.[22]

This reasoning is fallacious. It depends upon the false premise that if two actions, x and y, are the only doxastically possible options, then if the payoff of x is greater than that of y, it is rational to choose x. To see that this premise is false, suppose that an agent i has, and believes she has, three actions open to her: x, y, and z, where her preference ordering is $x>y>z$, and '$>$' represents a strict preference relation. Imagine now that what i believes to be a seer tells her that she will not choose x – not that it is causally impossible for her to do so, but just that she won't do so. She continues to believe that her choice of x is causally possible, but she now believes that she will not choose x. As a result there are only two doxastically possible actions for i, namely the choices of y and of z. It does not, however, follow that it is rational for her to choose y. The rational thing to do is still to take x. This would not be the case if i came to believe that the choice of x is not causally possible, since it is reasonable to ignore what one takes to be causally impossible actions in one's deliberation. So long, however, as she does believe that x is causally possible, her believing that she will not take x does not make it rational for her to do something else.[23]

Not only is the above reasoning in favor of the cooperative solution fallacious, but the conclusion – that it is rational in the envisioned model for both players to play C – appears to be false. The reason is that a provisional commitment to perform a certain action is rational only if that commitment is stable under consideration of alternative possibilities, and Player 1's provisional commitment to play C, were she to make one, is not stable in this way. To see this, suppose that Player 1 has formed the plan to play C, and takes it that Player 2 will play C as well. If Player 1 now asks herself what would happen were she to play D instead, the answer would appear to be that Player 2 would nevertheless play C. For on a familiar semantical construal of subjunctive conditionals, 'Were A the case, then B would be the case', is true at world w iff in the world most similar to w in which A is true, B is true as well. In the world now in question, both players play C. Because by assumption there is no causal interaction between the two players, the most similar world to w in which Player 1 plays D (and has the same beliefs and utility function he has in w) is one in which Player 2 continues to play C.[24] Since, however, (D,C) nets Player 1 more than does (C,C), it follows that Player 1's provisional commitment to perform C is not rational. Because the cooperative choice is not robust under consideration of alternative possibilities, both players will reach the conclusion that D is the only rational choice. We conclude that models of the game incorporating a doxastic version of the Identicality assumption are not, as such, ones in which it is rational for players to cooperate.

IV. THE IDENTICALITY ASSUMPTION: NOMIC INTERPRETATIONS

Having considered a model of the PD game that incorporates a version of the Identicality assumption and finding it unable to support an argument for the cooperative solution, let us contemplate a second model, in which players believe that it is causally necessary that both do the same thing. Specifically, this is a model in which both players believe there is no possible world consistent with the laws of nature in which they do different things.[25] The reasoning in favor of the cooperative solution might go as follows. Since Player 1, for instance, believes that the set $\{(C,C), (D,D)\}$ contains all and only causally possible outcomes, then given that the payoff of the former outcome is greater than that of the latter, the cooperative outcome is the rational choice. Player 2 will reason in the same way and conclude in favor of playing C. Moreover, the players' conclusions will be common belief.

The premises of this reasoning conflict with the fact that players have common belief in the Independence assumption,[26] which requires that off-diagonal outcomes are causally possible. On the present interpretation of the Identicality assumption, however, players must also believe that off-diagonal outcomes are *not* causally possible. This pair of beliefs is ruled out by our requirement that the doxastic accessibility relation is serial, that is, that players' beliefs are internally consistent. There would appear, then, to be no model of the PD in which we construe the Identicality assumption in terms of causal necessity.

One can, however, distinguish two notions of independence, each one corresponding to a different conception of causal dependence. For just as the notion of causation itself admits of importantly distinct elucidations[27], so too may the notion of causal independence. Moreover, on one account of causal independence, we find that the Independence assumption is consistent with the nomic interpretation of the Identicality assumption. To see this consider the following construals of independence:

I_1: No matter what one player does, it is causally possible that the other plays either C or D.

I_2: No matter what one player does, that choice has no causal influence on the choice of the other player.

I_2 requires that there be no causal interaction between the players' actions, so that if these two actions are spacelike separated then I_2 will be satisfied. On the other hand, as we shall see presently there may be indeterministic point-events that, in spite of being spacelike separated, are such that it is not causally possible that they both occur. I_1 therefore implies I_2, but not *vice versa.*

The nomic version of the Identicality assumption is inconsistent with I_1, since according to I_1 it must be causally possible that (C,D) occur and causally possible that (D,C) occur. Independence interpreted according to I_2, however, is consistent with this version of Identicality. For on this nomic construal of Identicality there need be no causal influence between the acts of the two players since we may take the two choice-events to be spacelike separated. Yet in such a case it might nevertheless be causally impossible that the two players do different things, and this would be so in spite of the fact that the action of neither one has any effect on the action of the other.

Our question is thus whether there could be an example in which although it is causally necessary that the acts of the two agents are the same, their actions are nonetheless causally independent in the sense of I_2. The answer depends upon whether there could be an agentive analogue to the Einstein–Podolsky–Rosen phenomenon in quantum mechanics.[28] The phenomenon involves a pair of particles, #1 and #2, each of which could either go spin-up or spin-down. For each particle neither outcome is causally determined by states of affairs in its light cone. However, in the EPR situation it is causally necessary that if #1(#2) is spin-up, then #2(#1) is spin-down, and if #1(#2) is spin-down, then #2(#1) is spin-up, even though there is no causal interaction between the two particles in that no light ray could connect either point-event to the other.

An agentive analogue of EPR in the context of the PD would be a model of this game played by players 1 and 2, and satisfying the following conditions:

(i) Given that Player 1 chooses *C*, it is causally necessary that Player 2 chooses *C*, and given that Player 1 chooses *D*, it is causally necessary that Player 2 chooses *D*.

(ii) Given that Player 2 chooses *C*, it is causally necessary that Player 1 chooses *C*, and given that Player 2 chooses *D*, it is causally necessary that Player 1 chooses *D*.

(iii) The choice events are causally independent, in that the choice events are spacelike separated.

(iv) Neither agent's choice is causally determined by antecedent states of affairs.

(v) The above four conditions are common belief among the players.

We shall not try to construct such a model here, since doing so would require building enough spatiotemporal structure into each possible world in the model to capture some basic features of relativity and quantum indeterminacy. Yet we see no reason why such a construction could not be carried out.[29] Further, we readily grant that any case satisfying the above five conditions would be mysterious. Were we ever confronted with what seemed to be a case of this kind we would be justified in searching for a "hidden variable" that explains the correlation between the actions of the two players. Our contention, however, is only that a case of this kind, an "EPRPD", is for all we know a possibility.[30] Indeed, it appears to be the only kind of case in which Identicality (causally interpreted) and Independence are jointly satisfied.

Now although such an EPRPD does violate Independence construed as I_1, it does not violate that assumption construed as I_2. Yet we know of no good reason for preferring the first construal of the Independence postulate over the second. For this reason we are in no position to assert that the PD version of the EPR involves a violation of causal independence. We shall therefore treat the EPRPD as a case that does not violate the Independence assumption. Let us imagine, then, a model that satisfies the above five conditions. Such a model may be graphically represented by means of what in Section III we called the Mirror's Choice. Since the only two causally possible outcomes in this model are (C,C) and (D,D), and the players are aware of this fact, Rationality implies that each agent will choose C.

There appears to be no separate argument for the conclusion that the rational choice for each player in the model in question is to play D. Such an argument requires the premise that each of the off-diagonal outcomes is causally possible, and this premise is false in the present model. For while one of the off-diagonal outcomes is preferable to each of the players, neither is causally possible in at least one sense of 'causally possible', and it is reasonable to ignore what one believes to be a causally impossible outcome in one's deliberation. Observe further that a provisional commitment to play C is robust under consideration of alternatives. The reason is that because the actions of the two players are causally correlated, each player is able to see that were he to play D instead, the other player would play D as well. It would thus seem that if there can be a model of the PD satisfying the five postulates that we have used to characterize the EPRPD, then in such a model it is rational for agents to cooperate.[31]

V. DOES THE SYMMETRY OF THE GAME IMPLY IDENTICALITY?

Proponents of the argument for cooperation will perhaps now urge that we have misconstrued their aim. Their reason might be that one need not treat Identicality as a postulate that is satisfied only in certain models of the PD. Rather, Identicality, appropriately construed, is just the product of symmetry and rationality and so holds for all models of the PD. Thus Rapoport writes

> ...because of the symmetry of the game, rationality must prescribe *the same choice to both* [players]. (Rapoport 1966, 142.)

In the most detailed formulation of this approach to date, Davis (1977) largely follows Rapoport, replacing rationality with common knowledge thereof. Speaking of one player's belief that the other player's choice will be a "mirror reflection" of his own, Davis says,

> To the contrary, it seems an obvious entailment of the assumption that each knows that each knows that each is rational, together with the symmetry of the situation. (Davis 1977, 322.)

Now rationality, even when it is common knowledge, does not alone imply that two players endowed with it will make the same choice, even if they are playing a symmetrical game.[32] A game may be symmetrical and both players may have two equivalent strategies that dominate all others, with the result that they could end up choosing different rational strategies. Symmetry and rationality (or common knowledge of rationality) alone do not entail that players will make the same choice. In the PD, however, symmetry and rationality entail the same choice for both players because there is a unique rational choice for both.

How, then, might one reason from the premise that both players will do the same thing to the conclusion that the rational action for each player is to play *C*? Further, we must consider whether this reasoning reveals a fallacy in the standard argument for the (*D*,*D*) solution, or whether it shows that there are models of the PD in which it is *both* rational to play *D* and to play *C*.

In the passage from which the above quotation is drawn, Rapoport reasons in favor of the cooperative solution:

> ...because of the symmetry of the game, rationality must prescribe *the same choice* to both. But if both choose the same, then (*C*,*C*) and (*D*,*D*) are the only possible outcomes. Of these (*C*,*C*) is clearly the better. Therefore I should choose *C*.[33]

Rapoport might mean by 'possible' here either 'causally possible' or 'doxastically possible'. Suppose that he intends causal possibility. Then if Rapoport is correct in concluding that (*C*,*C*) and (*D*,*D*) are the only causally possible outcomes, the PD reduces to the Mirror's Choice discussed above, in which *C* is indeed the rational choice. It does not seem, however, that symmetry and rationality could conspire to imply that (*C*,*C*) and (*D*,*D*) are the only causally possible outcomes of the game. For as we have remarked, a given player could do something irrational even if we theorists are confident that she should not. On the other hand, Rapoport might intend to pick out doxastic possibility in his use of 'possible'. As we have seen, however, there is no good inference from the premise that it is doxastically necessary for both players that both play *C*, to the conclusion that the rational choice for each of them is to play *C*. There are, then, two interpretations of Rapoport's claim that (C,C) and (D,D) are the only possible alternatives. On the first interpretation (in terms of causal necessity) the claim is unjustified, while on the second interpretation (in terms of doxastic necessity) the claim does not imply that the rational choice for both agents is to cooperate.

Assuming that we have exhausted the relevant alternatives for interpretation of Rapoport's notion of possibility, we must conclude that this author's argument for cooperation is defective. We turn therefore to a consideration of Davis' more detailed argument for cooperation.[34] In preparing to argue that the rational choice is for players to cooperate, Davis contends first that it will be the case that the players do the same thing. Both players are rational, and both rationality and the structure of the game are common knowledge. Further, the game is symmetrical, so that players have the same number of strategies, and the payoff to Player 1 of (C,D) = payoff to Player 2 of (D,C). As we have seen, a game may be symmetrical with both players having two equivalent strategies that dominate all the others; in such a case both players could end up choosing different strategies with no violation of their rationality. However, in the PD game there is only one dominant strategy for each player, and it is the same for both. Hence the symmetry of the PD game, together with the rationality of both players, imply that both players will do the same thing, that is, that the outcome will be an element of the set $\{(C,C),(D,D)\}$. Let us agree as well that both players know that the outcome will be an element of the set $\{(C,C),(D,D)\}$. Davis now argues that the rational solution to the game is (C,C). Abbreviated, his argument runs as follows:

1. An alternative x is rationally prescribed for an agent i if i knows that there are just two possible outcomes m and n, such that if i takes x then the outcome is m, if i does not take x then the outcome is n, and m is better (in i's judgment) than n.
2. Each player knows that the set $\{(C,C),(D,D)\}$ includes the only possible outcomes, and that if he takes C, then the outcome is (C,C), and that if he does not choose C, then the outcome is (D,D).
3. Each player knows that he judges (C,C) to be better than (D,D).

Davis infers from 1, 2, 3 that

4. C is rationally prescribed for each player.

We follow Davis in treating the conditionals in (1) and (2) as material rather than as subjunctive conditionals,[35] since if read as subjunctive, one of the conditionals contained in premise (2) would be false in all but EPR-style models of the PD. Furthermore, unlike what we found in Rapoport's discussion, there is no ambiguity in Davis' use of 'possible' in steps (1) and (2) of his argument; he makes it clear that the notion in question is one of epistemic possibility:

> The desired sense is *epistemic*, and *relative*: the outcomes (C,C) and (D,D) are said to be alone possible relative to each agent's *relevant information*.[36]

Davis' argument explicated in terms of material conditionals and epistemic possibility is thus

1*. An alternative x is rationally prescribed for an agent i if i knows that there are just two epistemically possible outcomes m and n, such that if i takes x then the outcome is m, if i does not take x then the outcome is n, and m is better (in i's judgment) than n.

2*. Each player knows that the set $\{(C,C),(D,D)\}$ includes the only epistemically possible outcomes, and that if he takes C, then the outcome is (C,C), and that if he does not choose C, then the outcome is (D,D).

3*. Each player knows that he judges (C,C) to be better than (D,D).

4*. C is rationally prescribed for each player.

The difficulty is that, even if the inference to (4*) is valid, premise (1*) interpreted in terms of material conditionals is implausible. Observe that the expression 'just two' can here mean only 'at most two' rather than 'exactly two'; but thus disambiguated premise (1*) leads to an absurd consequence. For suppose an agent can choose either \$10 or \$100, and that she knows that rationality prescribes choosing the \$100, and that she will do what rationality prescribes. Suppose the agent also knows that no one will give her \$1000, whatever she chooses. So it follows that she knows that either she chooses the \$100 and no one will give her \$1000, or she chooses the \$10 and someone gives her \$1000. She knows this since this disjunction is an obvious consequence of what she knows. So all the agents' epistemic possibilities are included in the set {choose \$100 and get \$100, choose \$10 and get \$1010}. The agent prefers the latter, so rationality prescribes that she chooses \$10.[37]

Premise (1*) would become plausible only if the conditionals appearing in it were interpreted as subjunctive conditionals. But if (1*) and (2*) were interpreted in terms of subjunctive conditionals, premise (2*) would be false. We must therefore conclude that the argument is unsound, no matter how it is interpreted.

ACKNOWLEDGMENT

We are grateful to R. Stalnaker for comments on an earlier draft of this paper, as well as to N. Belnap, I. Gilboa, W. Harper, D. Hubin, P. Humphreys, A. Margalit and B. Skyrms for helpful discussions of many of the issues treated here.

Cristina Bicchieri
Carnegie Mellon University

Mitchell S. Green
University of Virginia

NOTES

[1] The utilities are taken to be von Neumann–Morgenstern utilities (1944), which also reflect each player's decision under a situation of risk, but this additional assumption is not needed here.

[2] Examples of formal models are Kaneko (1987), Bicchieri (1988a, 1988b, 1993), and Kaneko–Nagashima (1990).

[3] Note that in extensive-form games of perfect information, our Independence assumption would be violated. Consider a game in which Player 1 moves first, and Player 2 can observe Player 1's move. All choices available to Player 2 at her decision node are causally possible once Player 1 has made his choice, and hence Player 2's choice *at her decision node* is not causally necessitated by Player 1's choice, even if Player 1's choice may give Player 2 a reason to choose a particular action at her decision node. However, Player 1's choice has restricted Player 2's choice-set by cutting off some initially possible paths along the decision tree. Thus, although Player 2's choices are causally independent of Player 1's choices at every (local) decision node at which Player 2 is called upon to play, Player 2's initial choice-set as defined by the game (i.e., Player 2's initial strategy-set) is not causally independent of Player 1's choices. If we construe causal independence as applying to choice-sets, then Independence as we define it is not satisfied in extensive-form games of perfect information. It is easy to verify that it is always satisfied in strategic-form games.

[4] We shall presently consider the Independence postulate with greater care, for one can distinguish between Independence as we have formulated it and a version of that principle according to which two acts are independent of one another just in case neither event influences the other. Which of these two construals of independence we adopt will affect our assessment of certain symmetry arguments.

[5] We are here referring to games in strategic form, where Independence as we define it applies. It might be objected that our formulation of Independence is too weak to guarantee the rationality of a dominant strategy, since a player's choice may alter the probability of another player's choices. To answer this objection, we need to distinguish between a player believing that her choice is evidence that the other player is making a similar choice, and her belief that her choice is *causing* the other player to make a similar choice. Causal decision theory allows us to make this distinction, whereas evidential decision theory does not. As it will be clear from the following discussion, we side with causal decision theory (Gibbard and Harper 1978).

Savage's classical formulation of decision theory does not allow beliefs that make an agent regard his choices as evidence about which of the states, on which the outcome depends, obtain. For Savage, acts are not even in the algebra of propositions for which his model provides subjective probabilities (1954, 8–17). Jeffrey (1965, 2nd edition) introduced a decision theory in which acts are in the algebra of propositions that the beliefs are defined for. His theory allows for the expression of the belief that other agents probably made the same choice as you did, and it makes such beliefs relevant to deliberation.

In a PD, the so-called evidential expectations $E(C)$ and $E(D)$ of C and D are thus:

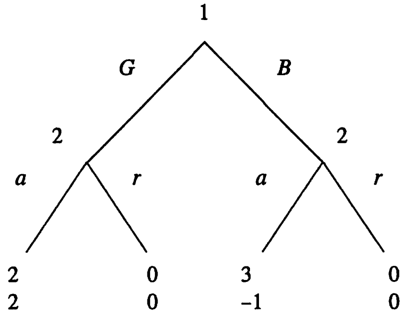
$$E(C) = p\ (C|C)\ u(C,C) + p\ (D|C)\ u(C,D)$$

$$E(D) = p\ (C|D)\ u(D,C) + p\ (D|D)\ u(D,D)$$

If one is convinced that what he does counts as evidence that the other player will do the same thing, then his personal evidential expectation for C could be higher than his evidential expectation for D. The dominance built into the Savage expectation, where for each action (C or D) of your opponent the same probability multiplier is used in the expectation for C as in the expectation for D (i.e., $p(C|C) = p(C|D) = p(C)$; and $p(D|C) = p(D|D) = p(D)$), can be broken up in the evidential expectation if $p(C|C)$ favors low defection (thus, it is greater than $p(D|C)$) and $p(D|D)$ favors high defection (thus, it is greater than $p(C|D)$). In the latter case, one could assign conditional probabilities that make the evidential expectation for C higher than the evidential expectation for D.

According to evidential expectation, sure thing reasoning can lead to fallacies unless the probabilities of the states are the same no matter which act is performed. There are two ways in which a player's probability of choosing a given strategy might be altered by another player's choice. One is the case in which players move sequentially, and Player 2 can observe what Player 1 did. Another is a situation in which a player believes her choice to be evidence of what the other player's choice is going to be. We want to argue that in both cases sure thing reasoning applies, even if in the first case Independence (as we define it) is violated.

Here is an example (suggested to us by Harper) of what is considered a fallacious application of Savage's sure thing principle: Player 1 is deciding which of two offers to make to Player 2, who will then have the options to either accept or reject the offer. The extensive form may be represented as follows:



Player 1 can offer a good (*G*) or a bad (*B*) contract, and Player 2 can accept (*a*) or reject (*r*) the offer. If the states on which the outcomes of the acts depend are whether or not Player 2 accepts the offer, the decision matrix for Player 1 would be:

	a	*r*
G	2	0
B	3	0

If this were a correct formulation of the states, then Savage's sure thing principle would recommend offering the bad contract, *B*. For according to the sure thing principle (1954, 21–26), states on which the acts have the same outcome can be ignored so that preference between them goes by conditional preference given the states in which they differ. However, this application of sure thing reasoning would be fallacious. It ignores the fact that an offer of *G* would be accepted, whereas an offer of *B* would be rejected (as can be seen in the extensive form with perfect information). Evidential decision theory would represent this by having $p(a|G)$ close to 1 and $p(a|B)$ close to 0. According to our formulation of Independence, Player 2's choices are not causally independent of Player 1's choices, as Player 1's move does restrict Player 2's choice-set. Note however that Player 2 has four, not just two, strategies available, thus *a* and *r* are not the correct representations of the states on which the outcomes of Player 1's decision depends. If we give the correct representation of Player 2's strategies, then Savage's theory also avoids the fallacy. The correct strategic form for the above extensive form representation is:

	a	*r*	*a* if *G*/*r* if *B*	*r* if *G*/*a* if *B*
G	2, 2	0, 0	2, 2	0, 0
B	3, -1	0, 0	0, 0	3, -1

These are the correct Savage states, based on the structure built into the tree (Harper (1988) calls it the tree's "causal structure"). If we correctly specify the strategic form corresponding to the above extensive form, there is no sure thing fallacy. Player 1's fallacious sure thing argument for B has been rejected by adding the new states corresponding to the strategies for conditional choices available to Player 2. In the new matrix, sure thing reasoning applies to make "a if G/r if B" the unique rational choice for Player 2.

Let us now consider the second case, in which Player 1's choice is taken to be evidence of another player's choice, and where Independence (as we define it) holds. In the PD, choices are causally independent, in the sense that each player chooses in isolation, no previous communication is possible and each knows that his choice is not going to influence the way the other chooses. Causal independence, however, does not prevent a player from regarding his choice as evidence that the other will probably choose similarly. Now, evidential decision theory would recommend cooperation to any player who regarded his choice to cooperate as better evidence that the other will cooperate than the choice to defect would be. According to causal decision theory, the argument from such an assignment of probabilities to cooperation is a fallacy if the states are causally independent of the acts. One can have evidential relevance and causal independence together. Choosing C might count as evidence that the other player chooses C, but in no way influences his choice. Causal decision theory (CDT) argues that an agent's epistemic conditional probabilities are not always sufficient to represent the relevant beliefs about what his choices can or cannot influence. Following Stalnaker's idea of using subjunctive conditionals to represent deliberation, CDT considers the following conditionals in the PD game:

"If I were to do C, you would do C", ... and so on.

Agents in CDT assess unconditional probabilities of conditionals, that is, they assess the probability of conditionals such as "If I were to do C, you would do C". Player 1's belief that whether the other player does C or D is not influenced by Player 1's choice makes the probabilities of the conditionals equal the probabilities of their consequents, so that the causal expectation reduces to the Savage's expectation:

$U(C) = p$ (If I were to do C, you would do C) $u(C,C) + p$ (If I were to do C, you would do D) $u(C,D)$

which in turn reduces to:

$U(C) = p$ (you would play C) $u(C,C) + p$ (you would play D) $u(C,D)$

This supports the sure thing argument for D. According to CDT, the argument in favor of playing C because $p(C|C)$ is greater than $p(D|C)$ is akin to deciding to play C to bring about evidence that the desired outcome (C,C) obtains even though one knows that this can in no way help to bring it about. To reiterate the above point: if you believe that the other player's action is causally independent of your choice, then your evaluation of the probabilities of the conditionals will go by your evaluation of the probabilities of their consequents. Thus: p (If I were to do C, you would do C) $= p$ (you would do C) $= p$ (If I were to do D, you would do C).

[6] A strategy s for player i is strictly dominant iff, for all combinations of other players' strategies, the payoff of s is strictly better than the payoff of any other srategy s' available to i.

[7] Common belief may be defined in a way that is analogous to that for common knowledge, which has been characterized by Lewis (1969) and Aumann (1976).

[8] Outcome x Pareto-dominates outcome y just in case the payoff for each player in x is strictly better than the payoff for each player in y.

[9] See for example Binmore (1994, 204–205).

[10] See Hubin and Ross (1985) for an elaboration of the point.

[11] In a strategic-form game a pure strategy for player i is an action that i may choose. In an extensive-form game a pure strategy for player i is a function that assigns an action to each information set of player i. For a fuller account see Fudenberg and Tirole (1992, 4 and 83).

[12] In what follows we will employ a slightly modified definition of a model for a game provided by Stalnaker (1993).

[13] Note that a two-person strategic-form game is always represented by a matrix.

[14] R is a Euclidean relation iff for any x, y, z, if xRy and xRz then yRz. Transitivity and euclidity each correspond to reflection principles, namely that if a player believes that p, then she believes that she believes that p (transitivity), and if a player does not believe that p, then she believes that she does not believe that p (euclidity). Stipulating that R be transitive and Euclidean implies that a player's beliefs in any world w are the same as they are in any other world w' accessible to w.

[15] There are well known difficulties with such a "coarse grained" construal of propositions, but the idealization if adequate for our purposes, and one wishing to do so may replace propositions as we conceive them with more fine-grained entities.

[16] We use the expression 'doxastic' as opposed to 'epistemic' because we are considering beliefs and not knowledge.

[17] Note that we are referring here to so-called games of complete information. In such games, there is no uncertainty as to the other players' strategies or payoffs. If a player were uncertain, say, about the strategies available to the other players, they might not be playing the same game and in such a case it would be appropriate to assume that a player can falsely believe some outcome to be causally possible.

[18] On the present treatment of C, the converse is not true. If a proposition is causally impossible relative to a world w then it will be doxastically impossible for all players in that world as well.

[19] For more on this point see Stalnaker (1993). A player's belief is *about the structure of the game* just in case it is a belief she holds in every model of that game.

[20] This way of speaking may be unfamiliar, for there are some games for which the question of what it is rational to do is well posed independent of any model for that game. For instance, in a game that has a unique Pareto-optimal Nash equilibrium in dominant strategies, it is natural to say of a player to whom that equilibrium strategy is available that it is rational *simpliciter* for her to follow that strategy. Thus in the following game:

		Player 2	
		left	*right*
Player 1	*top*	2,2	3,3
	bottom	1,1	0,2

{*Top*,*Right*} is the unique Nash equilibrium in which the two players play dominant strategies, and it is also Pareto-optimal. It is obviously rational for each player in the above game to follow her dominant strategy in *all* models of this game.

[21] Robert Stalnaker has suggested to us that the doxastic version of the Identicality assumption might also be interpreted as the assumption that players "necessarily" believe that they choose alike. That is, in all possible worlds players believe that they choose alike. This assumption would be inconsistent with the Independence assumption, since the latter implies that there is a possible world w' in which one player chooses differently from the other player, and is aware of it. Consequently, this strengthened version of the doxastic form of the Identicality assumption is not one that we employ.

[22] The reasoning here resembles an argument considered, but not endorsed, in Nozick (1974, 131).

[23] Our contention – that even if a player believes that he and his opponent will cooperate, it does not follow that rationality precludes the consideration of an off-diagonal outcome – runs afoul of arguments presented by Schick (1979) and Levi (1992) aiming to show that once one has made up one's mind as to how to act, one no longer has any choice to do otherwise. Although consideration of these arguments would take us too far afield we would suggest that these arguments depend upon conflation of causal and epistemic modalities.

[24] The worlds in which Player 1 plays D that are most similar to the world in which both players

play *C* will keep the laws of nature constant. Because the Identicality assumption is not here being construed to be a law of nature, it need not be held constant in worlds similar to the actual one. Instead those features of Player 2's psychology that result (though perhaps not deterministically) in her playing *C* in the world in which both players play *C* should be held fixed in considering the world in which Player 1 plays *D*.

[25] That is, this is a model in which in the actual world w^*, for each player i, and each world w' such that w^*R_iw', there is no w'' where players do different things and $w'Cw''$. Each player might, for instance, believe that her action has a causal propensity to "trigger" the same kind of choice in the opponent. Whether such a belief is justified is a matter that goes beyond the present discussion.

[26] Analogous reasoning for an inconsistency in "perfect predictor" versions of the Newcomb problem may be found in Hubin and Ross (1985).

[27] See Skyrms (1984) for a list of seven of them.

[28] See Skyrms (1984) for a philosophical treatment of the phenomenon. In private conversations both W. Harper and A. Margalit reported having considered giving an agentive analogue to the EPR phenomenon in treating the so-called "paradox of the twins" in the prisoner's dilemma context.

[29] The tools for such a construction have, in fact, been provided in Belnap (1992).

[30] Sobel (1988) asks us to consider a version of the Newcomb problem according to which the predictor is *in principle* incapable of error. He argues that any agent who is aware of the infallibility of the predictor could not have a choice between taking one box and taking both. For Sobel holds that it is a necessary condition of the agent's having a decision problem that he be sure, of at least two actions, that he can do either of them; and he argues that this condition does not hold in the infallible predictor version of the Newcomb Problem. Even though the case that Sobel considers involves conceptual necessity whereas the EPRPD involves only causal necessity, one might wonder whether Sobel's reasoning, or an appropriate variation thereon, could undercut the intelligibility of the EPRPD. It might seem on the face of it that each player in the EPRPD only *appears* to be making a choice when in fact her hand is in some sense tied by the act of the other player. There is, however, a crucial difference between the EPRPD and the infallible predictor version of the Newcomb problem. In the latter there is a moment *prior to the moment of choice* (or putative choice), such that as of that moment there is only one action that the agent can perform. Further, the agent in the Newcomb problem is (or at least can come to be) aware of this fact. By contrast, in the EPRPD there is no such pre-choice moment. There is no moment prior to the moment of choice (or putative choice) as of which it is even causally necessary that, say, Player 1 do either *C* or *D*. For we have stipulated that the two point-events of choosing are spacelike separated. Because of this, each agent in the EPRPD can be sure that as of every moment leading up to that of his making his choice, he can either perform *C* or perform *D*. This, in turn, seems to be enough to show that each agent in this scenario can be sure that there will be two actions open to him when it comes time to choose between *C* and *D*.

[31] Note that if we were to cast the EPRPD in terms of causal decision theory, we would have to conclude that each player would maximize her causal expected utility by cooperating if her probabilities for conditionals of the form, "If I were to play *C*, the other player would play *C* as well," are high enough for each player's choice of *C* to be a ratifiable alternative in the sense of Harper (1988).

[32] In a symmetric game, the players have the same number of pure strategies, and the payoff to player i of (a_i,b_j) = payoff to j of (a_j,b_i). In other words, a symmetric game is one that looks the same no matter which player one is. In the PD, if both players do the same thing, they get the same payoff. If they choose differently from each other, the *D*-chooser gets 4 and the *C*-chooser gets 0. So the payoff is determined by what a player does, not by who she is.

[33] Other authors who have endorsed reasoning along similar lines are Gauthier (1974), Watkins (1972), and Hardin (1982).

[34] Pettit (1986) has shown the fallacy in his reconstruction of Davis' argument while leaving open the question whether every reconstruction of the argument is fallacious. We here consider another such version of the argument, perhaps closer to Davis' intent.

[35] Davis (1977, 325).

[36] *Ibid.*, p. 325.

[37] We are grateful to R. Stalnaker for suggesting this example to us in correspondence.

REFERENCES

Aumann, R.J. (1976), "Agreeing to Disagree," *Annals of Statistics* **4**, 1236–1239.

Belnap, N. (1992), "Branching Space-time," *Synthese* **92**, 385–434.

Belnap, N., and M. Green (1993), "Indeterminism and the Thin Red Line" in J. Tomberlin (ed.), *Philosophical Perspectives VII: Logic and the Philosophy of Language*. Atascadero, CA., Ridgeview, pp. 365–388.

Bicchieri, C. (1988a), "Strategic Behavior and Counterfactuals," *Synthese* **76**, 135–169.

Bicchieri, C. (1988b), "Common Knowledge and Backwards Induction: A Solution to the Paradox" in M. Vardi (ed.), *Theoretical Aspects of Reasoning about Knowledge*. Los Altos, CA., Morgan Kaufmann, pp. 381–393.

Bicchieri, C. (1993), *Rationality and Coordination*. Cambridge, U.K., Cambridge University Press.

Bicchieri, C. and M. Dalla Chiara (eds.) (1992), *Knowledge, Belief, and Strategic Interaction*. Cambridge, U.K., Cambridge University Press.

Binmore, K. (1994), *Playing Fair: Game Theory and the Social Contract*, Vol. I. Cambridge, MA., M.I.T.

Brams, S. (1975), "Newcomb's Problem and Prisoners' Dilemma," *Journal of Conflict Resolution* **19**, 597–612.

Campbell, R.K. (1989), "The Prisoner's Dilemma and the Symmetry Argument for Cooperation," *Analysis* **49**, 60–65.

Campbell, R. and L. Sowden (eds.) (1985), *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*. Vancouver, University of British Columbia Press.

Davis, L. (1977), "Prisoners, Paradox, and Rationality," *American Philosophical Quarterly* **14**, no. 4, 319–327.

Davis, L. (1985), "Is the Symmetry Argument Valid?" in Campbell and Sowden (1985), pp. 255–262.

Fudenberg, D. and J. Tirole (1992), *Game Theory*. Cambridge, MA., M.I.T. Press.

Gauthier, D. (1974), "The Impossibility of Rational Egoism," *Journal of Philosophy* **71**.

Gibbard, A. and W. Harper (1978), "Counterfactuals and two kinds of expected utility" in *Foundations and Applications of Decision Theory*. Dordrecht, D. Reidel, pp. 125–162.

Hardin, R. (1982), *Collective Action*. Johns Hopkins University Press.

Harper, W. (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 in Decision, Belief Change, and Statistics", Dordrecht, Kluwer Academic Publishers.

Hubin, D. and G. Ross (1985), "Newcomb's Perfect Predictor," *Nous* **19**, 439–447.

Jeffrey, R. (1965, 1983), *The Logic of Decision*. Chicago, University of Chicago Press.

Kaneko, M. (1987), "Structural Common Knowledge and Factual Common Knowledge," *RUEE Working Paper* no. 87–27, Hitotsubashi University.

Kaneko, M. and T. Nagashima (1990), "Game Logic I: Players' Deductions and Knowledge of Deductive Abilities," E90-3-1, VPI-SU.

Körner, S. (ed.) (1974), *Practical Reason*. New Haven.

Leslie, J. (1991), "Ensuring Two Bird Deaths With One Throw," *Mind* **100**, 73–86.

Levi, I. (1992), "Feasibility" in Bicchieri and Dalla Chiara (1992), pp. 1–20.

Lewis, D. (1969), *Convention*. Cambridge, MA., Harvard University Press.

Lewis, D. (1979), "Prisoners' Dilemma is a Newcomb's Problem," *Philosophy and Public Affairs* **8**.

Nozick, R. (1974), "Newcomb's Problem and Two Principles of Choice" in N. Rescher *et al.*, *Essays in Honor of Carl G. Hempel*. Dordrecht, D. Reidel, pp. 114–146.

Pettit, P. (1986), "Preserving the Prisoner's Dilemma," *Synthese* **86**, 181–184.

Rapoport, A. (1966), *Two-Person Game Theory*. Ann Arbor, University of Michigan Press.

Savage, L.J. (1954), *The Foundations of Statistics*. New York, John Wiley and Son.

Schick, F. (1979), "Self-knowledge, Uncertainty, and Choice," *British Journal for the Philosophy of Science* **30**, 235–252.

Schlesinger, G. (1974), "The Unpredictability of Free Choice," *British Journal for the Philosophy of Science* **25**, 209–221.

Skyrms, B. (1984), "EPR: Lessons for Metaphysics" in French, Uehling, and Wettstein (eds.), *Midwest Studies in Philosophy IX: Causation and Causal Theories*. Minneapolis, University of Minnesota Press, pp. 245–255.

Sobel, J.H. (1988), "Infallible Predictors," *Philosophical Review* **97**, 3–24. "Dilemma," *Synthese* **55**, 347–352.

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

Stalnaker, R. (forthcoming), "Knowledge, Belief, and Counterfactual Reasoning in Games," in C. Bicchieri, R. Jeffrey and B. Skyrms (eds.), *The Logic of Strategy*. New York, Oxford University Press.

von Neumann, J. and O. Morgenstern (1944), *Theory of Games and Economic Behavior*. Princeton, NJ., Princeton University Press.

Watkins, J. (1972), "Imperfect Rationality," in R. Borger and F. Cioffi (eds.), *Explanation in the Behavioral Sciences*. Cambridge, Cambridge University Press.