PHILOSOPHY OF GAME THEORY

Till Grüne-Yanoff and Aki Lehtinen

1 INTRODUCTION

Consider the following situation: when two hunters set out to hunt a stag and lose track of each other in the process, each hunter has to make a decision. Either she continues according to plan, hoping that her partner does likewise (because she cannot bag a deer on her own), and together they catch the deer; or she goes for a hare instead, securing a prey that does not require her partner's cooperation, and thus abandoning the common plan. Each hunter prefers a deer shared between them to a hare for herself alone. But if she decides to hunt for deer, she faces the possibility that her partner abandons her, leaving her without deer or hare. So, what should she do? And, what will she do?

Situations like this, where the outcome of an agent's action depends on the actions of all the other agents involved, are called *interactive.* Two people playing chess is the archetypical example of an interactive situation, but so are elections, wage bargaining, market transactions, arms races, international negotiations, and many more. Game theory studies these interactive situations. Its fundamental idea is that an agent in an interactive decision should and does take into account the deliberations of the other players involved, who, in turn, take her deliberations into account. A rational agent in an interactive situation should therefore not ask: "what should I do, given what is likely to happen?" but rather: "what will they do, given their beliefs about what I will do; and how should I respond to that?"

In this article, we discuss philosophical issues arising from game theory. We can only sketch the basic concepts of the theory in order to discuss some of their philosophical implications and problems. We will thus assume that our readers have some familiarity with the basic concepts. For those who are primarily looking for an introduction to the basics of game theory, we recommend Binmore [2007; 2008] or Kreps [1990], both of which also consider philosophical issues. Osborne and Rubinstein [1994] and Fudenberg and Tirole [1991] are textbooks that put more emphasis on the mathematical proofs. Hargreaves-Heap & Varoufakis [2001], Ross [2006b] and Grüne-Yanoff [2008b] provide philosophical accounts of game theory.[1]

Philosophy and game theory are connected in multiple ways. Game theory has been used as a tool in philosophical discussions, and some crucial game theoretical

[1] This paper is based on Grüne-Yanoff's earlier paper.

Handbook of the Philosophy of Science. Volume 13: Philosophy of Economics.
Volume editor: Uskali Mäki. General editors: Dov M. Gabbay, Paul Thagard and John Woods.

concepts have been developed by philosophers.[2] Game theory also has been the object of philosophical inquiry itself. Our discussion will concentrate on the latter. Since game theory relies heavily on mathematical models, the standard epistemic issues concerning modelling and unrealistic assumptions in philosophy of economics are also relevant for game theory. But since game theory transcends economics, a number of other philosophical issues also arise. Perhaps the most important of these is the interpretation of the theory: is game theory to be understood mainly as a tool for *recommending* rational choices, for *predicting* agents' behaviour, or for merely providing an abstract *framework for understanding* complex interactions (e.g., [Blackburn, 1998; Aydinonat, 2008])? If we settle for the first interpretation, the issue of whether the rationality concept employed by the theory is justifiable becomes pressing. Is it intuitively rational to choose as the theory prescribes? If the second interpretation is adopted, one must ask whether the theory can in principle be a good predictive theory of human behaviour: whether it has empirical content, whether it is testable and whether there are good reasons to believe that it is true or false. If the third interpretation is adopted, the question arises concerning which qualities of the theory contribute to this understanding, and to what extent these qualities are different from the prescriptive or predictive function discussed in the first two interpretations.

We will address this central question in sections 3 and 4. In order to do so, a number of game-theoretical concepts that are particularly important in a philosophical assessment must be discussed first, viz. payoffs, strategies, and solution concepts.

2 SOME BASIC CONCEPTS

Decision theory, as well as game theory, assesses the rationality of decisions in the light of preferences over outcomes and beliefs about the likelihood of these outcomes. The basic difference between the two lies in the way they view the likelihood of outcomes. Decision theory treats all outcomes as exogenous events, 'moves of nature'. Game theory, in contrast, focuses on those situations in which outcomes are determined by interactions of deliberating agents. It proposes that agents consider outcomes as determined by other agents' reasoning, and that each agent therefore assesses the likelihood of an outcome by trying to figure out how the other agents they interact with will reason. The likelihoods of outcomes therefore become "endogenous" in the sense that players take their opponents' payoffs and rationality into account when considering the consequences of their strategies.

The formal theory defines a game as consisting of a set of *players*, a set of *pure strategies* for each player, an *information set* for each player, and the players' *payoff functions*. A player's pure strategy specifies her choice for each time she has to choose in the game. If a player's strategy requires choices at more than one time,

[2] For example, David Lewis [1969] introduced the notion of common knowledge, and Allan Gibbard [1973] that of the game form.

we say that the strategy contains a number of *actions*. Games in which players choose between actions simultaneously and only once are called *static games*. In *dynamic games* players choose between actions in a determined temporal order. All players of a game together determine a consequence. Each chooses a specific strategy, and their combination (which is called a *strategy profile*) yields a specific consequence. The consequence of a strategy profile can be a material prize — for example money — but it can also be any other relevant event, like being the winner, or feeling guilt. Game theory is really only interested in the players' *evaluations* of this consequence, which are specified in each players' payoff or utility function.

The part of the theory that deals with situations in which players' choice of strategies cannot be enforced is called the theory of *non-cooperative* games. *Cooperative* game theory, in contrast, allows for pre-play agreements to be made binding (e.g. through legally enforceable contracts). This article will not discuss cooperative game theory. More specifically, it will focus — for reasons of simplicity — on non-cooperative games with two players, finite strategy sets and precisely known payoff functions. The first philosophical issue with respect to these games arises from the interpretation of their payoffs.

2.1 Payoffs

Static two-person games can be represented by m-by-n matrices, with m rows and n columns corresponding to the players' strategies, and the entries in the squares representing the payoffs for each player for the pair of strategies (row, column) determining the square in question. As an example, Figure 1 provides a possible representation of the stag-hunt scenario described in the introduction.

		Col's choice	
Row's choice		$C1$	$C2$
	$R1$	2,2	0,1
	$R2$	1,0	1,1

Table 1. The stag hunt

The 2-by-2 matrix of Figure 1 determines two players, Row and Col, who each have two pure strategies: $R1$ and $C1$ (go deer hunting) and $R2$ and $C2$ (go hare hunting). Combining the players' respective strategies yields four different pure strategy profiles, each associated with a consequence relevant for both players: $(R1, C1)$leads to them catching a deer, $(R2, C1)$ leaves Row with a hare and Col with nothing, *(R2,C2)* gets each a hare and $(R1, C2)$ leaves Row empty-handed and Col with a hare. Both players evaluate these consequences of each profile. Put informally, players rank consequences as 'better than' or 'equally good as'. In the stag-hunt scenario, players have the following ranking:

This ranking can be quite simply represented by a numerical function u, according to the following two principles:

Row	Col
1. $(R1, C1)$	1. $(R1, C1)$
2. *(R2,C1); (R2,C2)*	2. *(R1,C2); (R2,C2)*
3. $(R1, C2)$	3. $(R2, C1)$

Figure 1. The hunters' respective rankings of the strategy profiles

1. For all consequences $X, Y : X$ is better than Y if and only if $u(X) > u(Y)$
2. For all consequences $X, Y : X$ is equally good as Y if and only if $u(X) = u(Y)$

A function that meets these two principles (and some further requirements that are not relevant here) is called an *ordinal utility function.* Utility functions are used to represent players' evaluations of consequences in games. One of the most important methodological principles of game theory is that *every* consideration that may affect a player's choice is included in the payoffs. If an agent, for example, cared about the other players' well-being, this would have to be reflected in her payoffs. The payoffs thus contain all other behaviour-relevant information except beliefs.

Convention has it that the first number represents Row's evaluation, while the second number represents Col's evaluation. It is now easy to see that the numbers of the game in Figure 1 represent the ranking of Figure 2. Note, however, that the matrix of Figure 1 is not the only way to represent the stag-hunt game. Because the utilities only represent rankings, there are many ways how one can represent the ranking of Figure 2. For example, the games in figure 3 are identical to the game in Figure 1.

	$C1$	$C2$
$R1$	-5,-5	-7,-6
$R2$	-7,-7	-6,-6

(a)

	$C1$	$C2$
$R1$	100,100	1,99
$R2$	99,1	99,99

(b)

	$C1$	$C2$
$R1$	-5,100	-7,99
$R2$	-6,1	-6,99

(c)

Figure 2. Three versions of the stag hunt

In Figure 3a, all numbers are negative, but they retain the same ranking of consequences. And similarly in 3b, only that here the proportional relations between the numbers (which do not matter) are different. This should also make clear that utility numbers only express a ranking for one and the same player, and do not allow a comparison of different players' evaluations. In 3c, although the numbers are very different for the two players, they retain the same ranking as in Figure 1. Comparing, say, Row's evaluation of $(R1, C1)$ with Col's evaluation of $(R1, C1)$ simply does not have any meaning.

Note that in the stag-hunt game, agents do not gain if others lose. Everybody is better off hunting deer, and lack of coordination leads to losses for all. Games with this property are therefore called *coordination games.* They stand in stark contrast to games in which one player's gain is the other player's loss. Most social games are of this sort: in chess, for example, the idea of coordination is wholly absent. Such games are called *zero-sum games.* They were the first games to be treated theoretically, and the pioneering work of game theory, von Neumann and Morgenstern's [1947] *The Theory of Games and Economic Behaviour* concentrates solely on them. Today, many of the games discussed are of a third kind: they combine coordination aspects with conflictual aspects, so that players may at times gain from coordinating, but at other times from competing with the other players. A famous example of such a game is the Prisoners' Dilemma, to be discussed shortly.

Players can create further strategies by *randomizing* over pure strategies. They can choose a randomization device (like a dice) and determine for each chance result which of their pure strategies they will play. The resultant probability distribution over pure strategies is called a *mixed strategy* σ. For example, Row could create a new strategy that goes as follows: toss a (fair) coin. Play $R1$ if heads, and $R2$ if tails. Because a fair coin lands heads 50% of the time, such a mixed strategy is denoted $\sigma_R = (0.5, 0.5)$. As there are no limits to the number of possible randomization devices, each player can create an infinite number of mixed strategies for herself. The players' evaluation of mixed strategies profiles is represented by the *expected values* of the corresponding pure-strategy payoffs. Such an expected value is computed as the weighted average of the pure-strategy payoffs, where the weights are given by the probabilities with which each strategy is played. For example, if Row in Figure 1 plays her mixed strategy $\sigma_R = (0.5, 0.5)$, and Col plays a strategy $\sigma_C = (0.8, 0.2)$, then Row's expected utility will be computed by:

$$u_R(\sigma_R, \sigma_C) = 0.5(0.8 \times 2 + 0.2 \times 0) + 0.5(0.8 \times 1 + 0.2 \times 1) = 1.3$$

With the same mixed strategies, Col's expected utility, $u_C(\sigma_R, \sigma_C) = 1$. For the payoffs of mixed strategy to be computable, the utility function has to carry *cardinal* information. That is, now it is also important how much a player prefers a consequence X to a consequence Y, in comparison to another pair of consequences X and Z. Because mixed strategies are a very important technical concept in game theory (although, as we will argue, the interpretation of this notion is often problematic), it is generally assumed that the utility functions characterizing the payoffs are cardinal.

It is important to note that the cardinal nature of utilities does not by itself allow making interpersonal comparisons. In fact, such interpersonal comparisons play no role in standard game theory at all. There are several reasons for this. The first is that the standard way how payoffs are measured does not permit interpersonal comparisons. Payoffs are usually interpreted as von Neumann-Morgenstern utility functions (NMUFs), which are constructed (in theory at least) with the so-called reference lottery technique. In this technique, an agent is asked to state probabilities p with which he or she is indifferent between obtaining an intermediately preferred outcome for sure, and a lottery involving the best and the worst outcomes with probabilities p and 1-p (see e.g., [Hirshleifer and Riley, 1992, pp. 16-7] for a more thorough account). Both indifference judgments, and the judgments concerning what is the (overall) best and the worst outcome, are subjective assessments of one individual, and cannot be transferred to other individuals. Thus, when using NMUFs, it is meaningless to compare different persons' utility schedules. (And although we do not discuss them here, this meaninglessness verdict also applies to other standard utility measures.)

The second reason is that standard accounts of strategic thinking do not require the players to make interpersonal comparisons. They only maximise their own utility, and they predict other players' choices by supposing that they also maximise their respective utilities. Thus, comparisons are only made between one player's evaluation of outcomes, and not between evaluations of different players.

Steven Kuhn [2004], however, has argued that standard accounts of evolutionary dynamics and equilibrium in evolutionary game theory require interpersonal comparisons. Evolutionary game theory takes a population perspective, in which different strategies in a population compete for higher rates of replication. Payoffs in such evolutionary games represent proportions of replication — that is, how much more a strategy replicates in a certain population, when compared to its competitors. Such proportional payoffs obviously compare across strategies. This may be unproblematic in biological applications, where payoffs are interpreted as Darwinian fitness. But in many social applications of evolutionary game theory, strategies are linked to individuals, and strategy payoffs to individuals' preferences. Applying standard evolutionary dynamics and equilibria to these cases, under a natural selection interpretation, then implies the interpersonal comparability of these preferences [Grüne-Yanoff, 2008a].

2.2 Strategies

A pure strategy denotes a choice of an available action in games in strategic form. This is a relatively straightforward concept, at least insofar as the notions of availability and actions are well understood. But the concept of strategy also includes pure strategies in extensive games and mixed strategies. Both of these strategy kinds are philosophically problematic and will be discussed here.

In extensive games, a strategy specifies an action for each node in the game tree at which a player has to move. Take the following example. Player 1 is on a diet and wishes to avoid eating baked goods. When she is leaving work, she can choose whether to take the direct way home (L), which leads past a bakery, or take a detour (R). Player 2 (the bakery owner) then decides, without knowing which route player 1 intends to take, whether to spray a 'freshly baked bread aroma' in front of her bakery (l) or not (r). Deploying this aerosol is costly, but may influence player 1's preferences over cakes. If player 1 chose L, he now has to decide whether to buy a bun (b) or not (d).

The standard strategy notion in extensive games requires that actions are spec-

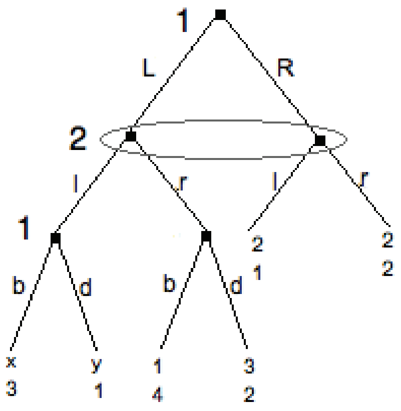


Figure 3. The baker's game

ified for each of the players' decision nodes. This has two counterintuitive implications. Let us only focus on the game tree of Figure 4 to discuss the first implication (leaving the payoffs aside for a moment). Even if player 1 chooses R at the first decision node, he also has to specify what he would choose had he chosen L, and player 2 had made her choice. As Rubinstein [1991, p. 911] points out, this is not in accord with our intuition of a '*plan of action'*. Such a plan, as commonly understood, would for this game require player 1 to decide between L and R, and *only if* he chose L, to make provisional choices for when player 2 has chosen l or r. A strategy in this and other extensive form games thus goes beyond a player's 'plan of action'. Further, these unintuitive aspects of strategies are crucial for game theory. In order to assess the optimality of player 2's strategy — for the case that player 1 should deviate from his plan — we have to specify player 2's expectations regarding player 1's second choice. For this reason, Rubinstein argues, it is more plausible to interpret this part of player 1's strategy as player 2's *belief* about player 1's planned future play.

According to this interpretation, extensive game strategies comprise of a player's plan *and* of his opponent's beliefs in the event that he does not follow the plan. This has important consequences for the interpretation of game theoretic results. In many cases (for example in sequential bargaining) it is assumed that strategies are stationary — i.e. that the history of the game has no influence on players' responses to their opponents' choices. Yet under the new perspective on strategies, this means that beliefs about opponents' play are also stationary. This, Rubinstein argues, eliminates a great deal of what sequential games are intended to model, namely the changing pattern in players' behaviour and beliefs, as they accumulate experience.

In addition to this stationarity issue, this view on strategies also has a problematic uniformity consequence. If a player's strategy necessarily comprises of opponents' beliefs about her choices, then the attribution of one strategy to the agent implies that all opponents hold the same belief about that player's future behaviour. This may be an implausibly strong built-in assumption, as it in particular concerns the player's behaviour off the equilibrium path.

The second implausible implication of the standard strategy notion concerns possible preference changes during the time of play. According to the standard notion, every strategy profile has a unique payoff for each player. That implies that player 1 at the initial node knows what the payoff for each of his strategies are, given the strategies of player 2. Even under incomplete information, he knows the probability distributions over possible payoffs. Yet there are intuitively plausible cases in which players may try to influence their opponents' preferences, in order to obtain better results. This introduces a strategic element into the payoff information that cannot be adequately represented by a probability distribution.

Take the baker's game as an example. According to the standard strategy notion, player 1 knows all strategy profile payoffs at the initial node. Because he is on a diet, he will at the initial node have a preference for not consuming a bun (d over b). Hence, independently of whether player 2 chooses l or $r, y > x$, and more specifically $y = 3$ and $x = 1$. From that perspective, (Ld, r) is the only sub-game perfect Nash equilibrium – the baker should never try to manipulate the dieter's preferences. Yet that is an implausible conclusion — such manipulations, after all, are often successful. Somehow, the influence of the baker's strategy on the dieter's preferences should be taken into account, that is, if player 1 chooses L and player 2 chooses l, then $x > y$. But the standard strategy notion does not allow for such an influence of actual play on payoffs; and biases standard game theory to ignore such strategic preference manipulations.

A mixed strategy is a probability distribution over all pure strategies in a strategic form game. We have already discussed their simplest interpretation, namely that players randomise their pure strategy choice. The idea is that randomisation may be a conscious decision, or may develop as an unconscious habit. Critics have objected that 'real world decision makers do not flip coins'. Such a criticism is too strong, as there are plausible cases where players randomly choose an action. Often cited examples include choices when to bluff in Poker, or how to serve in Tennis. In each of these cases, the randomising player avoids being correctly predicted — and hence outguessed — by her opponent. Yet, as Rubinstein [1991, p. 913] has pointed out, these are not mixed strategies in which actions like 'always bluff' and 'never bluff' are the pure strategies. In a mixed strategy equilibrium, the players are indifferent between the mixed strategy and a pure component of that mixed strategy. In the poker or the tennis game, in contrast, the player is not indifferent between 'always bluff' (in which case she soon will become predictable and hence exploitable) and a randomising strategy.

This feature of mixed strategy equilibria has always made them 'intuitively problematic' [Aumann, 1985, p. 43]. Why should a player choose to randomise a strategy, if she is indifferent in equilibrium between the randomisation and any pure component of the randomisation — in particular, if such randomisation is costly in terms of time or attention?

Because of these grave problems of the randomising account of mixed strategies, two alternative interpretations have been offered. The first reinterprets mixed strategies as the distribution of pure choices in a population. If populations instead of players interact in a game, a player is chosen from each population at random. The mixed strategy then specifies the probabilities with which the pure choices are drawn from the population. That is, mixed strategies are defined over sufficiently large population of players, each of which plays a pure strategy. This is an important context that we will examine in section 4, when discussing evolutionary games. But it is a rather specific one, which does not justify employing the notion of mixed strategy in a context where players are unique individuals; and it only holds in particular situations in this context anyway [Maynard Smith, 1982, pp. 183-88].

The second alternative is to reinterpret mixed strategies as the way in which games with incomplete information appear to outside observers. Each player's payoff function is subjected to a slight random perturbation, the value of which is known only to the player herself, but the other players only know the mean of her payoff function. Thus, each player will choose a pure strategy component of her mixed strategy in the resulting incomplete information game. Harsanyi [1973] showed that this incomplete information game has pure strategy equilibria that correspond to the mixed strategy equilibria of the original game. The point here is that to outside observers, the game appears as one in which players use mixed strategies, and the concept of a mixed strategy essentially represents one player's uncertainty concerning the *other* players' choice (see also Aumann 1987). This 'purification' account of mixed strategies provides an example of a game-theoretical concept, that of the mixed strategy, which is plausible only under some interpretations of game theory. The normative problem of justifying its use led to a reformulation which is sensible only if game theory is interpreted as a framework of analysis but not if it is taken to be a prescriptive theory.

2.3 Solution Concepts

When interactive situations are represented as highly abstract games, the objective of game theory is to determine the outcome or possible outcomes of each game, given certain assumptions about the players. To do this is to *solve* a game. Various *solution concepts* have been proposed. The conceptually most straightforward solution concept is the *elimination of dominated strategies*. Consider the game in Figure 5. In this game, no matter what Col chooses, playing $R2$ gives Row a higher payoff. If Col plays $C1$, Row is better off playing $R2$, because she can obtain 3 utils instead of two. If Col plays $C2$, Row is also better off playing $R2$, because she can obtain 1 utils instead of zero. Similarly for Col: no matter what Row chooses, playing $C2$ gives her a higher payoff. This is what is meant by saying that $R1$ and $C1$ are strictly dominated strategies.

More generally, a player A's pure strategy is *strictly dominated* if there exists another (pure or mixed) strategy for A that has a higher payoff for each of A's

	$C1$	$C2$
$R1$	2,2	0,3
$R2$	3,0	1,1

Figure 4. The Prisoners' Dilemma

opponent's strategies. To solve a game by eliminating all dominated strategies is based on the assumption that players do and should choose those strategies that are best for them, in this very straightforward sense. In cases like in that depicted in Figure 5, where each player has only one non-dominated strategy, the elimination of dominated strategies is a straightforward and plausible solution concept. Unfortunately, there are many games without dominated strategies, for example the game of Figure 6.

	$C1$	$C2$	$C3$
$R1$	3,4	2,5	1,3
$R2$	4,8	1,2	0,9

Figure 5. A game without dominated strategies

For these kinds of games, the *Nash equilibrium* solution concept offers greater versatility than dominance or maximin (as it turns out, all maximin solutions are also Nash equilibria). In contrast to dominated strategy elimination, the Nash equilibrium applies to strategy profiles, not to individual strategies. Roughly, a strategy profile is in Nash equilibrium if none of the players can do better by *unilaterally* changing her strategy. Take the example of matrix 6. Consider the strategy profile $(R1, C1)$. If Row knew that Col would play $C1$, then she would play $R2$ because that's the best she can do against $C1$. On the other hand, if Col knew that Row would play $R1$, he would play $C2$ because that's the best he can do against $R1$. So $(R1, C1)$ is not in equilibrium, because at least one player (in this case both) is better off by unilaterally deviating from it. Similarly for $(R1, C3)$, *(R2, C1), (R2, C2)* and $(R2, C3)$: in all these profiles, one of the players can improve her or his lot by deviating from the profile. Only *(R1, C2)* is a pure strategy Nash equilibrium — neither player is better off by unilaterally deviating from it.

There are games without a pure strategy Nash equilibrium, as matrix 7 shows. The reader can easily verify that each player has an incentive to deviate, whichever pure strategy the other chooses.

However, there is an equilibrium involving mixed strategies. Randomizing between the two strategies, assigning equal probability to each, yields a payoff of $0.5(0.5 \times 1 + 0.5 \times -1) + 0.5(0.5 \times 1 + 0.5 \times -1) = 0$ for both players. As mutually best responses, these mixed strategies constitute a Nash equilibrium. As one of

	$C1$	$C2$
$R1$	1,-1	-1,1
$R2$	-1,1	1,-1

Figure 6. Matching pennies

the fundamental results of game theory, it has been shown that *every* finite static game has a mixed-strategy equilibrium [Nash, 1950]. As discussed in the previous section, the interpretation of this equilibrium is problematic. If Row knew that Col plays a mixed strategy, she would be indifferent between randomising herself and playing one of the pure strategies. If randomisation came with a cost, she would prefer playing a pure strategy. So the mixed equilibrium seems unstable. If Col knew which pure strategy Row would play, he would exploit this knowledge by choosing a pure strategy himself. But that would give Row incentives again to randomise. So the mixed equilibrium would be re-installed.

Many games have several Nash equilibria. Take for example Figure 1. There, neither player has an incentive to deviate from $(R1, C1)$, nor to deviate from $(R2, C2)$. Thus both strategy profiles are pure-strategy Nash equilibria. With two or more possible outcomes, the equilibrium concept loses much of its appeal. It no longer gives an obvious answer to the normative, explanatory or predictive questions game theory sets out to answer. The assumption that one specific Nash equilibrium is played relies on there being some mechanism or process that leads all the players to expect the same equilibrium. Various *equilibrium refinements* try to rule out some of the many equilibria by capturing these underlying intuitions.

Schelling's [1960] theory of *focal points* suggests that in some "real-life" situations players may be able to coordinate on a particular equilibrium by using information that is abstracted away by the payoffs in the strategic form. Focal points are equilibria that are somehow salient. Names of strategies and past common experiences of the players provide examples of information that has such salience. It will remain very difficult to develop systematic work on the "focalness" of various strategies because what the players take to be focal depends on their cultural and personal backgrounds and salience is by definition not reflected in the payoffs. This fact makes it very hard to incorporate these concepts into the formal structure of game theory (but see [Bacharach, 1993; Sugden, 1995]).

Other refinement notions might appear to evade such context-dependence. Two prominent examples are *payoff dominance* and *risk dominance*. Consider the following coordination games:

We say that the strategy profile $(R1, C1)$ *payoff dominates* $(C2, R2)$ if and only if the payoffs of $(R1, C1)$ for each player are equal or larger than the payoffs for $(R2, C2)$ and at least one of these inequalities is strict. The intuition behind this refinement idea is that players will be able to coordinate on playing a certain strategy profile if this strategy profile is Pareto-efficient for all.

In contrast to this position, it has been argued that players may not only take

	$C1$	$C2$
$R1$	5,5	0,4
$R2$	4,0	2,2

	$C1$	$C2$
$R1$	A,a	C,b
$R2$	B,c	D,d

Figure 7. A coordination game

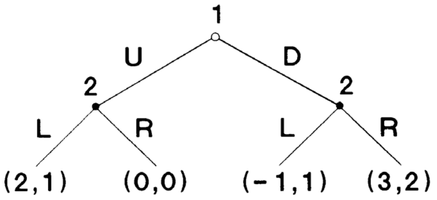
the payoff magnitude into account when selecting amongst multiple Nash equilibria, but that they also consider the risk of ending up in a non-equilibrium state. In other words, $(R1, C1)$ may be better for Row in Figure 8, but the possibility that Col makes a mistake and chooses $C2$ when Row chooses R1 bears such a risk that it is safer for Row to choose R2 (by symmetry, the same applies to Col). We say that $(R2, C2)$ *risk dominates* $(R1, C1)$ if and only if $(C-D)(c-d) \geq (B-A)(b-a)$ [Harsanyi and Selten, 1988, lemma 5.4.4]. Thus, in the same game of Figure 8, (R1, C1) is payoff dominant, while $(R2, C2)$ is risk dominant. Cases of such possible conflicts between refinement solution concepts are exacerbated by an embarrassment of riches. More and more competing refinements were developed, some of which imposed massive demands on the agents' ability to reason (and enormous faith that other agents will follow similar reasoning paths). Some were difficult to work with and their predictions were not always consistent with intuition, common sense or experimental evidence. Even more troubling, no proper basis was found from which to interpret these refinements or to choose between them.

As it will become clearer in section 3.2, the assumptions underlying the application of the Nash concept are somewhat problematic. The most important alternative solution concept is that of *rationalizability,* which is based on weaker assumptions. Players assign a subjective probability to each of the possible strategies of their opponents, instead of postulating their opponents' choices and then finding a best response to it, as in the Nash procedure. Further, knowing their opponent's payoffs, and knowing they are rational, players expect others to use only strategies that are best responses to some belief they might have about themselves. A strategy is rationalizable for a player if it survives indefinitely repeated selections as a best response to some rational belief she might have about the strategies of her opponent. A strategy profile is rationalizable if the strategies contained in it are rationalizable for each player. It has been shown that every Nash equilibrium is rationalizable. Further, the set of rationalizable strategies is nonempty and contains at least one pure strategy for each player [Bernheim, 1984; Pearce, 1984]. Rationalizability is thus often applicable, but there are often too many rationalizable strategies, so that this solution concept often does not provide a clear answer to the advisory and predictive questions posed to game theory, and it is thus seldom actually used in real-world applications.

All solution concepts discussed so far can be applied both to strategic and extensive form games. However, the extensive form provides more information than the strategic form, and this extra information sometimes provides the basis for further refinements. Take the example of Figure 9. The game has three Nash equilibria: *(U, (L,L)); (D, (L,R))* and *(D, (R,R))*. But the first and the third equilibria are suspect, when one looks at the extensive form of the game. After all, if player 2's *right* information set was reached, the he should play R (given that R gives him 3 utils while L gives him only -1 utils). But if player 2's *left* information set was reached, then he should play L (given that L gives him 2 utils, while R gives him only 0 utils). Moreover, player 1 should expect player 2 to choose this way, and hence she should choose D (given that her choosing D and player 2 choosing R gives her 2 utils, while her choosing U and player 2 choosing L gives her only 1 util). The equilibria *(U, (L,L))* and *(D, (R,R))* are not "credible', because they rely on an "empty threat" by player 2. The threat is empty because player 2 would never wish to carry it out. The Nash equilibrium concept neglects this sort of information, because it is insensitive to what happens off the path of play.

	L, L	L, R	R, L	R, R
U	2,1	2,1	0,0	0,0
D	-1,1	3,2	-1,1	3,2

(a)



(b)

Figure 8. Strategic and extensive form

The simplest way to formalise this intuition is the *backward-induction* solution concept, which applies to finite games of perfect information [Zermelo, 1913]. Since the game is finite, it has a set of penultimate nodes, i.e. nodes whose immediate successors are terminal nodes. Specify that the player who can move at each such node chooses whichever action that leads to the successive terminal node with the highest payoff for him (in case of a tie, make an arbitrary selection). So in the game of Figure 9b, player 2's choice of R if player 1 chooses U and her choice of L if player 1 chooses D can be eliminated, so that the players act as if they were faced with the following truncated tree:

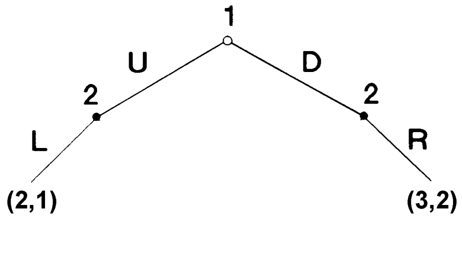


Figure 9. First step of backward induction

Now specify that each player at those nodes, whose immediate successors are the penultimate nodes, chooses the action that maximizes her payoff over the feasible successors, given that the players at the penultimate nodes play as we have just specified. So now player 1's choice U can be eliminated:

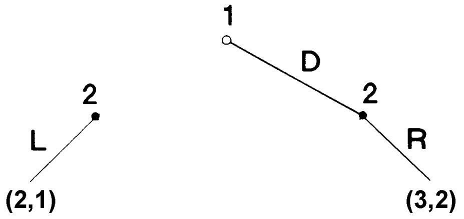


Figure 10. Second step of backward induction

Then roll back through the tree, specifying actions at each node (not necessary for the given example anymore, but one gets the point). Once done, one will have specified a strategy for each player, and it is easy to check that these strategies form a Nash equilibrium. Thus, each finite game of perfect information has a pure-strategy Nash equilibrium.

Backward induction fails in games with imperfect information. In a game like that in Figure 10, there is no way to specify an optimal choice for player 2 in his second information set, without first specifying player 2's belief about the previous choice of player 1.

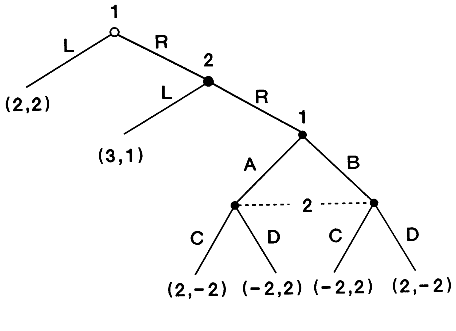


Figure 11. A game not solvable by backward induction

However, if one accepts the argument for backward induction, the following is also convincing. The game beginning at player 1's second information set is a simultaneous-move game identical to the one presented in Figure 7. The only Nash equilibrium of this game is a mixed strategy with a payoff of 0 for both players (as noted earlier in this section when we discussed the matching pennies game). Using this equilibrium payoff as player 2's payoff from choosing R, it is obvious that player 2 maximizes his payoff by choosing L, and that player 1 maximizes her payoff by choosing R. More generally, an extensive form game can be analyzed into *proper subgames*, each of which satisfies the definition of extensive-form games in their own right. Games of imperfect information can thus be solved by replacing a proper subgame with one of its Nash equilibrium payoffs (if necessary, repeatedly), and performing backward induction on the reduced tree. This equilibrium refinement technique is called *subgame perfection.*

Backward induction is based on the idea that players expect other players' behaviour to be rational in future decision nodes. *Forward induction* [Kohlberg and Mertens, 1986] is the converse of this: players expect others to have been rational in their previous choices. Consider game *G*', commonly known as the 'Battle of the Sexes', which is depicted in Figure 11a.

	LEFT	*RIGHT*
TOP	4,2	0,0
BOTTOM	0,0	2,4

Figure 12. Game G'

This game has no pure strategy equilibria, but we can compute that in a mixed strategy equilibrium (2/3, 1/3) the expected payoff is 4/3 for both players. Consider now how this game would be played if prior to playing game *G*' there was another game (depicted in Figure 11b) in which playing in G' was one of the possible strategies:

	IN	*OUT*
IN	G', G'	4,1
OUT	3,4	3,3

Figure 13. Game G

Since the expected payoff in G' is 4/3>1, the column player (he) has a dominant strategy of playing *IN*. The row player (she) then has a dominant strategy of playing *OUT*, so that the solution to G seems to be (OUT, *IN*). However, consider how she could rationalise a choice of *IN*. If she does enter the game *G*', she must be communicating her intention to obtain the best possible outcome (4, 2), and given that he understands this, he should choose *LEFT* if he were to find himself in this game. Notice that she could have secured a payoff of 3 by staying out, and that the intention of playing the (TOP, *LEFT*) equilibrium is the only reason for her to enter G'. One might object that she should simply never enter because the expected payoff in G' is lower than that from playing OUT. The forward induction argument thus asks us to consider a counterfactual world in which something that is unimaginable from the point of view of other game-theoretic principles happens.

As we saw in the case of static games, different solution concepts may sometimes give conflicting advice. A similar problem arises in the case of dynamic games: according to the forward induction argument, entering *G*' seems like a perfectly rational thing to do. Indeed, the very idea of forward induction is to interpret all previous choices as rational. If the choice of *IN* is taken as a mistake instead, it seems reasonable to continue to play the mixed strategy equilibrium. It is not very surprising that there is a fair amount of discussion on the plausibility of forward induction. As Binmore [2007, p. 426] suggests, this is because people's intuitions about how the relevant counterfactuals are to be interpreted depend on details concerning how exactly the game has been presented. If you were to find yourself in game G' as a column player, would you randomise? We will continue discussing the role of counterfactuals in backward and forward induction in section 3.3.

Because of the context-dependence and possibility of contradiction, game theorists are cautious about the use of refinements. Rather, they seem to have settled for the Nash equilibrium as the 'gold standard' of game-theoretic solution concepts (see [Myerson, 1999] for a historical account). Yet as we show in section 3.2, justifications of why players should or will play equilibrium strategies are rather shaky. Instead of privileging one solution concept, one needs to take a closer look at how the choice of solution concepts is justified in the application of game theory to particular situations. This leads us to the discussion of the architecture of game theory.

2.4 The Architecture of Game Theory

The structure of game theory is interesting from the perspective of the philosophy of science. Like many other theories, it employs highly abstract models, and it seeks to explain, predict and advice on real world phenomena by a theory that operates through these abstract models. What is special about game theory, however, is that this theory does not provide a general and unified mode of dealing with all kinds of phenomena, but rather offers a 'toolbox', from which the right tools must be selected.

Ken Binmore distinguishes between *modelling* and *analysing* a game (1994, pp. 27, 161-2, 169). Modelling means constructing a game model that corresponds to an imaginary or a real world situation. Analysing means choosing and applying a solution concept to a game model, and deriving a prediction of or a prescription for the players' choices.[3] Grüne-Yanoff and Schweinzer [2008] distinguish three main components of game theory. The *theory proper* (on the left hand side of Figure 12) specifies the concept of a game, provides the mathematical elements that are needed for the construction of a game form, and offers solution concepts for the thus constructed games. The *game structure* (left half of the central circle of

[3] Ross [2006b, p. 24] seems to suggest, however, that choosing the solution concept is part of modelling because the choice of a refinement depends on the 'underlying dynamics that equipped players with dispositions prior to commencement of a game'. But Ross does not specify how the choice is made in the end.

Figure 12) is a description of a particular game that is constructed using elements of the theory proper. The *model narrative* (the right half of the central circle of Figure 12) provides an account of a real or a hypothetical economic situation. Its account of the situation interprets the game.

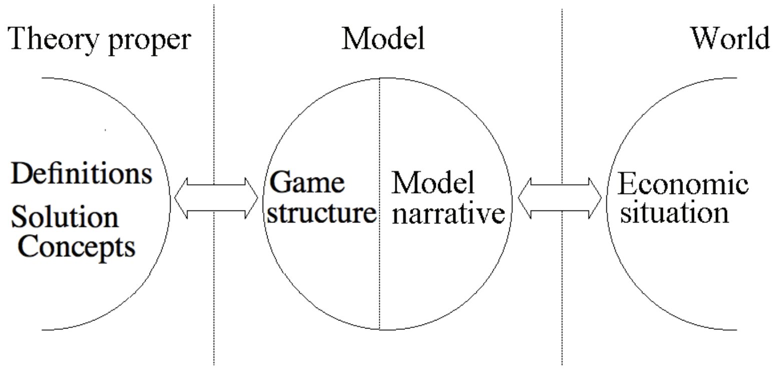


Figure 14. The architecture of game theory

A game model consists of a formal game structure and an informal model narrative. The game structure — formally characterised as a set-theoretic object — specifies the number of players, their strategies, information-sets and their payoffs.[4] The function of the theory proper is to constrain which set-theoretical structures can be considered as games, and to offer a *menu* of solution concepts for possible game structures. Game theorists often focus on the development of the formal apparatus of the theory proper. Their interest lies in proposing alternative equilibrium concepts or proving existing results with fewer assumptions, not in representing and solving particular interactive situations. "Game theory is for proving theorems, not for playing games" (Reinhard Selten, quoted in [Goeree and Holt, 2001, p. 1419]).

One reason for distinguishing between theory proper and the game model (or between analysing and modelling) is to force the game theorist to include all possible motivating factors in the payoffs. If this is assumed, introducing new psychological variables during the analysis is ruled out. Binmore argues [1994, pp. 161-162], for example, that Sen's arguments on sympathy and commitment should be written into the payoffs of a game, i.e. that they should be taken into account when it is being modelled. The point with the distinction is thus that critics of game theory should not criticise game-theoretical analyses by invoking issues that belong to modelling. This is surely a reasonable requirement. Indeed, Binmore's point is not new in the discussion. Game theorists and decision the-

[4] The term 'game' is also used for this mathematical object, but since it is also often used to refer to the combination of the game structure *and* the accompanying narrative (think of 'Prisoner's dilemma' for example), we hope that clarity is served by distinguishing between game structures and game models.

orists have always subscribed to the idea that payoffs should be interpreted as *complete descriptions* of all possible factors that may motivate the players (see e.g., [Kohlberg and Mertens, 1986]). Furthermore, it has been recognised that if payoffs are interpreted as complete descriptions, the theory proper is empirically empty.

> In our view, game theory ... should be regarded as a purely formal theory lacking empirical content. Both theories merely state what will happen if all participants have consistent preferences and follow their own preferences in a consistent manner – whatever these preferences may be. Empirical content comes in only when we make specific assumptions about the nature of these preferences and about other factual matters [Harsanyi, 1966, pp. 413-4].

Yet not only the theory proper lacks empirical content, but the game structure does too. Although they habitually employ labels like 'players', 'strategies' or 'payoffs', the game structures that the theory proper defines and helps solving are really only abstract mathematical objects. To acquire meaning, these abstract objects must be interpreted as representations of concrete situations. The interpretation is accomplished by an appropriate *model narrative* (cf. Morgan's [2005] discussion of *stories* in game theory). Such narratives are very visible in game theory — many models, like the chicken game or the prisoners' dilemma, are named after the story that comes with the model structure. The question is whether these narratives only support the use of models, or whether they are part of the model itself [Mäki, 2002, p. 14].

As regularly exemplified in textbooks, these narratives may be purely illustrative: they may describe purely fictional situations whose salient features merely help to exemplify how a particular model structure could be interpreted. Yet in other cases, narratives facilitate the relation of game structures to the real world. The narrative does this by first conceptualising a real world situation with game-theoretic notions. It identifies agents as players, possible plans as strategies, and results as payoffs. It also makes explicit the knowledge agents possess, and the cognitive abilities they have. Secondly, the narrative interprets the given game structure in terms of this re-conceptualised description of the real-world situation. Thus, model narratives fill model structures either with fictional or empirical content.

The model narrative also plays a third crucial role in game theory. As discussed in the previous sections, a specified game structure can be solved by different solution concepts. Sometimes, as in the case of minimax and Nash equilibrium for zero-sum games, the reasoning behind the solution concepts is different, but the result is the same. In other cases, however, applying different solution concepts to the same game structure yields different results. This was the case with payoff-dominance vs. risk dominance, as well as with backward and forward induction, which we discussed in section 2.3. Sometimes, information contained in the game structure alone is not sufficient for selecting between different solution concepts. Instead, the information needed is found in an appropriate account of the situation — i.e. in the model narrative. Thus, while it is true that stories (prisoner's dilemma, battle of the sexes, hawk-dove, etc.) are sometimes presented only for illustrative purposes, they take on a far more important function in these cases. They determine, together with constraints given by theory proper, the choice of the appropriate solution concept for a specific game [Grüne-Yanoff and Schweinzer, 2008]. Because model structures alone do not facilitate the choice of solution concepts, they are incomplete. Grüne-Yanoff and Schweinzer thus argue that model structure and model narrative together form the game model, and that model narratives are an essential part of game models.[5]

This conclusion raises the issue of model identity. It is quite common to hear economists identify a real-world situation with a particular game model; for example, to say that a situation X 'is a Prisoners' Dilemma'. According to the above analysis, such a claim not only implies that any suitable description of X can serve as an interpretation of the model structure. It also implies that this description of $X$is appropriately similar to the model narrative of the prisoners' dilemma — for example, in terms of the knowledge of the agents, their cognitive abilities, and their absence of sympathy and altruism. Without this additional requirement of similarity of the informal background stories, identification of game model with concrete situations may lead to the unjustifiable application of certain solution concepts to that situation, and hence to incorrect results.

More generally, the observations about the architecture of game theory and the role of informal model narratives in it have two important implications. First, it becomes clear that game theory does not offer a universal notion of rationality, but rather offers a menu of tools to model specific situations at varying degrees and kinds of rationality. Ultimately, it is the modeller who judges, on the basis of her own intuitions, which kind of rationality to attribute to the interacting agents in a given situation. This opens up the discussion about the various intuitions that lie behind the solution concepts, the possibility of contravening intuitions, and the question whether a meta-theory can be constructed that unifies all these fragmentary intuitions. Some of these issues will be discussed in section 3.

The second implication of this observation concerns the status of game theory as a positive theory. Given its multi-layer architecture, any disagreement of prediction and observation can be attributed to a mistake either in the theory, the game form or the model narrative. This then raises the question how to test game theory, and whether game theory is refutable in principle. These questions will be discussed in section 4.

3 GAME THEORY AS A NORMATIVE THEORY OF RATIONALITY

Game theory has often been interpreted as a part of a general theory of rational behaviour. This interpretation was already in the minds of the founders of game theory, who wrote:

[5] This third function of model narratives in game theory distinguishes it from earlier accounts of stories in economic models more generally (cf. [Morgan, 2001]).

> We wish to find the mathematically complete principles which define "rational behavior" for the participants in a social economy, and to derive from them the general characteristics of that behavior. [von Neumann and Morgenstern, 1944, p. 31]

To interpret game theory as a theory of rationality means to give it a prescriptive task: it recommends what agents *should* do in specific interactive situations, given their preferences. To evaluate the success of this rational interpretation of game theory is to investigate its justification, in particular the justification of the solution concepts it proposes. That human agents ought to behave in such and such a way does not of course mean that they will do so; hence there is little sense in testing rationality claims empirically. The rational interpretation of game theory therefore needs to be distinguished from the interpretation of game theory as a predictive and explanatory theory. The solution concepts are either justified by identifying sufficient conditions for them, and showing that these conditions are already accepted as justified, or directly, by compelling intuitive arguments.

3.1 Is Game Theory a Generalisation of Decision Theory?

Many game theorists have striven to develop a unifying framework for analysing games and single-person decision situations. Decision theory might provide foundations for game theory in several ways. (i) One can argue that payoffs are determined as revealed preferences in single-person decision problems (e.g., [Binmore, 2007, pp. 13-14]), or relatedly, that the payoffs are NMUFs. (ii) Another argument is to say that game-theoretical solution concepts can be reduced to the more widely accepted notion of rationality under uncertainty (e.g., [Aumann, 1987]). If such reduction is to be successful, one should be able to derive solution concepts from more primitive assumptions concerning individual rationality as in decision theory. In this section we will try to see whether this unificatory endeavour has been successful.[6]

We will point out several important differences: First, the interpretation of beliefs in decision theory is objective (vNM) or subjective (Savage), but game theoretical solution concepts imply restrictions on the players' beliefs, which in turn implies a 'logical' or 'necessitarian' interpretation (Section 3.1.1): the game determines what the relevant probabilities of rational agents ought to be. Second, the epistemic conditions for solution concepts are more stringent than those that derive from the decision-theoretic axioms (Section 3.1.2). The revealed preference arguments are discussed later in Section 4.4.

[6] See Mariotti [1995; 1996; 1997] for an argument that axioms of decision theory may conflict with game theoretical solution concepts. Hammond [1996; 1998; 2004] presents a thorough discussion of the role of individual utility in game theory. See also [Battigalli, 1996].

3.1.1 Common Priors and Bayesianism

To motivate the discussion, one may start by asking why the players do not simply maximise expected utility just as they do in single-person contexts [Kadane and Larkey, 1982; 1983]. A quick answer is that since the relevant probabilities often concern the other players' choices, those probabilities must be endogenously determined. In other words, one must analyse the whole game with a solution concept in order to determine the probabilities. This makes the interpretation of the beliefs a necessitarian one: arguments appealing to the players' rationality are used to determine constraints for the beliefs.

Bayesianism in game theory (e.g., [Aumann, 1987; Tan and Werlang, 1988]) can be characterised as the view that it is always possible to define probabilities for anything that is relevant for the players' decision-making. In addition, it is usually taken to imply that the players use Bayes' rule for updating their beliefs. If the probabilities are to be always definable, one also has to specify what players' beliefs are before the play is supposed to begin. The standard assumption is that such prior beliefs are the same for all players (see [Morris, 1995]). This *common prior assumption* (CPA) means that the players have the same prior probabilities for all those aspects of the game for which the description of the game itself does not specify different probabilities. Common priors are usually justified with the so called *Harsanyi doctrine* [Harsanyi, 1967-8], according to which all differences in probabilities are to be attributed solely to differences in the experiences that the players have had. Different priors for different players would imply that there are some factors that affect the players' beliefs even though they have not been explicitly modelled. The CPA is sometimes considered to be equivalent to the Harsanyi doctrine, but there seems to be a difference between them: the Harsanyi doctrine is best viewed as a metaphysical doctrine about the determination of beliefs, and it is hard to see why anybody would be willing to argue against it: if everything that might affect the determination of beliefs is included in the notion of 'experience', then it alone does determine the beliefs. The Harsanyi doctrine has some affinity to some convergence theorems in Bayesian statistics: if individuals are fed with similar information indefinitely, their probabilities will ultimately be the same, irrespective of the original priors.

The CPA however is a methodological injunction to include everything that may affect the players' behaviour in the game: not just everything that motivates the players, but also everything that affects the players' beliefs should be explicitly modelled by the game: if players had different priors, this would mean that the game structure would not be completely specified because there would be differences in players' behaviour that are not explained by the model. In a dispute over the status of the CPA, Faruk Gul [1998] essentially argues that the CPA does not follow from the Harsanyi doctrine. He does this by distinguishing between two different interpretations of the common prior, the 'prior view' and the 'infinite hierarchy view'. The former is a genuinely dynamic story in which it is assumed that there really is a prior stage in time. The latter framework refers to

Mertens and Zamir's [1985] construction in which prior beliefs can be consistently formulated. This framework however, is static in the sense that the players do not have any information on a prior stage, indeed, the 'priors' in this framework do not even pin down a player's priors for his own types. Thus, the existence of a common prior in the latter framework does not have anything to do with the view that differences in beliefs reflect differences in information only.

It is agreed by everyone (including [Aumann, 1998]) that for most (real-world) problems there is no prior stage in which the players know each other's beliefs, let alone that they would be the same. The CPA, if understood as a modelling assumption, is clearly false. Robert Aumann [1998], however, defends the CPA by arguing that whenever there are differences in beliefs, there must have been a prior stage in which the priors were the same, and from which the current beliefs can be derived by conditioning on the differentiating events. If players differ in their present beliefs, they must have received different information at some previous point in time, and they must have processed this information correctly [1999b]; see also [Aumann, 1999a; Heifetz, 1999]. Based on this assumption, he further argues that players cannot 'agree to disagree': if a player knows that his opponents' beliefs are different from his own, he should revise his beliefs to take the opponents' information into account. The only case where the CPA would be violated, then, is when players have different beliefs, and have common knowledge about each others' different beliefs and about each others' epistemic rationality. Aumann's argument seems perfectly legitimate if it is taken as a metaphysical one, but we do not see how it could be used as a justification for using the CPA as a modelling assumption in this or that application of game theory (and Aumann does not argue that it should).

3.1.2 Sufficient Epistemic Conditions for Solution Concepts

Recall that the various solution concepts presented in section 2 provide advice on how to choose an action rationally when the outcome of one's choice depends on the actions of the other players, who in turn base their choices on the expectation of how one will choose. The solution concepts thus not only require the players to choose according to *maximisation considerations*; they also require that agents maximise their expected utilities on the basis of certain beliefs. Most prominently, these beliefs include their expectations about what the other players expect of them, and their expectations about what the other players will choose on the basis of these expectations. Such epistemic conditions are not always made explicit when game theory is being discussed. However, without fulfilling them, players cannot be expected to choose in accord with specific solution concepts. To make these conditions on the agent's knowledge and beliefs explicit will thus further our understanding on what is involved in the solution concepts. In addition, if these epistemic conditions turn out to be justifiable, one would have achieved progress in justifying the solution concepts themselves. This line of thought has in fact been so prominent that the interpretation of game theory as a theory of rationality has often been called the *eductive* or the *epistemic* interpretation [Binmore, 1987]. In the following, the various solution concepts are discussed with respect to their sufficient epistemic conditions, and the conditions are investigated with regard to their acceptability.

For the solution of eliminating dominated strategies, nothing is required beyond the rationality of the players and their knowledge of their own strategies and payoffs. Each player can rule out her dominated strategies on the basis of maximization considerations alone, without knowing anything about the other player. To the extent that maximization considerations are accepted, this solution concept is therefore justified.

The case is more complex for *iterated elimination* of dominated strategies (this solution concept was not explained before, so don't be confused. It fits in most naturally here). In the game matrix of Figure 13, only Row has a dominated strategy, $R1$. Eliminating $R1$ will not yield a unique solution. Iterated elimination allows players to consecutively eliminate dominated strategies. However, it requires stronger epistemic conditions.

	$C1$	$C2$	$C3$
$R1$	3,2	1,3	1,1
$R2$	5,4	2,1	4,2
$R3$	4,3	3,2	2,4

Figure 15. A game allowing for iterated elimination of dominated strategies

If Col knows that Row will not play $R1$, she can eliminate $C2$ as a dominated strategy, given that $R1$ was eliminated. But to know that, Col has to know:

i. Row's strategies and payoffs

ii. That Row knows her strategies and payoffs

iii. That Row is rational

Let's assume that Col knows i.-iii., and that he thus expects Row to have spotted and eliminated $R1$ as a dominated strategy. Given that Row knows that Col did this, Row can now eliminate $R3$. But for her to know that Col eliminated $C2$, she has to know:

i. Row's (i.e. her own) strategies and payoffs

ii. That she, Row, is rational

iii. That Col knows i.-ii.

iv. Col's strategies and payoffs

v. That Col knows her strategies and payoffs

vi. That Col is rational

Let us look at the above epistemic conditions a bit more closely. i. is trivial, as she has to know her own strategies and payoffs even for simple elimination. For simple elimination, she also has to be rational, but she does not have to know it — hence ii. If Row knows i. and ii., she knows that she would eliminate $R1$. Similarly, if Col knows i. and ii., he knows that Row would eliminate $R1$. If Row knows that Col knows that she would eliminate $R1$, and if Row also knows iv.-vi., then she knows that Col would eliminate $C2$. In a similar fashion, if Col knows that Row knows i.-vi., she will know that Row would eliminate $R3$. Knowing this, he would eliminate $C3$, leaving $(R2, C1)$ as the unique solution of the game.

Generally, iterated elimination of dominated strategy requires that each player knows the structure of the game, the rationality of the players and, most importantly, that she knows that the opponent knows that she knows this. The depth of one player knowing that the other knows, etc. must be at least as high as the number of iterated eliminations necessary to arrive at a unique solution. Beyond that, no further "he knows that she knows that he knows..." is required. Depending on how long the chain of iterated eliminations becomes, the knowledge assumptions may become difficult to justify. In long chains, even small uncertainties in the players' knowledge may thus put the justification of this solution concept in doubt.

From the discussion so far, two epistemic notions can be distinguished. If all players know a proposition p, one says that they have *mutual knowledge* of p. As the discussion of iterated elimination showed, mutual knowledge is too weak for some solution concepts. For example, condition iii insists that Row not only know her own strategies, but also knows that Col knows. In the limit, this chain of one player knowing that the other knows that p, that she knows that he knows that she knows that p,etc. is continued *ad infinitum.* In this case, one says that players have *common knowledge* of the proposition p. When discussing common knowledge, it is important to distinguish *of what* the players have common knowledge. It is standard to assume that there is common knowledge of the structure of the game and the rationality of the players.

Analysing the epistemic conditions of other solution concepts requires more space and technical apparatus than available here. Instead of discussing the derivation, we list the results for the central solution concepts in Figure 14. As shown there, for the players to adhere to solutions provided by rationalizability, common knowledge is sufficient. Sufficient epistemic conditions for pure-strategy Nash equilibria are even stronger. Common knowledge of the game structure or rationality is neither necessary nor sufficient for the justification of Nash equilibria, not even in conjunction with epistemic rationality. Instead, it is required that all players know what the others will choose (in the pure-strategy case) or what the others will conjecture all players will be choosing (in the mixed-strategy case). This is rather counter-intuitive, and it shows the limitations of the epistemic interpretation of solution concepts. Alternative interpretations of the Nash equilibrium are discussed in the next section. For further discussion of epistemic conditions of

Solution Concept	**Structure of the game**	**Rationality**	**Choices or beliefs**
Simple elimination of dominated strategies	Each player knows her payoffs	Fact of rationality	—
Iterated elimination of dominated strategies	Knowledge to the degree of iteration	Knowledge to the degree of iteration	—
Rationalizability	Common Knowledge	Common Knowledge	—
Pure-strategy Nash equilibrium	—	Fact of rationality	Mutual knowledge of choices
Mixed-strategy equilibrium in two-person games	Mutual knowledge	Mutual knowledge	Mutual knowledge of beliefs

Figure 16. (adapted from [Brandenburger, 1992]): Epistemic requirements for solution concepts

solution concepts, see [Bicchieri, 1993, Chapter 2].

3.2 Justifying the Nash Equilibrium

The Nash equilibrium concept is often seen as "the embodiment of the idea that economic agents are rational; that they simultaneously act to maximize their utility" [Aumann, 1985, p. 43]. Yet the previous analysis of the Nash equilibrium's sufficient epistemic conditions showed how strong these conditions are, and that they are too strong to derive the Nash equilibrium from decision theoretic principles. Claiming the Nash equilibrium to be an embodiment of rationality therefore needs further justification. We discuss three kinds of justifications in different contexts: in one-shot games, in repeated games, and in the evolutionary context of a population.

3.2.1 Nash Equilibria in One-Shot Games

It seems reasonable to claim that once the players have arrived at an equilibrium pair, neither has any reason for changing his strategy choice unless the other player does too. But what reason is there to expect that they will arrive at one? Why should Row choose a best reply to the strategy chosen by Col, when Row does not know Col's choice at the time she is choosing? In these questions, the notion of equilibrium becomes somewhat dubious: when scientists say that a physical system is in equilibrium, they mean that it is in a stable state, where all causal forces internal to the system balance each other out and so leave it "at rest" unless it is disturbed by some external force. That understanding cannot be applied to the Nash equilibrium, when the equilibrium state is to be reached by rational computation alone. In a non-metaphorical sense, rational computation simply does not involve causal forces that could balance each other out. When approached from the rational interpretation of game theory, the Nash equilibrium therefore requires a different understanding and justification. In this section, two interpretations and justifications of the Nash equilibrium are discussed.

Often, the Nash equilibrium is interpreted as a *self-enforcing agreement*. This interpretation is based on situations in which agents can talk to each other, and form agreements as to how to play the game, prior to the beginning of the game, but where no enforcement mechanism providing independent incentives for compliance with agreements exists. Agreements are self-enforcing if each player has reasons to respect them in the absence of external enforcement.

It has been argued that being a self-enforcing agreement is neither necessary nor sufficient for a strategy to be in Nash equilibrium. That it is not necessary is obvious in games with many Nash equilibria: not all of the equilibria could have been self-enforcing agreements at the same time. It also has been argued that Nash equilibria are not sufficient. Risse [2000] argues that the notion of self-enforcing agreements should be understood as an agreement that provides *some* incentives for the agents to stick to it, even without external enforcement. He then goes on to argue that there are such self-enforcing agreements that are not Nash equilibria. Take for example the game in Figure 16.

	$C1$	$C2$
$R1$	0,0	4,2
$R2$	2,4	3,3

Figure 17.

Let us imagine the players initially agreed to play $(R2, C2)$. Now both have serious reasons to deviate, as deviating unilaterally would profit either player. Therefore, the Nash equilibria of this game are $(R1, C2)$ and $(R2, C1)$. However, in an additional step of reflection, both players may note that they risk ending up with nothing if they *both* deviate, particularly as the rational recommendation for each is to *unilaterally* deviate. Players may therefore prefer the relative security of sticking to the strategy they agreed upon. They can at least guarantee 2 utils for themselves, whatever the other player does, and this in combination with the fact that they agreed on $(R2, C2)$ may reassure them that their opponent will in fact play strategy 2. So $(R2, C2)$ may well be a self-enforcing agreement, but it nevertheless is not a Nash equilibrium.

Last, the argument from self-enforcing agreements does not account for mixed strategies. In mixed equilibria all strategies with positive probabilities are best replies to the opponent's strategy. So once a player's random mechanism has assigned an action to her, she might as well do something else. Even though the mixed strategies might have constituted a self-enforcing agreement *before* the mechanism made its assignment, it is hard to see what argument a player should have to stick to her agreement after the assignment is made [Luce and Raiffa, 1957, p. 75].

Another argument for one-shot Nash equilibria commences from the idea that agents are sufficiently similar to take their own deliberations as simulations of their opponents' deliberations.

> The most sweeping (and perhaps, historically, the most frequently invoked) case for Nash equilibrium... asserts that a player's strategy must be a best response to those selected by other players, because he can deduce what those strategies are. Player i can figure out j's strategic choice by merely imagining himself in j's position. [Pearce, 1984, p. 1030]

Jacobsen [1996] formalizes this idea with the help of three assumptions. First, he assumes that a player in a two-person game imagines himself in both positions of the game, choosing strategies and forming conjectures about the other player's choices. Second, he assumes that the player behaves rationally in both positions. Thirdly, he assumes that a player conceives of his opponent as similar to himself; i.e. if he chooses a strategy for the opponent while simulating her deliberation, he would also choose that position if he was in her position. Jacobsen shows that on the basis of these assumptions, the player will choose his strategies so that they and his conjecture on the opponent's play constitute a Nash equilibrium. If his opponent also holds such a Nash equilibrium conjecture (which she should, given the similarity assumption), the game has a unique Nash equilibrium.

This argument has met at least two criticisms. First, Jacobsen provides an argument for Nash equilibrium conjectures, not for Nash equilibria. If each player ends up with a multiplicity of Nash equilibrium conjectures, an additional coordination problem arises over and above the coordination of which Nash equilibrium to play: now first the conjectures have to be matched *before* the equilibria can be coordinated.

Secondly, when simulating his opponent, a player has to form conjectures about his own play from the opponent's perspective. This requires that he predict his own behaviour. However, Levi [1998] raises the objection that to deliberate excludes the possibility of predicting one's own behaviour. Otherwise deliberation would be vacuous, since the outcome is determined when the relevant parameters of the choice situation are available. Since game theory models players as deliberating between which strategies to choose, they cannot, if Levi's argument is correct, also assume that players, when simulating others' deliberation, predict their own choices.

Concluding this sub-section, it seems that there is no general justification for Nash equilibria in one-shot, simultaneous-move games. This does not mean that there is no justification to apply the Nash concept to any one-shot, simultaneous-move game — for example, games solvable by iterated dominance have a Nash equilibrium as their solution. Also, this conclusion does not mean that there are no exogenous reasons that could justify the Nash concept in these games. However, the discussion here was concerned with endogenous reasons — i.e. reasons derived from the information contained in the game structure alone. And there the justification seems deficient.

3.2.2 Learning to Play Nash Equilibrium

People may be unable to play Nash equilibrium in some one-shot games, yet they may *learn* to play the equilibrium strategy if they play the same game repeatedly.[7] Playing the same game repeatedly may have different learning effects, depending on the cognitive abilities of the players and the variability of the matches. *Myopic* learners observe the results of past stage games and adjust their strategy choices accordingly. They are myopic because (i) they ignore the fact that their opponents also engage in dynamic learning, and (ii) they do not care about how their deviations from equilibrium strategies may affect opponents' future play. *Sophisticated* learners take this additional information into account when choosing a strategy. Yet most game theory abstracts from the effects of type (ii) information by focussing on games in which the incentive to influence opponents' future play is small enough to be negligible.

An important example of modelling sophisticated learners is found in Kalai and Lehrer [1993]. In an n player game (with a finite strategy set), each player knows her payoffs for every strategy taken by the group of players. Before making her choice of a period's strategy, the player is informed of all the previous actions taken. The player's goal is to maximise the present value of her total expected payoff.

Players are assumed to be subjectively rational: each player commences with subjective beliefs about the individual strategies used by each of her opponents. She then uses these beliefs to compute her own optimal strategy. Knowledge assumptions are remarkably weak for this result: players only need to know their own payoff matrix and discount parameters. They need not know anything about opponents' payoffs and rationality; furthermore, they need not know other players' strategies, or conjectures about strategies. Knowledge assumptions are thus weaker for learning Nash equilibria in this kind of infinite repetition than those required for Nash solutions or rationalizability in one-shot games.

Players learn by updating their subjective beliefs about others' play with information about previously chosen strategy profiles. After each round, all players observe each other's choices and adjust their beliefs about the strategies of their opponents. Beliefs are adjusted by *Bayesian updating*: the prior belief is conditioned

[7] People may also be able to learn the equilibrium strategy in a game G from playing a game similar but not identical to G. Because similarity between games is not sufficiently conceptualised, the literature has largely eschewed this issue and focussed almost exclusively on the case of identity (for exceptions, see [LiCalzi, 1995; Rankin *et al.*, 2000]).

on the newly available information. Kalai and Lehrer portray Bayesian updating as a direct consequence of expected utility maximisation [Kalai and Lehrer, 1993, p. 1021]. Importantly, they do not assume common priors, but only that players' subjective beliefs do not assign zero probability to events that can occur in the play of the game. On the basis of these assumptions, Kalai and Lehrer show that (i) after repeatedly playing a game, the real probability distribution over the future play of the game is arbitrarily close to what each player believes the distribution to be, and (ii) the actual choices and beliefs of the players, when converged, are arbitrarily close to a Nash equilibrium. Nash equilibria in these situations are thus justified as potentially self-reproducing patterns of strategic expectations.

Kalai and Lehrer model sophisticated learners. Unlike myopic learners, who assume that their opponents' strategies are fixed, these sophisticated learners attempt the strategies of the infinitely repeated game. These strategies, which remain fixed, contain the reaction rules that govern all players' choices. Thus Kalai and Lehrer's model deals with the problem that players' opponents also engage in dynamic learning.

However, as Fudenberg and Levine [1998, p. 238] point out, Kalai and Lehrer's model assumes that the players' prior beliefs are such that there is a plausible model that is observationally equivalent to opponents' actual strategies — in the sense that the probability distribution over histories is the same (the so-called absolute continuity assumption). For players to 'find' these beliefs in principle requires the same kind of fixed point solution that finding a Nash equilibrium does. Thus the problem of justifying the Nash equilibrium has not been solved, but only transferred to the problem of finding appropriate beliefs.

3.2.3 Nash Equilibrium in a Population

The epistemic and cognitive assumptions underlying the Nash equilibrium under the standard, individualist interpretation have led some to look for an alternative interpretation based on ideas from biology:

> Maynard Smith's book *Evolution and the Theory of Games* directed game theorists' attention away from their increasingly elaborate definitions of rationality. After all, insects can hardly be said to think at all, and so rationality cannot be so crucial if game theory somehow manages to predict their behavior under appropriate conditions. (Binmore, foreword in [Weibull, 1995, x])

Thus, the *evolutive* approach proposed that the driving force behind the arrival and maintenance of equilibrium states was a non-cognitive mechanism — a mechanism that operated in population of interacting individuals, rather than a cognitive effort of the individual (Binmore 1987). If it is valid to model people as maximisers, this can only be because 'evolutionary forces, biological, social and economic, [are] responsible for getting things maximised' [Binmore, 1994, p. 11].

This leads to an evolutionary perspective on the Nash equilibrium. Evolutionary game theory studies games that are played over and over again by players drawn from a population. These players do not have a choice between strategies, but rather are "programmed" to play only one strategy. It is thus often said that the strategies themselves are the players. Success of a strategy is defined in terms of the number of replications that a strategy will leave of itself to play in games of future generations. Rather than seeing equilibrium as the consequence of strategic reasoning by rational players, evolutionary game theory sees equilibrium as the outcome either of resistance to mutation invasions, or as the result of a dynamic process of natural selection. Its interpretation of the equilibrium concept is thus closely related to the natural scientific concept of the stable state, where different causal factors balance each other out, than that under the eductive interpretation.

Evolutionary game theory offers two ways to model this evolutionary mechanism: a static and a dynamic one. The former specifies strategies that are *evolutionary stable* against a mutant invasion. Imagine a population of players programmed to play one (mixed or pure) strategy A. Imagine further that a small fraction of players "mutate" — they now play a strategy B different from A. Let the proportion of mutants in the population be p. Now pairs of players are repeatedly drawn to play the game, each player with equal probability. Thus, for any player that has been drawn, the probability that the opponent will play $B$is p, and the probability that the opponent will play A is 1-p. A strategy A is evolutionary stable if it does better when playing against some player of the invaded population than the mutant strategy itself. More generally, a strategy is an *evolutionary stable strategy* (ESS) if for every possible mutant strategy B different from A, the payoff of playing A against the mixed strategy $\sigma(1\text{-}p,p)$ is higher than the payoff of playing B against $\sigma(1\text{-}p,p)$.

With these assumptions, the players' cognitive abilities are reduced to zero: they simply act according to the strategy that they are programmed to play, persevere if this strategy is stable against mutants, or perish. It has been shown that every ESS is a strategy that is in Nash equilibrium with itself [van Damme, 1991, p. 224]. However, not every strategy that is Nash equilibrium with itself is an ESS.

The dynamic approach of evolutionary game theory considers a selection mechanism that favours some strategies over others in a continuously evolving population. Imagine a population whose members are programmed to play different strategies. Pairs of players are drawn at random to play against each other. Their payoff consists in an increase or decrease in fitness, measured as the number of offspring per time unit. Each 'child' inherits the parent's strategy. Reproduction takes place continuously over time, with the birth rate depending on fitness, and the death rate being uniform for all players. Long continuations of tournaments between players then may lead to *stable states* in the population, depending on the initial population distribution. This notion of dynamic stability is wider than that of evolutionary stability: while all evolutionary stable strategies are also dynamically stable, not all dynamically stable strategies are evolutionary stable.

In the standard literature, these results have often been interpreted as a justification of the Nash equilibrium concept (e.g., [Mailath, 1998]). This was foreshadowed by Nash himself, who proposed a 'mass action interpretation' in his Ph.D.

thesis [Leonard, 1994]. Yet there are at least two criticisms that can be put forward against such an interpretation. First, one can question why the Nash equilibrium, which is based on rational deliberation, should match with the evolutionary concepts, even though completely different causal mechanisms operate in the rational choice and the evolutionary scenarios. Against this criticism, game theorists have offered an 'as if defence': Although there is a more fundamental story 'behind' human behaviour, they claim, it is perfectly justifiable to treat this behaviour 'as if' it was indeed driven by cognitive maximisation efforts.

> Even if strategically interacting agents do not meet these epistemic conditions, their long-run aggregate behavior will nevertheless conform with them because of the workings of biological or social selection processes. [Weibull, 1994, p. 868]

Just as Friedman [1953] had used an evolutionary argument to defend the profit maximisation assumption, evolutionary ideas are used in game theory to prop up the classical theory - with the fine difference that formal identity proofs for results from evolutionary and classical game theory now seem to offer a much more precise foundation (cf. [Vromen, forthcoming]).

The second criticism of this interpretation concerns the functions of the Nash equilibrium that are thus justified. Sometimes, the claim is that the evolutionarily justified Nash equilibrium has a predictive function: it shows that people do play Nash equilibrium. This claim is somewhat dubious, however, because it is ultimately an empirical claim that cannot be established by investigating highly stylised models. It seems common practice to accept the evolutionary interpretation as a justification of the *normative* functions of the Nash equilibrium (see [Sugden, 2001] for anecdotal evidence of this claim). In the evolutionary models, players are not assumed to have preferences that they want to maximise, and for whose efficient maximisation game theory could prescribe the most efficient course of action. When it is claimed that evolutionary stability lends legitimacy to the Nash equilibrium concept, and when this concept is then used in order to prescribe efficient behaviour, the danger of committing Hume's naturalistic fallacy is obvious — an 'ought' is derived from an 'is'.

3.3 Backward Induction

Backward induction is the most common Nash equilibrium refinement for non-simultaneous games. Backward induction depends on the assumption that rational players remain on the equilibrium path because of what they anticipate would happen if they were to deviate. Backward induction thus requires the players to consider out-of-equilibrium play. But out-of-equilibrium play occurs with zero probability if the players are rational. To treat out-of-equilibrium play properly, therefore, the theory needs to be expanded. Some have argued that this is best achieved by a theory of counterfactuals [Binmore, 1987; Stalnaker, 1999] which gives meaning to sentences of the sort "if a rational player found herself at a node out of equilibrium, she would choose ...". Alternatively, for models where uncertainty about payoffs is allowed, it has been suggested that such unexpected situations may be attributed to the payoffs' differing from those that were originally thought to be most likely [Fudenberg *et al.*, 1988].

The problem of counterfactuals cuts deeper, however, than a call for mere theory expansion. Consider the two-player non-simultaneous perfect information game in Figure 17, called the "centipede". For representational convenience, the game is depicted as progressing from left to right (instead of from top to bottom as is usual in extensive-form games). Player 1 starts at the leftmost node, choosing to end the game by playing *down*, or to continue the game (giving player 2 the choice) by playing *right.* The payoffs are such that at each node it is best for the player who has to move to stop the game if and only if she expects that the game will end at the next stage if she continues (by the other player stopping the game or by termination of the game). The two zigzags stand for the continuation of the payoffs along those lines. Now backward induction advises to solve the game by starting at the last node z, asking what player 2 would have done if he ended up here. A comparison of player 2's payoffs for his two choices implies that he would have chosen *down,* given that he is rational. Given common knowledge of rationality, the payoffs that result from player 2 choosing *down* can be substituted for node z. One now moves backwards to player 1's decision node. What would she have done had she ended up at node y? She would have chosen *down.* This line of argument then continues all the way back to the first node. Backward induction thus recommends player 1 to play *down* at the first node.

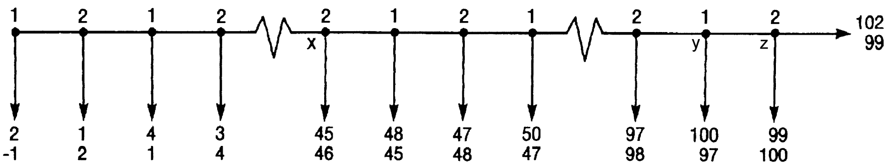


Figure 18.

So what should player 2 do if he actually found himself at node x? Backward induction tells him to play "down', but backward induction also tells him that if player 1 was rational, he should not be facing the actual choice at node $x$in the first place. So either player 1 is rational, but made a mistake ('trembled' in Selten's terminology) at each node preceding x, or player 1 is not rational [Binmore, 1987]. But if player 1 is not rational, then player 2 may hope that she will not choose *down* at her next choice either, thus allowing for a later terminal node to be reached. This consideration becomes problematic for backward induction if it also affects the counterfactual reasoning. It may be the case that the truth of the indicative conditional "If player 2 finds himself at x, then player 2 is not rational" influences the truth of the counterfactual "If player 2 were to find himself at x, then player 2 would not be rational". Remember that for backward induction to work, the players have to consider counterfactuals like this: "If player 2 found himself at x, and he was rational, he would choose *down*". Now the truth of the first counterfactual makes false the antecedent condition of the second: it can never be true that player 2 found himself at x and be rational. Thus it seems that by engaging in these sorts of counterfactual considerations, the backward induction conclusion becomes conceptually impossible.

This is an intensely discussed problem in game theory and philosophy. Here only two possible solutions can be sketched. The first answer insists that common knowledge of rationality implies backward induction in games of perfect information [Aumann, 1995]. This position is correct in that it denies the connection between the indicative and the counterfactual conditional. Players have common knowledge of rationality, and they are not going to lose it regardless of the counterfactual considerations they engage in. Only if common knowledge was not immune against evidence, but would be revised in the light of the opponents' moves, then this sufficient condition for backward induction may run into the *conceptual problem* sketched above. But common knowledge by definition is not revisable, so the argument instead has to assume *common belief* of rationality. If one looks more closely at the versions of the above argument (e.g., [Pettit and Sugden, 1989], it becomes clear that they employ the notion of common belief, and not of common knowledge.

Another solution of the above problem obtains when one shows, as Bicchieri [1993, Chapter 4] does, that limited knowledge of rationality and of the structure of the game suffice for backward induction. All that is needed is that a player, at each of her information sets, knows what the next player to move knows. This condition does not get entangled in internal inconsistency, and backward induction is justifiable without conceptual problems. Further, and in agreement with the above argument, she also shows that in a large majority of cases, this limited knowledge of rationality condition is also *necessary* for backward induction. If her argument is correct, those arguments that support the backward induction concept on the basis of common knowledge of rationality start with a flawed hypothesis, and need to be reconsidered.

3.4 Bounded Rationality in Game Players

Bounded rationality is a vast field with very tentative delineations. The fundamental idea is that the rationality which mainstream cognitive models propose is in some way inappropriate. Depending on whether rationality is judged inappropriate for the task of rational advice or for predictive purposes, two approaches can be distinguished. Bounded rationality which retains a normative aspect appeals to some version of the "ought implies can" principle: people cannot be required to satisfy certain conditions if *in principle* they are not capable to do so. For game theory, questions of this kind concern computational capacity and the complexity-optimality trade-off. Bounded rationality with predictive purposes, on the other hand, provides models that purport to be better descriptions of how people actually reason, including ways of reasoning that are clearly suboptimal and mistaken. The discussion here will be restricted to the normative bounded rationality.

The outmost bound of rationality is computational impossibility. Binmore [1987; 1993] discusses this topic by casting both players in a two-player game as Turing machines. A Turing machine is a theoretical model that allows for specifying the notion of computability. Very roughly, if a Turing machine receives an input, performs a finite number of computational steps (which may be very large), and gives an output, then the problem is computable. If a Turing machine is caught in an infinite regress while computing a problem, however, then the problem is not computable. The question Binmore discusses is whether Turing machines can play and solve games. The scenario is that the input received by one machine is the description of another machine (and vice versa), and the output of both machines determines the players' actions. Binmore shows that a Turing machine cannot predict its opponent's behaviour perfectly *and* simultaneously participate in the action of the game. Roughly put, when machine 1 first calculates the output of machine 2 and then takes the best response to its action, and machine 2 simultaneously calculates the output of machine 1 and then takes the best response to its action, the calculations of both machines enter an infinite regress. Perfect rationality, understood as the solution to the outguessing attempt in "I thank that you think that I think. . . " is not computable in this sense.

Computational impossibility, however, is very far removed from the realities of rational deliberation. Take for example the way people play chess. Zermelo [1913] showed long ago that chess has a solution. Despite this result, chess players cannot calculate the solution of the game and choose their strategies accordingly. Instead, it seems that they typically "check out" several likely scenarios and that they employ some method for evaluating the endpoint of each scenario (e.g., by counting the chess pieces). People differ in the depth of their inquiry, the quality of the "typical scenarios" they select, and the way in which they evaluate their endpoint positions.

The justification for such "piecemeal" deliberation is that computing the solution of a game can be very costly. Deliberation costs reduce the value of an outcome; it may therefore be rational to trade the potential gains from a full-blown solution with the moderate gains from "fast and frugal" deliberation procedures that are less costly (the term "fast and frugal" heuristics was coined by the ABC research group [Gigerenzer, Todd and ABC Research Group, 1999]. Rubinstein [1998] formalizes this idea by extending the analysis of a repeated game to include players' sensitivity to the *complexity* of their strategies. He restricts the set of strategies to those that can be executed by finite machines. He then defines the complexity of a strategy as the number of states of the machine that implements it. Each player's preferences over strategy profiles increase with her payoff in the repeated game, and decrease with the complexity of her strategy's complexity (He considers different ranking methods, in particular unanimity and lexicographic preferences). Rubinstein shows that the set of equilibria for complexity-sensitive games is much smaller than that of the regular repeated game.

4 GAME THEORY AS A PREDICTIVE THEORY

Game theory can be a good theory of human behaviour for two distinct reasons. First, it may be the case that game theory is a good theory of rationality, that agents are rational and that therefore game theory predicts their behaviour well. If game theory was correct for this reason, it could reap the additional benefit of great stability. Many social theories are inherently unstable, because agents adjust their behaviour in the light of its predictions. If game theory were a good predictive theory because it was a good theory of rationality, this would be because each player expected every other player to follow the theory's prescriptions and had no incentive to deviate from the recommended course of action. Thus, game theory would already take into account that players' knowledge of the theory has a causal effect on the actions it predicts [Bicchieri, 1993, chapter 4.4]. Such a *self-fulfilling theory* would be more stable than a theory that predicts irrational behaviour.[8] Players who know that their opponents will behave irrationally (because a theory tells them) can improve their results by deviating from what the theory predicts, while players who know that their opponents will behave rationally cannot. However, one should not pay too high a premium for the prospect that game theoretical prescriptions and predictions will coincide; evidence from laboratory experiments as well as from casual observations often cast a shadow of doubt on it.

Second, and independently of the question of whether game theory is a good theory of rationality, game theory may be a good theory because it offers the relevant tools to unify one's thought about interactive behaviour [Gintis, 2004; 2007]. This distinction may make sense when separating our intuitions about how agents behave rationally from a systematic account of our observations of how agents behave. Aumann for example suggests that

> [P]hilosophical analysis of the definition [of Nash equilibrium] itself leads to difficulties, and it has its share of counterintuitive examples. On the other hand, it is conceptually simple and attractive, and mathematically easy to work with. As a result, it has led to many important insights in the applications, and has illuminated and established relations between many different aspects of interactive decision situations. It is these applications and insights that lend it validity. [Aumann, 1985, p. 49]

These considerations may lead one to the view that the principles of game theory provide an approximate model of human deliberation that *sometimes* provides insights into real phenomena (this seems to be Aumann's position). Philosophy of Science discusses various ways of how approximate models can relate to real

[8] This was Luce and Raiffa's [1957] justification of the Nash Equilibrium.

phenomena, each of which has its specific problems which cannot be discussed here.

Aumann's considerations can also lead one to seek an alternative interpretation of the Nash concept that does not refer to human rationality, but retains all the formally attractive properties. In section 3.3.3 we already discussed *evolutive* approaches to game theory as a possible way to justify the normative use of the Nash equilibrium. While this normative use was marred by a number of serious doubts, the positive use of the evolutionary stability concepts seems more promising.

4.1 The Evolutive Interpretation

Evolutionary game theory was developed in biology; it studies the appearance, robustness and stability of behavioural traits in animal populations. Biology, obviously, employs game theory only as a positive, not as a normative theory; yet there is considerable disagreement whether it has contributed to the study of particular empirical phenomena, and whether it thus has any predictive function. One may get the impression that many biologists consider evolutionary game theory useful merely to studying what general evolutionary dynamics are or are not possible; or that, at best, evolutionary game theory provides an abstract mathematical language in terms of which the empirical study of biological phenomena may be described.

In contrast to this widespread scepticism in biology, many economists seem to have subscribed to the evolutive interpretation of game theory (Binmore 1987 proposed this term in order to distinguish it from the eductive approaches discussed in Section 3), and to accept it as a theory that contributes to the prediction of human behaviour. Proponents of the evolutive interpretation claim that the economic, social and biological evolutionary pressure directs human agents to behaviour that is in accord with the solution concepts of game theory, even while they have no clear idea of what is going on.

This article cannot do justice even to the basics of this very vibrant and expanding field [Maynard Smith, 1982; Weibull, 1995; Gintis, 2000], but instead concentrates on the question of whether and how this reinterpretation may contribute to the prediction of human behaviour.

Recall from section 3.3.3 that evolutionary game theory studies games that are played over and over again by players who are drawn from a population. Players are assumed to be 'programmed' to play one strategy. In the biological case, the relative fitness that strategies bestow on players leads to their differential reproduction: fitter players reproduce more, and the least fittest will eventually go extinct. Adopting this model to social settings presents a number of problems, including the incongruence of fast social change with slow biological reproduction, the problematic relation between behaviour and inheritable traits, and the difference between fitness and preference-based utility (as already discussed in section 2.1). In response to these problems, various suggestions have been made concerning how individual players could be 're-programmed', and the constitution of the population thus changed, without relying on actual player reproduction.

One important suggestion considers players' tendency to imitate more successful opponents (Schlag 1998, see also Fudenberg and Levine 1998, 66f.). The results of such models crucially depend on what is imitated, and how the imitation influences future behaviour. More or less implicitly, the imitation approach takes the notion of a meme as its basis. A meme is "a norm, an idea, a rule of thumb, a code of conduct – something that can be replicated from one head to another by imitation or education, and that determines some aspects of the behaviour of the person in whose head it is lodged" [Binmore, 1994, p. 20]. Players are mere hosts to these memes, and their behaviour is partly determined by them. Fitness is a property of the meme and its capacity to replicate itself to other players. Expected utility maximization is then interpreted as a result of evolutionary selection:

> People who are inconsistent [in their preferences] will necessarily be sometimes wrong and hence will be at a disadvantage compared to those who are always right. And evolution is not kind to memes that inhibit their own replication. [Binmore, 1994, p. 27]

This is of course a version of the replicator dynamics approach. To that extent, the theory of the fittest memes becoming relatively more frequent is an analytic truth, as long as "fitness" is no more than high "rate of replication". But Binmore then transfers the concept of strategy fitness to player rationality. Critics have claimed that this theory of meme fitness cannot serve as the basis for the claim that the behaviour of human individuals as hosts of memes will tend towards a rational pattern. The error occurs, Sugden [2001] argues, when Binmore moves from propositions that are true for memes to propositions that are true for people. In the analogous biological case — which is based on genes instead of memes — the reproductive success of phenotype depends on the *combination* of genes that carry it. Genes have positive consequences in combination with some genes but bad consequences in combination with others. A gene pool in equilibrium therefore may contain genes which, when brought together in the same individual by a random process of sexual reproduction, have bad consequences for that individual's survival and reproduction. Therefore, genes may be subject to natural selection, but there may be a stable proportion of unfit phenotypes produced by them in the population. It is thus not necessarily the case that natural selection favours phenotype survival and reproduction. The same argument holds for memes: unless it is assumed that an agent's behaviour is determined by one meme alone, natural selection on the level of memes does not guarantee that agents' behavioural patterns are rational in the sense that they are consistent with expected utility theory. But the relation between memes and behaviour is ultimately an empirical question (once the concept of meme is clarified, that is), which remains largely unexplored. It therefore remains an *empirical* question whether people behave in accord with principles that game theory proposes.

Of course, the imitation/meme interpretation of strategy replication is only one possible approach among many. Alternatives include reinforcement learning

[Börgers and Sarin, 1997] and fictitious play [Kandori *et al.*, 1993]. But the lesson learned from the above discussion also applies to these approaches: buried in the complex models are assumptions (mainly non-axiomatised ones like the meme-behaviour relation mentioned above), which ensure the convergence of evolutionary dynamics to classic equilibria. Until these assumptions are clearly identified, and until they are shown to be empirically supported, it is premature to hail the convergence results as support for the predictive quality of game theory, either under its eductive or its evolutive interpretation.

4.2 The Problem of Alternative Descriptions

While intuitions about rational behaviour may be teased out in fictional, illustrative stories, the question of whether prediction is successful is answerable only on the basis of people's observed behaviour. *Behavioural game theory* observes how people behave in experiments in which their information and incentives are carefully controlled. With the help of these experiments, and drawing on further evidence from psychology, it hopes to test game-theoretic principles for their correctness in predicting behaviour. Further, in cases where the tests do not yield positive results, it hopes that the experiments suggest alternative principles that can be included in the theory.[9] To test game theory, the theory must be specified in such detail that it may predict particular behaviour. To construct specific experimental setups, however, particular interactive phenomena need to be modelled as games, so that the theory's solution concepts can be applied. The problem of interpretation discussed in section 2.4 then surfaces. The most contentious aspect of game modelling lies in the payoffs. The exemplary case is the disagreement over the relevant evaluations of the players in the Prisoners' Dilemma.

Some critics of the defect/defect Nash equilibrium solution have claimed that players would cooperate because they would not only follow their selfish interests, but also take into account non-selfish considerations. They may cooperate, for example, because they care about the welfare of their opponents, because they want to keep their promises out of feelings of group solidarity or because they would otherwise suffer the pangs of a bad conscience. To bring up these considerations against the prisoners' dilemma, however, would expose a grave misunderstanding of the theory. A proper game uses the players' evaluations, captured in the utility function, of the possible outcomes, not the material payoff (like e.g. money). The evaluated outcome must be described with those properties that the players find relevant. Thus either the non-selfish considerations are already included in the players' payoffs (altruistic agents, after all, also have conflicting interests — e.g. which charitable cause to benefit); or the players will *not* be playing the Prisoners' Dilemma. They will be playing some other game with different payoffs.

Incorporating non-material interests in the payoffs has been criticized for making game theory empirically empty. The critics argue that with such a broad

[9] For more details on Behavioural Game Theory, their experimental methods and results, see [Camerer, 2003].

interpretation of the payoffs, any anomaly in the prediction of the theory can be dissolved by a re-interpretation of the agents' evaluations of the consequences. Without constraints on re-interpretation, the critics claim, the theory cannot be held to any prediction.

To counter this objection, many economists and some game theorists claim to work on the basis of the *revealed preference* approach. At a minimum, this approach requires that the preferences — and hence the utility function — of an agent are *exclusively* inferred from that agent's choices.[10] This ostensibly relieves game modellers from the need to engage in "psychologising" when trying to determine the players' subjective evaluations.

However, it has been argued that the application of the revealed preference concept either trivializes game theory or makes it conceptually inconsistent. The first argument is that the revealed preference approach completely neglects the importance of beliefs in game theory. An equilibrium depends on the players' payoffs and on their beliefs of what the other players believe and what they will do. In the stag hunt game of Figure 1, for example, Row *believes* that if Col *believed* that Row would play $R2$, then he would play $C2$. But if the payoff numbers represented revealed preferences, Hausman [2000] argues, then they would say how individuals would choose, given what the other chose, period. The payoffs would already incorporate the influence of belief, and belief would play no further role. Game theory as a theory of rational deliberation would have lost its job.

The second criticism claims that it is conceptually impossible that games can be constructed on the basis of revealed preferences. Take as an example the simple game in Figure 18.

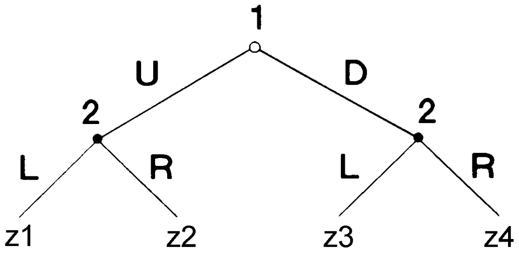


Figure 19. A game tree

How can a modeller determine the payoffs $z1 - z4$ for both players according to the revealed preference method? Let us start with player 2. Could one construct two choice situations for player 2 in which he chooses between $z1$ and $z2$ and between $z3$ and $z4$ respectively? No, argues Hausman [2000]: the two thus constructed choice situations differ from the two subgames in Figure 17 in that

[10] For a discussion of the revealed preference account, see [Grüne, 2004]. Binmore [1994, pp. 105-6, 164, 268] discusses revealed preferences in game theory. See also [Binmore, 1998, pp. 280, 358-362] and [Ross, 2005, pp. 128-136; 2006b].

they are not preceded by player 1's choice. Hence it is perfectly possible that player 2 chooses $z1$ over $z2$ in the game but chooses $z2$ over $z1$ in the constructed choice situation. Assume, for example, that player 2 considers player 1's choice of U unfair and chooses L in the game in order to take revenge. In that case she may prefer $z1$ over $z2$, but if there is no preceding choice by player 1, she may prefer $z2$ over $z1$. Thus, her choice of L merely reflects the relative desirability of $z1$ over $z2$. The problem here is that the players have state-dependent preferences: player 2 prefers $z1$ over $z2$ in one set of circumstances but $z2$ over $z1$ in another.[11] What makes this problem particularly vicious is the fact that the relevant circumstance is another player's choice in the game.

More problematically still, player 2 must be able to compare $z1$ with $z3$ and $z2$ with $z4$if one is to assign a utility function for him over all these outcomes on the basis of his choices. But it is logically impossible that she will ever face such a choice in the game, as player 1 will choose either U or D, and he will choose either between $z1$ and $z2$ or between $z3$ and $z4$. A similar argument applies to player 1. She never faces a choice between the final outcomes of this game at all, only between U and D. So the revealed preference theorist cannot assign preferences over outcomes to player 1 at all, and to player 2 only partially. This difficulty is clearly exposed in some recent efforts to provide revealed-preference conditions under which the players' choices rationalise various solution concepts.[12] These accounts start from the premise that preferences cannot be observed, and aim to provide conditions under which the players' choices may falsify or verify the *solution concept.*

Finally, assuming that the game has been properly modelled, what the modeller really can observe in a game are only its equilibria. Thus, by observing actual play, it would be possible to observe just that, say, player 1 chose U, and player 2 chose L.

We conclude that it seems conceptually impossible to construct players' payoffs by observing their choices in actual play. Further, preference elicitation procedures that partition the game into simpler subgames and infer revealed preferences from choices in those subgames are constrained by the required state-independence of preferences. As we showed, state-dependence prevents the success of such a elicitation procedure. As we have already seen, there are good reasons to believe that such state-independence is not likely because real people often care about how the other player has affected the outcome.

Further, determining whether or not preferences are state-dependent poses a problem itself. Even if the modeller were able to elicit preferences for 'Ling-with-revenge' and distinguish this from 'Ling-without-revenge' and 'Ling', he will not be able to elicit preferences for 'Ling-with-revenge-in-the-game-in-Figure-18-where-player-1-played-U' without assuming state-independence of some sort. The reason is that the only way of not making a state-independence assumption is to

[11] See Drèze and Rustichini [2004] for an overview on state-dependence.

[12] See [Sprumont, 2000] for an account of normal form games, and [Ray & Zhou, 2001] for extensive form games. Carvajal *et al.* [2004] provide an overview and additional references.

provide the game itself as the context of choice.

These problems may have contributed to a widespread neglect of the problem of preference ascription in game theoretic models. As Weibull [2004] observes:

> While experimentalists usually make efforts to carefully specify to the subject the *game form* ... they usually do not make much effort to find the subject's preferences, despite the fact that these preferences constitute an integral part of the very definition of a game. Instead, it is customary to simply hypothesize subjects' preferences. [Weibull, 2004]

Despite the problems of applying revealed preferences to game theory, the methodological rationale for the injunction to include all motivating factors into the payoffs is sound. It is just important to see its proper role in different contexts. If theoretical game theory has something to contribute, it is in providing interesting analyses of solution concepts in interesting games. For this purpose, the injunction is perfectly legitimate, and it matters very little whether or not anybody is able to find some actual situation in which preferences corresponding to the game could be elicited. It would perhaps be best to drop reference to revealed preferences and formulate the methodological argument in terms of the distinction between modelling and analysing games. One can then interpret payoffs as *dispositions* to choose (cf. [Ross, 2006a]).

The problem of preference identification has been insufficiently addressed in rational choice theory in general and in game theory in particular. But it is not unsolvable. One solution is to find a criterion for outcome individuation. Broome offers such a criterion *by justifiers*: "outcomes should be distinguished as different if and only if they differ in a way that makes it rational to have a preference between them" (Broome 1991, 103). This criterion, however, requires a concept of rationality independent of the principles of rational choice. A rational choice is no longer based on preferences alone, but preferences themselves are now based on the rationality concept. This constitutes a radical departure of how most rational choice theorists, including game theorists, regard the concept of rationality. Another option that Hausman [2005] suggests is that economists can use game theoretic anomalies to study the factors influencing preferences. By altering features of the game forms and, in particular, by manipulating the precise beliefs each player has about the game and about others' conjectures, experimenters may be able to make progress in understanding what governs choices in strategic situations and hence what games people are playing.

4.3 Testing Game Theory

Whether game theory can be tested depends on whether the theory makes any empirical claims, and whether it can be immunized against predictive failure.

Does the theory make testable claims? At first, it does not seem to do so. The solution concepts discussed in section 2.3 mainly takes the form of *theorems.*

Theorems are deductive conclusions from initial assumptions. So to test game theory, these assumptions need to be tested for their empirical adequacy. In this vein, Hausman [2005] claims that game theory is committed to contingent and testable axioms concerning human rationality, preferences, and beliefs. This claim remains controversial. Many economists believe that theories should not be tested with regard to their assumptions, but only with respect to their predictions (a widespread view that was eloquently expressed by Friedman [1953]). But the theory only makes empirical claims in conjunction with its game models.

Further, testing game theory through its predictions is difficult as such tests must operate through the mediation of models that represent an interactive situation. Here the issue of interpreting the modelled situation (see section 2.4) and of model construction drives a wedge between the predicting theory and the real world phenomena, so that predictive failures can often be attributed to model misspecification (as discussed in section 4.2).

Francesco Guala [2006] recently pointed to a specific element of game theory that seems to make an empirical claim all by itself, and independent of auxiliary hypotheses. For this purpose, he discusses the phenomenon of *reciprocity.* Agents reciprocate to other agents who have exhibited "trust" in them because they want to be kind to them. Reciprocation of an agent 1 to another agent 2 is necessarily dependent on 2 having performed an action that led 1 to reciprocate. Reciprocation is thus clearly delineated from general altruism or justice considerations.

The question that Guala raises is whether reciprocity can be accounted for in the payoff matrix of a game. The 'kindness' of an action depends on what could have been chosen: I think that you are kind to me because you could have harmed me for your benefit, but you chose not to. This would mean that the history of chosen strategies would *endogenously* modify the payoffs, a modelling move that is explicitly ruled out in standard game theory. Guala shows that the exclusion of reciprocity is connected right to the core of game theory: to the construction of the expected utility function.

All existing versions of the proofs of the existence of a utility function rely on the so-called *rectangular field assumption.* It assumes that decision makers form preferences over every act that can possibly be constructed by combining consequences with states of the world. However, if reciprocity has to be modelled in the consequences, and reciprocity depends on others' acts that in turn depend on the players' own acts, then it is *conceptually impossible* to construct acts in accord with the rectangular field assumption, because the act under question would be caught in an infinite regress. The problem is that in these cases, the Savagean distinction between consequences, states and acts cannot be consistently maintained in game theory. It follows from this that reciprocity is not the only consideration that game theory cannot put into the consequences. Things like revenge, envy, and being-suckered-in-Prisoner's-Dilemma suffer from the same problem (see also [Sugden, 1991; 1998]).

If Guala's argument is correct, it seems impossible to model reciprocity in the payoffs, and game theory is not flexible enough to accommodate reciprocity considerations into its framework. Game theory then could be interpreted as asserting that reciprocity is irrelevant for strategic interaction, or at least that reciprocity could be neatly separated from non-reciprocal strategic considerations. With this claim, game theory would be testable, and - if reciprocity were indeed an integral and non-separable factor in strategic decisions, as the evidence seems to suggest – would be refuted.

5 CONCLUSION

Game theory, this survey showed, does not provide a general and unified theory of interactive rationality; nor does it provide a positive theory of interactive behaviour that can easily be tested. These observations have many implications of great philosophical interest, some of which were discussed here. Many of the questions that arise in these discussions are still left unanswered, and thus continue to require the attention of philosophers.

BIBLIOGRAPHY

[Aumann, 1999a] R. J. Aumann. Interactive Epistemology I: Knowledge, *International Journal of Game Theory,* vol. 28, no. 3, pp. 263-300, 1999.

[Aumann, 1999b] R. J. Aumann. Interactive Epistemology II: Probability, *International Journal of Game Theory,* vol. 28, no. 3, pp. 301-314, 1999.

[Aumann, 1998] R. J. Aumann. Common Priors: A Reply to Gul, *Econometrica,* vol. 66, no. 4, pp. 929-938, 2998.

[Aumann, 1995] R. J. Aumann. Backward Induction and Common Knowledge of Rationality, *Games and Economic Behavior,* vol. 8, no. 1, pp. 6-19, 1995.

[Aumann, 1987] R. J. Aumann. Correlated Equilibrium as an Expression of Bayesian Rationality, *Econometrica,* vol. 55, no. 1, pp. 1-18, 1987.

[Aumann, 1985] R. J. Aumann. What is Game Theory Trying to Accomplish? In *Frontiers of Economics*, K.J. Arrow & S. Honkapohja, eds., Basil Blackwell, Oxford, pp. 28-76, 1985.

[Aydinonat, 2008] N. E. Aydinonat. *The invisible hand in economics: how economists explain unintended social consequences*, Routledge, London, 2008.

[Bacharach, 1993] M. Bacharach. Variable Universe Games. In *Frontiers of game theory*, K. Binmore, A.P. Kirman & P. Tani, eds., MIT Press, Cambridge Mass., pp. 255-275, 1993.

[Battigalli, 1996] P. Battigalli. The Decision-Theoretic Foundations of Game Theory: Comment. In *The rational foundations of economic behaviour: Proceedings of the IEA Conference held in Turin*, K. J. Arrow *et al.*, eds., Macmillan Press, Hampshire, pp. 149-154, 1996.

[Bernheim, 1984] D. Bernheim. Rationalizable Strategic Behavior, *Econometrica,* vol. 52, no. 4, pp. 1007-1028, 1984.

[Bicchieri, 1993] C. Bicchieri. *Rationality and coordination*, Cambridge University Press, Cambridge England; New York, USA, 1993.

[Binmore, 2008] K. Binmore. *Game theory: A very short introduction*, Oxford University Press, New York, 2008.

[Binmore, 2007] K. Binmore. *Playing for Real*, Oxford University Press, New York, 2007.

[Binmore, 1998] K. Binmore. *Game theory and the social contract: Just playing*, The MIT Press, London, 1998.

[Binmore, 1994] K. Binmore. *Game theory and the social contract: Playing fair*, MIT Press, Cambridge, Mass, 1994.

[Binmore, 1993] K. Binmore. De-Bayesing Game Theory. In *Frontiers of game theory*, K. Binmore, A.P. Kirman & P. Tani, eds., MIT Press, Cambridge Mass., pp. 321-339, 1993.

[Binmore, 1987] K. Binmore. Modeling Rational Players: Part 1, *Economics and Philosophy,* vol. 3, no. 2, pp. 179-214, 1987.

[Blackburn, 1998] S. Blackburn. *Ruling passions: a theory of practical reasoning*, Clarendon Press, Oxford; New York, 1998.

[Börgers and Sarin, 1997] T. Börgers and R. Sarin. Learning through Reinforcement and Replicator Dynamics, *Journal of Economic Theory,* vol. 77, no. 1, pp. 1-14, 1997.

[Brandenburger, 1992] A. Brandenburger. Knowledge and Equilibrium in Games, *The Journal of Economic Perspectives,* vol. 6, no. 4, pp. 83-101, 1992.

[Broome, 1991] J. Broome. *Weighing goods: equality, uncertainty and time*, Basil Blackwell, Cambridge, Mass, 1991.

[Camerer, 2003] C. Camerer. *Behavioral game theory: experiments in strategic interaction*, Princeton University Press, Princeton, N.J. ; Woodstock, 2003.

[Carvajal *et al.*, 2004] A. Carvajal, I. Ray, and S. Snyder. Equilibrium Behavior in Markets and Games: Testable Restrictions and Identification, *Journal of Mathematical Economics,* vol. 40, no. 1-2, pp. 1-40.

[Dréze and Rusichini, 2004] J. H. Drèze and A. Rustichini. State-Dependent Utility and Decision Theory. In *Handbook of utility theory*, eds. S. Barberà, P.J. Hammond & C. Seidl, Kluwer Academic Publishers, Boston, pp. 839-892, 2004.

[Friedman, 1953] M. Friedman. The Methodology of Positive Economics*, in Essays in Positive Economics*, University of Chicago Press, Chicago, pp. 3-43, 1953.

[Fudenberg *et al.*, 1988] D. Fudenberg, D. M. Kreps, and D. K. Levine. On the Robustness of Equilibrium Refinements, *Journal of Economic Theory,* vol. 44, no. 2, pp. 354-380, 1988.

[Fudenberg and Levine, 1998] D. Fudenberg and D. K. Levine. *The Theory of Learning in Games*, MIT Press, Cambridge, Mass, 1998.

[Fudenberg and Tirole, 1991] D. Fudenberg and J. Tirole. *Game theory*, MIT Press, Cambridge, Mass, 1991.

[Gibbard, 1973] A. F. Gibbard. Manipulation of Voting Schemes: A General Result, *Econometrica,* vol. 41, no. 4, pp. 587-601, 1973.

[Gigerenzer *et al.*, 1999] G. Gigerenzer, P. M. Todd, and ABC Research Group. *Simple heuristics that make us smart*, Oxford University Press, Oxford ; New York, 1999.

[Gintis, 2007] H. Gintis. A Framework for the Unification of the Behavioral Sciences, *Behavioral and Brain Sciences,* vol. 30, no. 1, pp. 1-16, 2007.

[Gintis, 2004] H. Gintis. Towards the Unity of the Human Behavioral Sciences, *Politics, Philosophy and Economics,* vol. 3, no. 1, pp. 37-57, 2004.

[Gintis, 2000] H. Gintis. *Game theory evolving: A problem-centered introduction to modeling strategic behavior*, Princeton University Press, Princeton, 2000.

[Goeree and Holt, 2001] J. K. Goeree and C. A. Holt. Ten Little Treasures of Game Theory and Ten Intuitive Contradictions, *American Economic Review,* vol. 91, no. 5, pp. 1402-1422, 2001.

[Grüne, 2004] T. Grüne. The Problem of Testing Preference Axioms with Revealed Preference Theory, *Analyse & Kritik,* vol. 26, no. 2, pp. 382-397, 2004.

[Grüne-Yanoff, 2008a] T. Grüne-Yanoff. Evolutionary Game Theory, Interpersonal Comparisons and Natural Selection, mimeo, University of Helsinki, Department of Social and Moral Philosophy, 2008.

[, Grüne-Yanoff] 008b T. Grüne-Yanoff. Game theory, *Internet encyclopedia of philosophy,* , pp. 29.4.2008. `http://www.iep.utm.edu/g/game-th.htm`.

[Grüne-Yanoff and Schweinzer, 2008] T. Grüne-Yanoff and P. Schweinzer. The Role of Stories in Applying Game Theory, *Journal of Economic Methodology,* vol. 15, no. 2, pp. 131-146, 2008.

[, Guala] 006 G. Guala. Has Game Theory been Refuted?, *Journal of Philosophy,* vol. 103, pp. 239-263, 2006.

[Gul, 1998] F. Gul. A Comment on Aumann's Bayesian View, *Econometrica,* vol. 6, no. 4, pp. 923-927, 1998.

[Hammond, 1998] P. J. Hammond. Consequentialism and Bayesian Rationality in Normal Form Games. In *Game theory, experience, rationality: foundations of social sciences, economics and ethics; in honor of John C. Harsanyi*, W. Leinfellner & E. Körner , eds., Kluwer, Dordrecht, pp. 187-196, 1998.

[Hammond, 1996] P. J. Hammond. : Consequentialism, structural rationality and game theory. In *The rational foundations of economic behaviour: Proceedings of the IEA Conference held in Turin, Italy,* K.J. Arrow, E. Colombatto & M. Perlman, ed., St. Martin's Press; Macmillan Press in association with the International Economic Association, New York; London, pp. 25, 1996.

[Hargreaves-Heap and Varoufakis, 2001] S. Hargreaves-Heap and Y. Varoufakis. *Game theory: a critical introduction*, 2nd edn, Routledge, London, 2001.

[Harsanyi, 1973] J. C. Harsanyi. Games with Randomly Disturbed Payoffs: A New Rationale for Mixed Strategy Equilibrium Points, *International Journal of Game Theory,* vol. 2, pp. 1-23, 1973.

[Harsanyi, 1967-8] J. C. Harsanyi. Games with Incomplete Information Played by 'Bayesian' Players, I-III, *Management Science,* vol. 14, pp. 159-182, 320-334, 486-502, 1967-8.

[Harsanyi, 1966] J. C. Harsanyi. A General Theory of Rational Behavior in Game Situations, *Econometrica,* vol. 34, no. 3, pp. 613-634, 1966.

[Harsanyi and Selten, 1988] J. C. Harsanyi and R. Selten. *A general theory of equilibrium selection in games*, MIT Press, Cambridge, Mass, 1988.

[Hausman, 2005] D. Hausman. 'Testing' Game Theory, *Journal of Economic Methodology,* vol. 12, no. 2, pp. 211-23, 2005.

[Hausman, 2000] D. M. Hausman. Revealed Preference, Belief, and Game Theory, *Economics and Philosophy,* vol. 16, no. 1, pp. 99-115, 2000.

[Heifetz, 1999] A. Heifetz. How Canonical is the Canonical Model? A Comment on Aumann's Interactive Epistemology, *International Journal of Game Theory,* vol. 28, no. 3, pp. 435-442, 1999.

[Hirshleifer and Riley, 1992] J. Hirshleifer and J. G. Riley. *The analytics of uncertainty and information*, Cambridge University Press, Cambridge ; New York, 1992.

[Jacobsen, 1996] H. J. Jacobsen. On the Foundations of Nash Equilibrium, *Economics and Philosophy,* vol. 12, no. 1, pp. 67-88, 1996.

[Kadane and Larkey, 1983] J. B. Kadane and P. D. Larkey. The Confusion of is and Ought in Game Theoretic Contexts, *Management Science,* vol. 29, no. 12, pp. 1365-1379, 1983.

[Kadane and Larkey, 1982] J. B. Kadane and P. D. Larkey. Subjective Probability and the Theory of Games, *Management Science,* vol. 28, pp. 113-120, 1982.

[Kalai and Lehrer, 1993] E. Kalai and E. Lehrer. Rational Learning Leads to Nash Equilibrium, *Econometrica,* vol. 61, no. 5, pp. 1019-1045, 1993.

[Kandori em et al., 1993] M. Kandori, G. J. Mailath, and R. Rob. Learning, Mutation, and Long Run Equilibria in Games, *Econometrica,* vol. 61, no. 1, pp. 29-56, 1993.

[Kohlberg and Mertens, 1986] E. Kohlberg and J.-F. Mertens. On the Strategic Stability of Equilibria, *Econometrica,* vol. 54, no. 5, pp. 10030, 1986.

[Kreps, 1990] D. M. Kreps. *Game theory and economic modelling*, Clarendon Press; Oxford University Press, Oxford; New York, 1990.

[Kuhn, 2004] S. T. Kuhn. Reflections on Ethics and Game Theory, *Synthese,* vol. 141, no. 1, pp. 1-44, 2004.

[Leonard, 1994] R. J. Leonard. Reading Cournot, Reading Nash: The Creation and Stabilisation of Nash Equilibrium, *Economic Journal,* vol. 104, no. 424, pp. 492-511, 1994.

[Levi, 1998] I. Levi. Prediction, Bayesian Deliberation and Correlated Equilibrium*, in Game Theory, Experience, Rationality*, eds. W. Leinfellner & E. Köhler, Kluwer Academic Publishers, Dordrecht, pp. 173-185, 1998.

[Lewis, 1969] D. K. Lewis. *Convention: a philosophical study*, Harvard University Press, Cambridge, Mass, 1969.

[LiCalzi, 1995] M. LiCalzi. Fictitious Play by Cases, *Games and Economic Behavior,* vol. 11, no. 1, pp. 64-89, 1995.

[Luce and Raiffa, 1957] D. R. Luce, and H. Raiffa. *Games and decisions; introduction and critical survey*, Wiley, New York, 1957.

[Mailath, 1998] G. J. Mailath. Do People Play Nash Equilibrium? Lessons from Evolutionary. Game Theory, *Journal of Economic Literature,* vol. 36, pp. 1347-1374, 1998.

[Mäki, 2002] U. Mäki. The Dismal Queen of the Social Sciences. In *Fact and Fiction in Economics*, ed. U. Mäki, Cambridge University Press, Cambridge, pp. 3-34, 2002.

[Mariotti, 1997] M. Mariotti. Decisions in Games: Why there should be a Special Exemption from Bayesian Rationality, *Journal of Economic Methodology,* vol. 4, no. 1, pp. 43-60, 1997.

[Mariotti, 1996] M. Mariotti. The Decision-Theoretic Foundations of Game Theory. In *The rational foundations of economic behaviour: Proceedings of the IEA Conference held in Turin*, ed. Arrow, Kenneth J. et al., Macmillan Press, Hampshire, pp. 133-148, 1996.

[Mariotti, 1995] M. Mariotti. Is Bayesian Rationality Compatible with Strategic Rationality?, *Economic Journal,* vol. 105, no. 432, pp. 1099, 1995.

[Maynard Smith, 1982] J. Maynard Smith. *Evolution and the theory of games*, Cambridge University Press, Cambridge ; New York, 1982.

[Mertens and Zamir, 1985] J.-F. Mertens and S. Zamir. Formulation of Bayesian Analysis for Games with Incomplete Information, *International Journal of Game Theory,* vol. 4, no. 1, pp. 1-29, 1985.

[Morgan, 2005] M. S. Morgan. The Curious Case of the Prisoner's Dilemma: Model Situation?*, in Science without laws*, eds. A. Creager, M. Norton Wise & E. Lunebeck, Duke University Press, Durham, 2005.

[Morgan, 2001] M. S. Morgan. Models, Stories and the Economic World, *Journal of Economic Methodology,* vol. 8, no. 3, pp. 361, 2001.

[Morris, 1995] S. Morris. The Common Prior Assumption in Economic Theory, *Economics and Philosophy,* vol. 1, pp. 227-253, 1995.

[Myerson, 1999] R. B. Myerson. Nash Equilibrium and the History of Economic Theory, *Journal of Economic Literature,* vol. 37, no. 3, pp. 1067-1082, 1999.

[Nash, 1950] J. F. Nash. Equilibrium Points in n-Person Games, *Proceedings of the National Academy of Science,* vol. 36, pp. 48-49, 1950.

[Osborne and Rubinstein, 1994] M. J. Osborne and A. Rubinstein. *A course in game theory*, MIT Press, Cambridge, Mass, 1994.

[Pearce, 1984] D. G. Pearce. Rationalizable Strategic Behavior and the Problem of Perfection, *Econometrica,* vol. 52, no. 4, pp. 1029-1050, 1984.

[Pettit and Sugden, 1989] P. Pettit and R. Sugden. The Backward Induction Paradox, *Journal of Philosophy,* vol. 86, no. 4, pp. 169-182, 1989.

[Rankin *et al.*, 2000] F. W. Rankin, J. B. Van Huyck, and R. C. Battalio. Strategic Similarity and Emergent Conventions: Evidence from Similar Stag Hunt Games, *Games and Economic Behavior,* vol. 32, no. 2, pp. 315-337, 2000.

[Ray and Zhou, 2001] I. Ray and L. Zhou. Game Theory Via Revealed Preferences, *Games and Economic Behavior,* vol. 37, no. 2, pp. 415-24, 2001.

[RIsse, 2000] M. Risse. What is Rational about Nash Equilibria?, *Synthese,* vol. 124, no. 3, pp. 361-384, 2000.

[Ross, 2006a] D. Ross. Evolutionary Game Theory and the Normative Theory of Institutional Design: Binmore and Behavioral Economics, *Politics, Philosophy & Economics,* vol. 5, no. 1, pp. 51-79, 2006.

[Ross, 2006b] D. Ross. Game theory, *Stanford Encyclopedia of Philosophy,* , pp. 1-78, 2006. `http://plato.stanford.edu/entries/game-theory/`.

[Ross, 2005] D. Ross. *Economic theory and cognitive science: microexplanation*, The MIT Press, Cambridge, Mass., London, 2005.

[Rubinstein, 1998] A. Rubinstein. *Modeling bounded rationality*, MIT Press, Cambridge, Mass, 1998.

[Rubinstein, 1991] A. Rubinstein. Comments on the Interpretation of Game Theory, *Econometrica,* vol. 59, no. 4, pp. 909-924, 1991.

[Schelling, 1960] T. C. Schelling. *The strategy of conflict*, Harvard University Press, Cambridge, 1960.

[Schlag, 1998] K. Schlag. Why Imitate, and if so, how? A Boundedly Rational Approach to Multi-Armed Bandits, *Journal of Economic Theory,* vol. 78, no. 1, pp. 130-156, 1998.

[Sprumont, 2000] Y. Sprumont. On the Testable Implications of Collective Choice Theories, *Journal of Economic Theory,* vol. 93, no. 2, pp. 205-232, 2000.

[Stalnaker, 1999] R. Stalnaker. Knowledge, Belief and Counterfactual Reasoning in Games*, in The logic of strategy*, eds. C. Bicchieri, R. Jeffrey & B. Skyrms, Oxford University Press, Oxford, 1999.

[Sugden, 2001] R. Sugden. The Evolutionary Turn in Game Theory, *Journal of Economic Methodology,* vol. 8, no. 1, pp. 113-130, 2001.

[Sugden, 1998] R. Sugden. Difficulties in the Theory of Rational Choice [Review Article], *Journal of Economic Methodology,* vol. 5, no. 1, pp. 157-163, 1998.

[Sugden, 1995] R. Sugden. A Theory of Focal Points, *Economic Journal,* vol. 105, no. 430, pp. 533-550, 1995.

[Sugden, 1991] R. Sugden. Rational Choice: A Survey of Contributions from Economics and Philosophy, *Economic Journal,* vol. 101, no. 407, pp. 751-785, 1991.

[Tan and Werlang, 1988] T. C. Tan and S. R. da C. Werlang. The Bayesian Foundations of Solution Concepts of Games, *Journal of Economic Theory,* vol. 45, pp. 370-339, 1988.

[van Damme, 1991] E. van Damme. *Stability and Perfection of Nash Equilibria*, 2nd rev. and enlarged ed. edn, Springer, Berlin, 1991.

[von Neumann and Morgenstern, 1947] J. von Neumann and O. Morgenstern. *Theory of games and economic behavior*, 2nd edn, Princeton University Press, Princeton, 1947.

[Vromen, forthcoming] J. J. Vromen. (forthcoming): Friedman's Selection Argument Revisited*, in The Methodology of Economics. Milton Friedman's Essay at 50*, ed. U. Mäki, Cambridge University Press, Cambridge, forthcoming.

[Weibull, 2004] J. W. Weibull. Testing Game Theory. In *Advances in understanding strategic behavior*, ed. S. Huck, Palgrave, New York, pp. 85-104, 2004.

[Weibull, 1995] J. W. Weibull. *Evolutionary game theory*, MIT Press, Cambridge, Mass, 1995.

[Wiebull, 1994] J. W. Weibull. The 'as if' Approach to Game Theory: Three Positive Results and Four Obstacles, *European Economic Review,* vol. 38, no. 3-4, pp. 868-881, 1994.

[Zermelo, 1913] E. Zermelo. Über Eine Anwendung Der Mengenlehre Auf Die Theorie Des Schachspiels*, in Proceedings of the Fifth International Congress on Mathematics*, eds. E.W. Hobson & A.E.H. Love, Cambridge University Press, Cambridge, pp. 501-504, 1913.