Gilboa, Itzhak

**Working Paper**

# Can Free Choice Be Known?

Discussion Paper, No. 1055

**Provided in Cooperation with:**
Kellogg School of Management - Center for Mathematical Studies in Economics and Management Science, Northwestern University

*Suggested Citation:* Gilboa, Itzhak (1993) : Can Free Choice Be Known?, Discussion Paper, No. 1055, Northwestern University, Kellogg School of Management, Center for Mathematical Studies in Economics and Management Science, Evanston, IL

This Version is available at:
https://hdl.handle.net/10419/221412

Discussion Paper No. 1055

# CAN FREE CHOICE BE KNOWN?*

by

Itzhak Gilboa**

July 1993

---

*This note was written in November 1991 as a draft. It was supposed to evolve into a joint work with Cristina Bicchieri. However, as time went by it became clear that very little agreement exists between the co-authors. We therefore decided to split. Cristina will shortly complete a separate paper on a related subject. I am grateful to Andy Fano, Eva Gilboa, Shlomo Gilboa, Eric Jones, Ehud Kalai, and Eitan Zemel for the discussions that motivated this work, as well as for comments and references.

**Department of Managerial Economics and Decision Sciences, J.L. Kellogg Graduate School of Management, Northwestern University, Evanston, Illinois 60208.

Abstract

In this note we reconsider an argument, borrowed from causal decision theory, according to which rational and identical players should cooperate in a one-shot prisoner's dilemma. We argue that, regardless of how one views this type of reasoning, the example points at a possible inconsistency in standard formulations of knowledge and decision. We suggest that when formalizing notions of "decision," "choice," and "rationality," care must be taken not to assume knowledge of one's own choice. Finally, the relationships to the classical problems of causal decision theory and of determinism versus free will are briefly discussed.

## 1. Introduction

The last decade has witnessed a proliferation of attempts to formalize the notions of knowledge, choice and rationality, especially in interactive situations. Two, admittedly related, types of problems arise in this context: the substantial philosophical questions which have a couple of millennia of experience in tormenting innocent minds, and the more modern formal questions. Among the former, the essence of rationality, choice and knowledge may lead one to causal decision theory, self-referential paradoxes and the age-old problem of determinism and free will. Among the latter, it seems that even when we have a very clear idea of some notion of rationality, a satisfactory formulation of it may be extremely evasive. At times, the formulation problem would indeed be rooted in a much more fundamental philosophical quandary.

The main goal of this paper is to address a question of the second type. We highlight a problem in the formal modeling of rational choice, and suggest a way to deal with it. The discussion seems to lead one inevitably to dangerous areas shadowed by ominous problems of "the first type." On our retreat from these areas to a safe haven, we will naturally watch the ogres carefully and relate their size and ghastliness as seen from our viewpoint, though we will not tackle them directly.

We start by presenting a well-known argument, according to which rational agents who play a one-shot prisoner's dilemma should choose to cooperate if they know they are identical. We then argue that, regardless of how convincing this type of reasoning is, it cannot be dismissed without refining our definition of rational choice. We propose a solution which consists of two main points. The first, more technical one, involves the

"synchronization" of proof steps. The second point is that one should ignore any self-knowledge of free choice when attempting to model the latter. After discussing this solution in general, we briefly relate it to causal decision theory and to the problem of determinism versus free will.

## 2. Motivating Example

Consider the classical "prisoner's dilemma" game:

| | | Player 2 C | Player 2 D |
|---|---|---|---|
| Player 1 | C | (3,3) | (0,4) |
| | D | (4,0) | (1,1) |

As customary, player 1 chooses a row and player 2 chooses a column. The choices can be thought of as simultaneous. The payoff matrix above gives, for each pair of choices of the two players, the utility of player 1 and that of player 2, respectively. These utility functions are assumed to be derived from past observed choices, so that the model implicitly assumes that, in a one-person decision problem under certainty, each player would maximize his/her utility by definition. (In general, the utilities are taken to be von Neuman-Morgenstern utilities (1944), which also reflect each player's decision under a situation of risk, but this additional assumption is not needed here.)

It is implicitly assumed that the game is played exactly once by rational agents. Furthermore, the game and the players' rationality are assumed to be common knowledge between the players. (For the definition of

common knowledge, see Lewis (1969) and Aumann (1976). The notion of rationality will remain vague for the time being and, alas, probably also at the end of this paper.) By the behavioral definition of the utility function, it already incorporates all psychological and sociological payoffs, altruism considerations and so forth.

We will not discuss the prisoner's dilemma at length here. (See, for instance, Aumann (1987b).) Let us only point out that, if one of the players, say player 1, had only one possible choice, player 2 would choose D by definition of the utility function. This claim implicitly assumes that the mere existence of player 1 does not change player 2's behavior, that is, she still views the decision problem as a single-person one. However, there is nothing in the definition of the utility function above, nor in that of vNM's utility function, which implies that player 2 would choose D in this game, where player 1 is not a "dummy" player. (This would follow, however, from the stronger assumptions used in subjective probability models as Savage (1954) and Anscombe-Aumann (1963).) Yet the non-cooperative choice (D) does follow from a very weak assumption of rationality, namely that a player would choose a strictly dominating strategy if such exists. (D is such a strategy since for every choice of the other player it guarantees a strictly higher utility level.)

Using this additional assumption one therefore concludes that both players would choose the non-cooperative outcome. (This has been a source of puzzlement and agony to the profession since (C,C) Pareto dominates (D,D), i.e., both players would have been better off had they decided to play C.)

In particular, one may impose an additional assumption at the outset, stating that the players are identical in some sense. The game is symmetric,

and the players may follow identical reasoning to conclude they should play D, i.e., make identical choices.

However, with the same assumptions one may suggest the following reasoning: "Suppose I am player 1. Knowing that player 2 is identical to me, she would end up making the same choice. That is, if I play C, so does she, and the same applies to D. To be precise, 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 using my information I conclude that C dominates D. To be precise, you could deduce that I will choose C over D by the mere definition of my utility function."

Thus, the argument goes, if identical rational players, who are aware of their identicality, play the prisoner's dilemma, they will choose to cooperate (C).

## 3. Modeling Free Choice

### 3.1 A Paradox

Most game theorists would reject the above reasoning as nonsensical, ignoring the independence of different players' choices.[1] They would argue that it actually corresponds to the following one-player game, to which we will refer as "the identical prisoners' choice":

---

[1]That is a result of an informal small survey. In fact, most first-year graduate students are already enlightened/indoctrinated enough to reach the same conclusion.

| | | |
|---|---|---|
| Player 1 | C | 3 |
| | D | 1 |

and which already incorporates the identicality of the two players, but it does not make any sense in the original game, even if the players in it end up making identical choices.

Further, they would continue, the fact that the players are identical does not mean that in their reasoning process they cannot conceive of situations in which they differ. Thus, the off-diagonal outcomes are conceivable, even if not possible, and, indeed, they are the distinction between the two games above. Moreover, in the process of rational decision making one has to consider the conceivable states of the world, even if the decision theory will end up ruling them out as "impossible." (See Gilboa (1990), Gilboa (1991).) Finally, rationality in the context of independent decision makers has to be applied in such a way, that a player compares states of nature (possible or merely conceivable) which differ only in his/her own decision.

While we agree with all these claims, we find that they fail to pinpoint the flaw in the argument of Section 2 above. To be more specific, consider some formal model of the game describing players' knowledge and choices. (See, for example, Kaneko (1987), Bicchieri (1988a, 1988b 1989, 1992), and Kaneko-Nagashima (1989).) Suppose that in such a model the following assumption of rationality is assumed to be common knowledge: if a player is faced with finitely many conceivable choices, and he/she knows the payoff

derived from each of them, then the player would choose an act which maximizes the payoff function. (This was dubbed the "common sense" assumption in Gilboa-Schmeidler (1988).)

Notice that this assumption does not follow directly from the definition of the utility function since it involves the player's knowledge, which will turn out to be a crucial issue. Note also that, as formulated, this axiom does not imply the choice of a dominant strategy in the presence of uncertainty, say, about other players' choices. We may either strengthen it to say that an action be chosen by a player whenever it is known (by him/her) to guarantee a higher payoff than all other actions, whatever other players may do, or we may additionally impose a separate axiom stating that dominant actions be chosen. The second option does not allow for iterative elimination of dominated strategies, but it simplifies exposition. Let us therefore refer to the conjunction of the "common sense" axiom and the "domination" axiom as "the rationality assumption."

For clarity of exposition, let us further impose the identicality assumption (we will later see that it is actually redundant). Its exact meaning may be problematic: to say that the players are identical typically means that for a certain class of predicates $\Psi$, $\psi$ (player 1) holds if and only if $\psi$ (player 2) does, for all $\psi$ in $\Psi$. Obviously, this may be hard to formulate in first-order-logic models. However, for our purposes, "identicality" may simply mean that "player 1 chooses C iff player 2 chooses C."

As explained above, there seems to be no reason for classical game theorists to reject this assumption in the prisoner's dilemma case, since it has a unique dominant strategy--indeed, the same one--for both players.

Yet, equipped with this assumption, we can now follow the reasoning of Section 2 in a formal proof: if identicality is known to both players, each knows that "I play C implies the other player plays C," and the rationality assumption concludes that each player would choose C. On the other hand, it also implies that each player would choose D (by the domination argument). In other words, the model is inconsistent even though the identicality assumption itself does not contradict the implication of the other assumptions (namely, the conclusion that the players would play D.)

### 3.2 Synchronization of the Proof

The reader may claim that the identicality assumption is unwarranted. Indeed, one may argue that there seems to be no compelling reason to believe two (or more) players will choose the same strategies even if the game is symmetric, unless this conclusion is derived from more primitive assumptions. (Consider, for instance, a game in which both players have two equivalent strategies which dominate all others.) According to this view, the assumption "If I play C, the other player does the same" may be imposed at the outset only if there is some direct ("causal") relationship between the players' choices. For instance, assume that player 2 is the image player 1 sees in a mirror. In this case, player 1 actually faces the "identical prisoners' choice," a game in which he will play C both by the common sense and by the domination axioms. Differently put, the very formulation of the game (in its normal form) presupposes that no conditional statement, "If A then B" may be imposed where A and B relate to different players' choices.

In the next sub-section we provide another argument by which the identicality assumption should be rejected. However, in the prisoners'

dilemma case, the identicality assumption is not necessary for an inconsistency to occur: since rationality implies a unique choice for each player, one can prove that the players' choice will be identical. To be precise, after having proven that each player will play D, both statements "If I play C, the other player will play C" and "If I play D, the other player will play D" would be logically correct for and known by both players. But then they may be used, in conjunction with the rationality assumption, to derive the contradiction.[2]

There seems to be a problem with the synchronization of the proof in the above reasoning. After the choice C was ruled out at some stage, vacuously true statements beginning with "If I play C" come back to haunt us and derive various unwarranted conclusions.

At this point the similarity to the backward induction problem is rather evident. It has been commonly believed for quite some time that if rationality is common knowledge (whatever that may mean), a backward induction solution is the only consistent one in finite extensive form games with perfect information. However, this belief has been challenged. Reny (1988) pointed out that common knowledge of rationality cannot hold at all nodes of the game tree, and Bicchieri (1988a, 1988b, 1989) further argued that common knowledge (of the game and of all behavioral axioms) and rationality are incompatible axioms.

Gilboa (1990) and Ben-Porath (1993) attempted to provide formulations of these axioms that will be consistent (at least as long as the backward induction solution is actually played.) The solution proposed by Gilboa

---

[2]As noted by Ehud Kalai, a similar inconsistency may arise also in a one-person decision problem.

(1990) deals with "synchronization" of the proof in such a way that proven impossibilities will not be allowed to be used again in later derivation.

It seems that a similar idea may be adopted here, and it would roughly run as follows: we first introduce a set of "possible" states of the world as a formal entity of the model.[3] The rationality assumption is qualified to hold only for a player's actions which are "possible," and it is used to conclude that some of them are "impossible." Put differently, the axioms are viewed as an operator, ascribing to each set of states of the world a subset thereof, with the interpretation that if the argument of the operator is known to include all possible states, so does its value.

A set is said to be "possible" if it is a fixed-point of this operator, and there exists a chain of applications of the operator, starting with the set of all states and the world and leading to it. (Gilboa (1990) and, in a different formulation, Ben-Porath (1993) note that the fixed-point requirement alone will not do for derivation of the backward induction solution. It seems that dropping the "chain condition" would result in unintuitive conclusions in our context as well.)

Note that for all finite games there exists a unique possible set according to this definition. However, it may well be empty if the behavioral axioms happen to be inconsistent.

In a way, then, this solution concept imitates the reasoning of the players or of an outside observer. Following this solution in our case, one first has to conclude that choosing C is impossible for both players, whence the identicality assumption follows. But then one cannot use it to show that

---

[3]The distinction between "conceivable" and "possible" states of the world was earlier suggested by Hintikka (1975).

rationality implies the choice C, because the rationality assumption has no bite after C was dubbed "impossible."

As herein described (and in Gilboa (1990)), this "solution concept" is imposed from outside the model, and only implicitly may one assume that the players indeed follow the reasoning implied by the chain of operator applications. One may well wish to formally include this "proof" in the model, using a three-place predicate such as "Player i considers state $\omega$ possible at step t" where the steps are a well-ordered set (say **N**) corresponding to steps of the proof. Then one may formulate the claim that all players start out by considering all conceivable states of the world, and the rationality assumption would allow us to conclude that some states, previously deemed possible, will no longer be so.

There is one rub, though: in this case one also needs an axiom stating that whatever was deemed possible would continue being so unless otherwise proven. (This is close to the "Frame Axioms" in the artificial intelligence literature.) That is, something along the lines of "For all i, $\omega$ and t, if i does not have a t-stage proof that $\omega$ is not possible, then i considers $\omega$ as possible at stage t + 1." It is not clear to us at this point whether there are satisfactory formulations of such a system which are also consistent.

At any rate, if we avoid this problem and leave the "proof" steps in the definition of the solution concept, we obtain a consistent model which seems to capture our basic game-theoretic intuition.

### 3.3 Two Kinds of Knowledge

We can therefore hope that there is a way to resolve the paradox, at least as long as the identicality assumption is ruled out. Notice, however,

that the "synchronization" solution does not suffice in the presence of the identicality assumption, since in this case the reasoning leading to the cooperative outcome may be carried out immediately, and it need not "wait" for the classical reasoning to first conclude that the non-cooperative solution will be played.

However, we will argue that the identicality assumption cannot be known by the players to begin with, for reasons that have nothing to do with the paradox discussed above. Ruling the identicality assumption out on theoretical grounds, coupled with the "synchronization" argument of the previous sub-section, will complete the resolution of the paradox.

Let us first analyze a simpler example. Consider the following one-person decision problem: Sir Isaac Newton is standing in a room, consdering the possibility of jumping out of the window. If he decides to do so, there are two possibilities (or should we write "conceivabilities"?). He may hover in the air, possibly enjoying the view, or he may fall down to the ground. The latter outcome is assumed highly undesirable.

Now let us assume Newton knows two facts about the world: that the law of gravity holds and that he is rational. The first implies that he may practically ignore the possibility of hovering in the air. The second one, by the same token, means that he is not going to jump out of the window.

This decision is certainly made in accordance with Newton's free will. Indeed, the rationality assumption attempts to model precisely these choices that are intuitively referred to as "free." Yet there is something rather troublesome in this formulation: if he "knows" that he is rational in the same way he "knows" that gravitation works, then prior to making up his mind Newton will rule out the possibility of jumping out of the window just as he

rules out the possibility of hovering in the air if he does jump. In this case, then, the preliminary analysis he conducts in order to make a "rational" decision leaves him with no choice whatsoever.

The problem here is not logical consistency. Even a "naive" formulation of the axiom of rationality does not seem to lead to any inconsistencies in this simple decision problem. However, we feel that this could hardly be counted as an intuitive modeling of "free choice." Treating one's knowledge of the outside world (which may include other players as well) and one's knowledge of oneself in the same way (formally) does not seem to correspond to our first-hand experience of "making a choice."

Thus, regardless of the problem discussed in subsection 3.1, we are tempted to suggest the following principle: when analyzing a (possibly interactive) decision problem from the subjective viewpoint of one individual, one should ignore that individual's knowledge regarding him/herself. To be precise, one may assume that the individual "knows" his/her tastes and beliefs in the same sense (s)he "knows" rules of nature, but does not "know" any behavioral assumptions that (s)he may satisfy. (Note that we use here the term "taste" rather then "utility." For clarity of exposition, we refer to one's tastes derived from direct introspection and <u>not</u> to utility which is a theoretical construct one uses to explain and predict other players' observed behavior. However, these tastes can still be behaviorally defined, as long as they relate to past choices and not to the choice which is about to be made. To be precise, there is no distinction in this respect between a player's reasoning about others and about oneself. The distinction is that regarding others one may "believe" or even "know" that they make choices in a consistent way with past ones, so that their utility functions may be used for

prediction. As for one's own self, however, no similar conistency axiom can be known or believed. We will come back to this point in Section 4 below.)

Formally speaking, an indiviudal may "know" conditional statements of the type "If I choose. . .then. . ." and, indeed, these statements are essential for the formalization of rational choice. But then (s)he may not have any non-trivial belief, let alone knowledge, regarding the conclusion of such a statement.[4] In particular, the latter may not be a strict subset of other players' (or nature's) normal form strategies.

With this interpretation in mind, Newton's decision not to jump out of the window is the result of his free choice, which relies on his knowledge of gravity as well as of his preference not to be smashed on the ground. Yet neither while making this decision nor afterwards does he "know" that he satisfies the rationality assumption in general or that he always follows his preferences. He may know that his beliefs satisfy certain axioms (such as the system S5), and he may know all about his tastes now or in the future. But he never "knows" anything about a choice while making it.

A similar interpretation problem was raised and discussed in Aumann (1987a) regarding the assumption that a state of the world specifies all players' actual choices. Our solution is similar to his. Dekel (1990) and Gilboa (1991) further elaborate on this point in the same spirit.

Considering the paradox of subsection 3.1 again, we first note that the "knowledge" referred to in the rationality assumption should be more carefully defined as the player's knowledge about all but his/her choices. (Other players' "knowledge" about our player are to be interpreted as their beliefs,

---

[4]By "non-trivial" we would like to exclude statements such as "If he chooses not to jump, 1 = 1," which may be known simply because "1 = 1" is known.

which may be known to the player.)

Thus, the identicality assumption, even if incorporated in the model, cannot be used in the proof in conjunction with the rationality assumption: a player cannot know, nor use in any further reasoning, the identicality assumption since it involves the player's own choices and relates them to events about which (s)he may have beliefs.

There are several ways to formulate this intuition. One may introduce, for each player, two knowledge designators (which may be characters, operators on propositions, operators on sets, predicates, or whatever entity they are in a formal model of knowledge): one of them will refer to the "outside" world, which, broadly understood, includes not only other players and "nature," but also results of introspection such as tastes and beliefs. The other knowledge designator will refer to an agent's knowledge of his/her own behavior. Alternatively, one may parameterize the knowledge operator by time and allow the agent to know his/her choices only after they were actually made. Yet another solution is to keep only one knowledge designator but to assume it does not apply to propositions involving the player's choices, where these are given by a formal choice function. (As opposed to a player's utility, which may be known by the player--say, interpreted as taste--and by others--interpreted as observed behavior--the player's actual choice in each state of the world can only be known by others.)

While the precise way one chooses to formalize this distinction may not be of paramount importance, it seems that some formal distinction between these two types of knowledge is needed for consistency and for more intutitive modeling of "free choice."

4. Causal Decision Theory

Causal decision theory distinguishes between "two kinds" of expected utility (see Gibbard and Harper (1978)). One would assume independence of beliefs from choices, while the other allows for a decision maker to have beliefs regarding his/her own choices and update beliefs regarding states of the world given a certain choice. Thus the decision maker may, indirectly, choose what to believe.

In the spirit of the above discussion, we find the second "kind" of expected utility problematic. This is so not only because it robs decision theory of one of its most cherished assets, namely, the theoretical dichotomy between states of the world (which cannot be controlled) and choices (regarding which there are no beliefs), but also because a choice about which the chooser has beliefs does not seem to comply with our intuition of "free" choice.

We see no problem with a decision maker who has beliefs regarding his/her own tastes, or "action tendencies," as well as states of the world and updates the beliefs regarding the latter given some information about the former. (This approach is discussed in Lewis (1981), but it is there rejected as a substitute to learning-from-one's-own-choices.) Indeed, Savage's (1954) classical approach would incorporate the uncertainty about one's tastes into that about the state of the world to begin with. However, we are generally happier with a model in which one cannot be said to have beliefs (let alone knowledge) of one's own choice while making this choice.

One may legitimately ask: Can you truly claim you have no beliefs about your own future choices? Can you honestly contend you do not believe--or even know--that you will not choose to jump out of the window? And is such a model

really more intuitive?

Indeed, the answer to these questions is probably a resounding (or reluctantly muttered) "No." But the emphasis should be on the timing: when one considers one's choice tomorrow, one may indeed be quite sure that one will not decide to jump out of the window. However, a future decision should actually be viewed as a decision by a different "agent" of the same decision maker. One's beliefs about one's behavior tomorrow probably derive from knowledge of one's past choices as well as past and present introspection. These, however, cannot rule out the possibility of evolving into a different person, as in the changing preference literature. Almost all authors who addressed these problems, of which the classical one (with "constant" preferences) are but a special case, seem to agree that the best way to conceptualize them is by considering different "agents" as separate players. (See Peleg-Yaari (1973), Hammond (1976), and Ferreira-Gilboa-Maschler (1992), which also includes additional references.)

Thus, when asked about Newton's suicidal choice, he is not actually making the choice. (Perhaps he makes a different choice, namely, which answer to give.) He may therefore have beliefs about the future choice without leading to any inconsistency, nor undermining the notion of "free choice." It is only at the time of choice, within an "atom of decision," that we wish to preclude beliefs about it.

In the same vein, one may relate the introspection into tastes and beliefs to the behavioral approach. That is to say, their separation from the actual decision need not necessarily leave them at the mercy of direct intuition or subject them to the risk of being considered "metaphysical nonsense" by the decision maker. The latter may ask him/herself hypothetical

questions about a variety of choice situations. As long as these are not the actual choices being made at present, our approach allows the individual to have very strong intuition regarding such choices. If these turn out to satisfy, say, the Savage axioms, the individual may elicit his/her own tastes and beliefs.

When introspection is considered, one may assume that hypothetical choices are the primitives about which a decision maker has intuition, and from which tastes and beliefs are derived, or vice versa, or even claim that the two are inextricably interrelated. All of these views (and probably many others as well) are consistent with our view of "free choice" as long as no beliefs about a choice are assumed to exist while it is actually made.

## 5. Determinism and Free Will

It would seem an outrageous cowardice to get that close to the problem of determinism versus free will and avoid even glancing at it. Let us therefore try to look at it from our viewpoint for what it is worth.

The traditional problem arises as a conflict between the undeniable intuition people have of making choices freely, and the belief that this choice is known or could in principle be known. The latter may arise from belief in a superbeing, or from monism coupled with the assumption that physical laws completely determine the future given present data, or from any other beliefs.

The common feature seems to be, therefore, that an individual, while making a seemingly "free" choice, also knows that someone else knows or could in principle know what this choice will end up being. Thus formulated, it is, indeed, very similar (though not identical) to the problem discussion in

Section 3.

True to our arguments above we should therefore contend that when modeling "free will" in a careful but intuitive way, one should make sure not to assume any knowledge, or knowledge of knowledge, of one's choice.

There is, of course, no problem if one is happy with this assumption. Indeed, the problem discussed in Section 3, which dealt only with the modeling of choice, ends here. However, the problem does not seem to have changed if the above has to be reconciled with the intuitive feeling that the choice is known. (We here interpret "predetermined" as "known" or "knowable" by some entity.)

While no resolution of this contradiction is to be expected from this paper, let us just carry the reformulation of the problem one step further. It can therefore be argued that "free will" or "free choice" refer to this very well-known feeling one has while making a choice, and having the clear impression that the choice is not knowable. If one generally believes that impression is false, one may treat this it as an illusion, maybe similar to the phenomenon of deja-vu. (Admittedly, the free will illusion happens more frequently than the deja-vu one and, alas, has weightier implications.)

Obviously, many other suggested reconciliations of determinism and free will are consistent with our notion of "free choice." For instance, one may argue that as long as the choice is knowable only in principle, rather than in actuality, it does not contradict free choice. Thus, as long as Newton does not know whether he is going to jump out of the window, his "free" choice is nicely modeled even if he realizes that "in principle" it could be deduced from data which are "in principle" available. (The qualification "in principle," however modeled, may suffice to distinguish the knowledge of his

choice from the knowledge that gravity works.) Similarly, if one believes that one's choice can only be known by a superbeing and not by oneself and, in particular, one will never know what exactly does the superbeing know of one's choice (apart from the very fact it knows it)--there may already be a qualitative distinction between knowledge of choice and of other facts so as to satisfactorily model "free choice."

Needless to say (partly because this has already been said too often), this paper does not attempt to suggest a resolution to the age-old problem of determinism and free will, nor even to make a claim regarding the existence of such a solution. We do hope, however, that a clearer understanding of what is meant by "free choice" may help thinking about this problem, which people keep addressing, either because they choose to or because they are pre-ordained to do so (or possibly both).

## References

Anscombe, F. J. and R. J. Aumann (1963), "A Definition of Subjective Probability," Annals of Mathematical Statistics, 34, 199-205.

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

Aumann, R. J. (1987a), "Correlated Equilibrium as an Expression of Bayesian Rationality," Econometrica, 55, 1-18.

Aumann, R. J. (1987b), "Game Theory," in The New Palgrave: Game Theory, J. Eatwell, M. Milgate and P. Newman (eds.), W. W. Norton: New York, London, pp. 1-53.

Ben-Porath, E. (1993), "Rationality, Nash Equilibrium, and Backward Induction in Perfect Information Games," mimeo.

Bicchieri, E. (1988a), "Strategic Behavior and Counterfactuals," Synthese, 76, 135-169.

Bicchieri, C. (1988b), "Common Knowledge and Backward Induction: A Solution to the Paradox," in the Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge, M. Vardi (ed.), Morgan-Kaufmann, 381-393.

Bicchieri, C. (1989), "Self-Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge," Erkenntnis, 30, 69-85.

Bicchieri, C. (1992), Rationality and Coordination, Cambridge University Press, forthcoming.

Dekel, E. (1990), "Discussion of 'Foundations of Game Theory' by Kenneth J. Binmore; 'Switching Away from Probability One Beliefs' by Georg Noldeke and Eric van Damme; and 'Global Games and Equilibrium Selection' by Hans

Carlsson and Eric van Damme," mimeo.

Ferreira, J.-L., I. Gilboa and M. Maschler (1992), "Changing Preferences and Credible Equilibria" (temporary title), mimeo.

Gibbard, A. and W. L. Harper (1978), "Counterfactuals and Two Kinds of Expected Utility," Foundations and Applications of Decision Theory, 1, 125-162.

Gilboa, I. (1990), "A Note on the Consistency of Game Theory," in the Proceedings of the Third Conference on Theoretical Aspects of Reasoning About Knowledge, R. Parikh (ed.), Morgan-Kaufmann, 201-208.

Gilboa, I. (1991), "Rationality and Ascriptive Science," mimeo.

Gilboa, I. and D. Schmeidler (1988), "Information Dependent Games: Can Common Sense Be Common Knowledge?", Economic Letters, 27, 215-221.

Hammond, P. J. (1976), "Changing Tastes and Coherent Dynamic Choice," Review of Economic Studies, 43, 159-173.

Hintikka (1975), "

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 the Common Knowledge of Deductive Abilities," Virginia Polytechnic Institute Working Paper No. E90-03-1.

Lewis, D. (1969), Conventions: A Philosophical Study, Cambridge: Cambridge University Press.

Lewis, D. (1981), "Causal Decision Theory," Australian Journal of Philosophy, 51, 5-30.

Peleg, B. and M. E. Yaari (1973), "On the Existence of a Consistent Course of Action when Tastes are Changing," Review of Economic Studies, 40, 391-

401.

Reny, P. (1988), "Rationality, Common Knowledge and the Theory of Games," mimeo.

Savage, L. J. (1954), The Foundations of Statistics, New York: Wiley.

von Neumann, I. and O. Morgenstern (1944), Theory of Games and Economic Behavior, Princeton: Princeton University Press.