Has Game Theory Been Refuted?*

Francesco Guala

Department of Sociology and Philosophy
University of Exeter
Amory Building
Exeter EX4 4RJ, UK
f.guala@ex.ac.uk

Abstract. The answer in a nutshell is: yes, five years ago, but nobody has noticed. Nobody has noticed because the majority of social scientists subscribe to one of the following views: (1) the 'anomalous' behaviour observed in standard prisoner's dilemma or ultimatum game experiments has refuted standard game theory a long time ago; (2) game theory is flexible enough to accommodate any observed behaviour by 'refining' players' preferences; or (3) it is just a piece of pure mathematics (a tautology). None of these views is correct. This paper defends the view that game theory as commonly understood is not a tautology, that it suffers from important (albeit very recently discovered) empirical anomalies, and that it is not flexible enough to accommodate all the anomalies in its theoretical framework. In particular it cannot accommodate the existing evidence about reciprocal behaviour in many classic game-theoretic contexts. I conclude trying to explain why it took so long for experimental game theorists to design experiments that could adequately test the theory.

* The ideas of this paper have been presented at seminars at the University of Turin, the 2005 Economic Science Association meeting, and the 2005 British Society for the Philosophy of Science conference. I should thank the audiences at these meetings, Jim Cox, John Dupré, and Jim Woodward for their comments and suggestions. All the remaining mistakes are mine.

Introduction

In his famous 'Rational Fools' article, Amartya Sen noticed that

> if today you were to poll economists of different schools, you would almost certainly find the coexistence of beliefs (i) that the rational behaviour theory is unfalsifiable, (ii) that it is falsifiable and so far unfalsified, and (iii) that it is falsifiable and indeed patently false (Sen 1977, p. 325).

Three decades later, Sen's poll would give pretty much the same results. But although the confusion highlighted by Sen was not unjustified in the seventies, this is not the case anymore. Or so I will argue in this paper. This paper defends the view that Game Theory (GT) as commonly understood is not unfalsifiable, that it suffers from important recently discovered empirical anomalies, and that it cannot accommodate the anomalies in its theoretical framework.

A few clarifications are due before we start. First, I'll be concerned with GT in its *descriptive* (or 'positive') interpretation only. This is not to suggest that the normative version of the theory of rational choice is philosophically uninteresting, or irrelevant for its appraisal as a model of actual behaviour. However, the discussion of the normative-descriptive hybrid status of GT is worth a separate paper.[1]

Second, asking whether GT has been refuted does not imply a commitment to some naïve form of falsificationism. 'Refuted theory' here is taken somewhat loosely to mean a theory that suffers from severe empirical anomalies. Minimally, a refuted theory cannot be interpreted as universally true in its intended domain of application. I take that many scientists are interested in refutations at least for this reason, quite independently of the pragmatic status of the theory itself. (Whether it should be rejected or whether it's been superseded by one of its rivals, for example.) Towards the end of the paper in fact I shall argue that, despite its shaky empirical record, one should not rush to any conclusion regarding GT's prospects as a research programme.

Finally, this paper completely sidesteps underdetermination and Duhem-Quine problems. It's not a paper about the methodological problems of empirical testing in general, but rather deals with the much more specific issue of GT's relation with empirical evidence. It takes the available evidence as given, and assumes that all the reported data have been collected properly, by competent scientists who have correctly designed their experiments. Of course there are interesting issues in the methodology of experimental game theory that would be worth discussing in more depth, but that must be the topic of another paper. This one is about the interpretation of GT, and its descriptive power in light of the best scientific evidence to date.

[1] On the interaction between normative and descriptive interpretations of the theory of individual decision making, see e.g. Guala (2000) and Starmer (2000).

Is game theory a tautology?

Like other 'successful' scientific theories, GT can be (and is) interpreted in different ways. Thus first of all it will be necessary to clarify what we are dealing with: *what* has been refuted (if at all)? Under a standard interpretation, GT is aimed at modelling and understanding people's behaviour, in particularly when the outcome of their actions depends not only on what they do, but also on other people's decisions. According to an influential textbook,

> Game theory is a bag of analytical tools designed to help us understand the phenomena that we observe when decision-makers interact. The basic assumptions that underlie the theory are that decision-makers pursue well-defined exogenous objectives (they are *rational*) and take into account their knowledge or expectations of *other* decision-makers' behavior (they reason *strategically*). (Osborne and Rubinstein 1994, p. 1)

This characterisation rules out, first of all, so-called 'mass-action' interpretations of GT, as for instance in evolutionary GT models. GT, under the standard 'epistemic' interpretation, models individual behaviour as determined by players' preferences and beliefs, under strong rationality assumptions. This does not mean that 'mass-action' interpretations are less interesting or promising than 'epistemic' ones, but simply that the latter is by far the most developed, used, and – crucially, for our purposes – empirically tested version of GT.[2]

The standard textbook definition also rules out the interpretation of GT as a branch of pure mathematics. Although game theorists are keen to distinguish 'pure' from 'applied' GT, the former is *not* a piece of uninterpreted formalism. The Osborne and Rubinstein quote significantly mentions decision makers, expectations, knowledge, objectives, and behaviour, as the subject matter of GT. Of course there's no mention of what these objectives may be, nor indeed of what the expectations are about. 'Decision-makers' is also left vague enough to be applicable to various real-world entities (individuals, firms, nations). But this just means that 'pure' GT is *very abstract*, rather than purely formal. [3] The business of specifying preferences, beliefs, and the identity decision-makers is left to applied social scientists, and presumably varies depending on the intended domain application.

There are some dissenters, however. 'Pure' game theorists often claim that their theory is just a *tautology*. Ken Binmore is a typical example:

> Mathematical theorems are tautologies. They cannot be false because they do not say anything substantive. They merely spell out the implications of how things

[2] The distinction between these two interpretations goes back to John Nash's seminal PhD dissertation.
[3] A similar point could be made (more elaborately, perhaps) by looking at the sort of solution-concepts that are explored and considered legitimate in GT. Rationality plays a prominent role in this activity, and rationality is not a purely formal concept. It is rationality for decision-makers, and more specifically decision-makers of the *human* kind.

> have been defined. The basic propositions of game theory have precisely the same character. (Binmore 1994, p. 96)

As with other claims of this kind, it is useful to ask first what is *meant* by 'tautology', and then *what* is supposed to be a tautology (Sober 1993, pp. 68-72). A tautology is, strictly speaking, a proposition that is true in virtue of the definition of its logical terms. This is probably not what Binmore has in mind. Most likely, he is thinking of *sets* of propositions such as:

(1) Rational agents always choose uniquely dominant strategies.
(2) 'Defect' is the only dominant strategy in the one-shot prisoner's dilemma game.
(3) Therefore, rational agents always defect in the one-shot prisoner's dilemma game.

This is an argument, not a proposition. It is a valid argument, in the standard logical sense: if the premises are true, the conclusion must also be true, independently of the truth of the premises. But whether each proposition in the argument is true or false is another question. In order to figure out whether it is true that rational agents always defect in the one-shot PD game, we need to check whether rational agents always choose uniquely dominant strategies, and whether 'defect' is the only dominant strategy in the one-shot PD game.

But how do we check that (1) and (2) are true? Binmore and the other supporters of the 'tautology' interpretation presumably argue that (1) and (2) are *analytic*: they are true in virtue of the meaning of the non-logical terms that are in them. So 'rational agent' is defined as the kind of decision-maker that always chooses uniquely dominant strategies. And the one-shot PD game (Table 1) is by definition a game where 'defect' is the only dominant strategy.[4]

	Cooperate	Defect
Cooperate	2, 2	0, 3
Defect	3, 0	1, 1

Table 1

Of course there would be nothing to test, if GT was nothing but this set of propositions. 'Tautologists' like Binmore however do not conclude that GT is empirically useless:

> One reason for emphasizing the tautological nature of these tools of the game theory trade is to direct attention of the critics away from the question of how a formal game is *analyzed* to the more rewarding question of how problems of human interaction should be *modeled* as formal games. (Binmore 1994, p. 169)

[4] In representing games in normal (or strategic) form, I follow the usual conventions: the first number in each cell (or consequence, or outcome) represents the payoff of the 'Row' player, the second number the payoff of the 'Column' player. Each row/column represents a strategy (or act) available to that particular player.

So what can be tested presumably is the claim that certain GT models can be used to understand (predict, explain) certain specific real-world situations.[5] This seems analogous to what some versions of the semantic view of scientific theories (e.g. Giere 1988) call a 'theoretical hypothesis' (but perhaps the expression 'empirical hypothesis' would be less misleading). A testable hypothesis would be, for example, that the one-shot PD game can be used to explain (predict, etc.) the behaviour of experimental subjects playing the game in Table 2.

	Left	Right
Up	\$2, \$2	\$0, \$3
Down	\$3, \$0	\$1, \$1

Table 2

So far it looks as if the tautologists are just quibbling about what a scientific theory is. According to them, a theory is just the formalism. According to most scientists and philosophers, it is an *interpreted* formalism, or a formalism plus a set of 'bridge principles' that connect it to empirical reality.

But in fact some tautologists are more radical than that. Take Binmore again: "Game theorists understand the notion of a payoff in a sophisticated way that makes it tautologous that players act as though maximizing their own payoffs" (1994, p. 98). This strategy is defended by drawing an analogy with biology:

> [in applications of GT in evolutionary biology] the payoffs are [...] defined in terms of an organism's "fitness". This makes it a tautology that the organisms that survive will seem to behave as though seeking to maximize their payoffs in the game under study. The same tautological flavor is also to be found when *rational* behavior in games is studied, provided that the payoffs in the game are interpreted according to the theory of *revealed preference* [...]. (Binmore 1994, p. 104)

Revealed preference theory was originally proposed as an attempt to *reduce* the notion of preference to that of observed choice (Samuelson 1938). There is remarkable confusion nowadays about what revealed preference theory really amounts to in neoclassical economics (cf. Hausman 2000, Dowding 2002), but tautologists like Binmore seem to stick close to the original understanding: "In revealed preference theory [...] choice behavior becomes the primitive. One observes some of the choices that a player makes and then argues that he is making choices *as though* he were equipped with a preference relation." (1994, pp. 105-6).

The first thing to notice is that Binmore's analogy with fitness in evolutionary biology is not unproblematic. According to a common interpretation, the fitness of a trait X is not whatever proportion of offsprings with X are observed in the next generation. Fitness is

[5] See also Binmore (1999, p. F18): "[a consolidating experimentalist] does not see as his task as asking whether economics works or not. He already knows that sometimes it does and sometimes it does not. The task is to classify economic environments into those where the theory works and those where it does not."

the *expected* reproductive success of a trait, or the probability – given the forces of natural selection – that a member of the next generation of offsprings will have trait X. Probability here can be (and usually is) interpreted as a *propensity*, rather than a frequency. This probability may or may not be reflected approximately in the frequency of X in the next generation. The reason it may not be is that other factors besides natural selection may interfere with the heritability of X: random genetic drift, typically, or mutation, migration, recombination. It's not true that "We know who will survive because they are fittest" and "We know who are fittest because they survive" (Binmore 1994, p. 98).

Interpreted this way, fitness is not a 'tautology'. This of course does not, by itself, prevent one from interpreting utility in terms of revealed preferences. But there are other, independent reasons not to do so. Preferences cannot be 'read off' directly from behaviour because choices are determined by individual preferences *and beliefs*. This point is perhaps obscured by the convention in GT of assuming common knowledge of the game that is being played ('I know that you know that I know ... (and so forth) that such and such strategies are available, with such and such payoffs, and that all players are rational'). Of course *if* we assume common knowledge of the game, then we can derive preferences from observed choices. Whether the derivation is correct, however, depends crucially on the correctness of the common knowledge assumption.

It's not clear however why one should want to take such a strong empiricist stance regarding preferences, if one is willing to simply assume a priori a certain configuration of players' beliefs (see also Hausman 2000). Why accord to the two key variables of GT a different status? In fact the fathers of GT were fully aware of the need to *mutually* determine, by means of observable procedures, the structure of individual preferences *and* beliefs.

Savage measurements

I'll keep using an old-fashioned philosophical distinction, according to which a scientific theory can be analysed in two sets of elements. On the one hand, a set of abstract concepts connected by theoretical relationships. On the other, some 'bridge-principles' that link the concepts with empirical reality or measurement procedures. We don't have to assume that these two elements are precisely separable. The distinction is useful only for heuristic purposes, and shouldn't be read as implying any deep philosophical commitment on the theory-observation distinction, or on the nature of scientific theories.[6]

What are the bridge principles of GT? In *The Theory of Games and Economic Behavior* von Neumann and Morgenstern famously introduce a procedure for measuring (or 'operationalising') cardinal utility. Expected Utility Theory (EUT) is the 'bridge' connecting GT with the real world.

[6] In fact, it should be broadly compatible both with old-fashioned 'syntactic' approaches (like the 'Standard' or 'Received View' of theories) and with more fashionable 'Semantic' approaches (where the 'bridge principles' are sometimes called 'theoretical hypotheses', see above).

The story has become part of the folklore of GT. Von Neumann was happy to build the theory on the generic notion of 'payoff'. Indeed, he conceived of payoffs basically as *objective* (material, possibly monetary) properties of outcomes or consequences. Morgenstern persuaded von Neumann to recast the theory in utility jargon, in order to appeal to the average neoclassical economist. Von Neumann quickly scribbled on the back of an envelope the measurement procedure that was to appear in the second (1947) edition of *The Theory of Games*.

Von Neumann and Morgenstern's measurement procedure of utility is well known: starting with three outcomes x, y and z such that $x > y > z$, the utility scale is normalized by assigning $U(x) = 1$ and $U(z) = 0$. A subject is then presented with the choice between the prospect y for sure and the prospect $[p, x; (1 - p), z]$; if she prefers either the risky prospect or the sure outcome, the probabilities in the former are varied until she is indifferent. At this stage, the utility of y can be computed by $U(y) = p$, and the procedure iterated for any value between x and z.

This clearly assumes that we already know subjects' probability assignments or beliefs. In practice, since von Neumann and Morgenstern restrict their discussion to lotteries with objective probabilities, we just assume that subjects' assignments mirror the objective probability measures (their beliefs are accurate representations of the world). When this is not the case, subjective probability assignments must be worked out from people's choices, in a way that is symmetric to the one above. Start with people's preferences about sets of consequences, and then infer which states of the world are considered more probable from the observation of individuals' acts (or choices of strategy). In practice, it is usually assumed that more money is preferred to less, and then individuals are offered money prizes associated with certain types of event. But again one has to take a firm fulcrum (preferences for the measurement of beliefs, beliefs for the measurement of preferences) for the lever of expected utility theory to lift any weight.

These procedures were first sketched by Frank Ramsey (1926) and then refined by Leonard Savage (1954). Since the latter's systematisation has become standard in contemporary GT, I shall refer to it as the *Savage measurement procedure*. According to 'kosher' GT, the application of the theory should start with a Savage measurement, i.e. with the identification of players' utilities and beliefs. Once this has been done, one can begin to figure out what kind of game they are playing, and apply the solution concepts of GT.[7] Most experimental GT is not kosher from this respect, however. This causes

[7] Sometimes Binmore writes (and speaks, in conversation) as if his revealed preference theory were nothing but an application of Savage's measurement procedure. He says for example ('hawk' here stands for 'defect', 'dove' for 'cooperate'): "the preference (*dove*, *hawk*) $\prec_A$ (*hawk*, *hawk*) is deduced from the prior information that, if Adam knew he had to choose between only (*dove, hawk*) and (*hawk, hawk*), then he actually would choose (*hawk, hawk*)" (1994, p. 106). Notice that according to this formulation one does *not* observe behaviour in a PD game first, and *on the basis of this*, define the preference scales. Rather, one elicits the preference scales by pair wise comparison of the various outcomes (as in a simplified Savage measurement – more about this below), and *then* uses such information to predict behaviour in the PD game. This does not make a 'tautology' out of GT, because of course there may well be a mismatch between the preferences as 'revealed' in the two situations. And it does not reduce the concept of

problems, because the results of experimental GT can be (and are) criticised for not testing the theory at all.

What kind of game are we playing?

Take the one-shot PD game again (Table 1). Here the free riding strategy is Down for Row and Right for Column, leading to a payoff of 2 units each. These strategies are uniquely dominant, and determine the unique rational solution (the Nash solution or Nash equilibrium) of the game. Informally, whatever Column will do, Row will be better off by playing Down; and similarly Column will always be better off by playing Right, given what the opponent does. But both would be better off if they played Up-Left. The Nash equilibrium in a one-shot PD game is not a Pareto-optimal solution.

	Left	Right
Up	2, 2	0, 3
Down	3, 0	1, 1

Table 1 (again)

As is well known, in a 'standard' PD experiment a considerable number of subjects play cooperatively (Up-Left, in the game above) (Rapoport and Chammah 1965). The game played in such experiments, however, is not necessarily the one in Table 1. It is the one already presented in Table 2, which is reproduced here for comparison.

	Left	Right
Up	\$2, \$2	\$0, \$3
Down	\$3, \$0	\$1, \$1

Table 2 (again)

The 'naive' interpretation of these experimental results is that many human beings (fortunately) do not behave as predicted by the theory. But a number of game theorists reject this interpretation. They say, for example, that the preference rankings of the subjects who play cooperatively in these experiments are inadequately represented by the numbers in the classic prisoner's dilemma game matrix. Or, in other words, that Table 1 and Table 2 are not necessarily isomorphic for all experimental subjects.

According to the orthodox interpretation of game theory, the numbers in Table 1 represent the (ordinal) structure of agents' preferences. Thus, the argument goes, if we observe anomalous behaviour in the experiment, the initial conditions postulated in the model were probably not instantiated in the experiment. Subjects were not *really* playing

---

preference to that of choice, because in order to identify preferences in the simplified Savage procedure you need to postulate a given belief structure.

the prisoner's dilemma game, but another game of their choice. A game like the one in Table 3, for example.

	Left	Right
Up	2, 2	0, 0
Down	0, 0	1, 1

Table 3

This game has *two* Nash equilibria. The transformation of Row's payoff from 3 to 0 in the Down-Left cell can be explained, for example, by Row's anticipated feelings of guilt in letting down Column, given that the latter was willing to cooperate (and vice-versa for Column in Up-Right). The PD game is then transformed in a coordination game, where Up-Left is a legitimate equilibrium.

Binmore has been a vocal supporter of this interpretation:

> [those who are anxious to deny that people seek only their own narrowly conceived selfish ends] argue that the players may care about the welfare of their opponents, or that they may actively want to keep their promises out of feelings of group solidarity or because they would otherwise suffer the pangs of a bad conscience. Such players will *not* be playing the Prisoners' Dilemma. They will be playing some other game with different payoffs. (Binmore 1992, pp. 313-314)

In other words: it is a common fallacy to interpret the payoffs in a GT model as representing physical properties or (typically) monetary outcomes. They represent utilities. And people may derive utility from (or, more precisely, may have preferences about) whatever features of the consequences they happen to care about – including other players' payoffs, one's own anticipated guilt, regret, and so forth.

This is in line with the liberal tradition that informs much of economic theory. According to this tradition, it is important that social scientists remain neutral about the contents of people's tastes and beliefs. From a normative viewpoint (and economics is a theory of rationality as well as a theory of actual behaviour) the requirements imposed on preferences and beliefs should be purely formal, of the *consistency* type. (Remember Hume: it is perfectly rational to prefer the end of the world to a scratch on my finger.) So, even though economists often assume for simplicity that outcomes are defined strictly in monetary terms, this ought not to be always the case: outcomes can (and should, whenever required) be redescribed or 'refined' to include whatever properties the agents find relevant.

This idea (the possibility of 'redefining' or 'refining' the consequences) has provoked controversy since the early days of expected utility theory.[8] An explicit attempt to

[8] A (probably incomplete) list of contributions to this debate includes Samuelson (1952), Tversky (1975), Sen (1985, 1993), Machina (1989), Bacharach (1988), Hammond (1988), Sugden (1991), Broome (1991), Anand (1992), and Munier (1996).

legitimise it is due to Peter Hammond, who puts the 'refining' strategy right at the core of his 'consequentialism'. Consequentialism, as usually understood, is the idea that rationality is forward-looking and independent of procedural considerations.

> Consequentialism requires everything which should be allowed to affect decisions to count as a relevant consequence – behavior is evaluated by its consequences, and nothing else. (Hammond 1988, p. 26)

Hammond's 'consequentialism' tries to transform genuine consequentialism into a trivial proposition:

> If regrets, sunk costs, even the structure of the decision tree itself, are relevant to [...] behavior, they are therefore already in the consequence domain. Indeed the content of a [...] theory of behavior is then largely a matter of what counts in practice as a relevant consequence, rather than whether consequentialism and other abstract axioms are satisfied. (1988, p. 26)

Many people find this problematic. It is often argued that if you are allowed to redefine consequences (or the argument of the utility functions) as you like, then economic theory becomes empty. This critique goes back a long time,[9] but is essentially misguided. It is a false dilemma to stipulate that *either* the payoffs are to be interpreted as money, *or* the theory is empty. People have preferences, and they are preferences *about* something. There's no reason a priori to assume that all individuals care about the same things. One difficult empirical task in testing GT is precisely to figure out what they care about. Once we know what they care about, we can check whether they are rational – i.e. whether the models of rational choice provide a correct account of human behaviour. The theory is not empty. It can be used to say something substantial about the real world.

To illustrate this point, consider an analogy with classical mechanics. Newton's law of universal gravitation ($F=ma$) can be interpreted (and usually is interpreted) as a description of the relation between mass, acceleration and force once *all* the forces at work in a given system have been taken into account. As such, it seems hard to falsify. Imagine a fictional classical (pre-relativity) physicist observing a prima-facie violation of the law; before coming to any conclusion about the validity of the universal law of gravitation, she would begin to search for some unobserved forces at work on the system under study. Such forces could be gravitational, but also of some other kind, not (yet) described by Newtonian mechanics (electromagnetic forces, for example). But the physicist's behaviour wouldn't make the law empty: the law says that *once* all the forces have been taken into account, they obey the universal law of gravitation – and in some cases we might well be wrong in computing the total force in the system.[10]

[9] Cf. e.g. Hutchison (1938, pp. 114-115); for recent versions, see Machina (1989, pp. 83-87); Rosenberg (1992, Ch. 5); and Anand (1993, pp. 105-106).

[10] This fictional example is not far from historical reality, as shown by Lakatos (1970), Putnam (1974) and others.

So, there is a division of labour in classical mechanics: on the one hand general principles like the law of universal gravitation are supposed to hold for all bodies and all forces. On the other, more specific laws connect specific kinds of forces to other quantities of interest. For example: gravitational forces are described by equations like $F=G(m_1m_2/r^2)$; electrostatic forces by Coulomb's law ($F=q_1q_2/4\pi\varepsilon_0r^2$), and so on. Similarly in economics one can envisage a division of labour between general principles of rational play like those of game theory, and the more specific task of connecting payoffs (utilities) with empirical reality, by specifying the argument of the utility function.

Jim Cox expresses this viewpoint clearly in a recent paper:

> In their seminal work on game theory, von Neumann and Morgenstern (1944, 1947) thought it necessary to simultaneously develop a theory of utility and a theory of play for strategic games. In contrast, much subsequent development of game theory has focused on analyzing the play of games to the exclusion of utility theory. In the absence of a focus by game theorists on utility theory, it is understandable that experimentalists testing the theory's predictions have typically assumed that agents' utilities are affine transformations of (only) their own monetary payoffs in the games. This interpretation of game theory incorporates the assumptions that agents do not care about others' (relative or absolute) material payoffs or about their intentions. There is a large experimental literature based on this special-case interpretation of the theory, which I shall subsequently refer to as the model of "self-regarding preferences." The part of the literature concerned with public goods experiments and trust and reciprocity experiments has produced replicable patterns of inconsistency with predictions of the model of self-regarding preferences. For example, the patterns of behavior that have been observed in one-shot trust and reciprocity games are inconsistent with the subgame perfect equilibria of that model. But this does *not* imply that the observed behavior is inconsistent with game theory, which is a point that has not generally been recognized in the literature. (Cox 2004, pp. 260-261)

The message is: one cannot proceed to make any inference about the empirical status of GT without figuring out first what individuals care about. If they care about fairness, for example, it is still entirely possible that they are rational maximisers of utility – once the latter has been defined appropriately to capture fairness considerations. The definition of the utility function as ranging exclusively over one's own monetary outcomes is unnecessarily restrictive and in principle can be relaxed to allow for more varied (and interesting) individual preferences.

Problems with refinements

The possibility of refining the options doesn't make the theory empty. There are still two ways in which GT can be refuted. First, suppose that cooperators in the one-shot PD game really do care only about their own monetary outcomes (and prefer more money to less). Perhaps they do not understand the concept of dominance; perhaps they do not pay

enough attention to the structure of the game. If either of these hypotheses were correct, the theory would turn out to be false.

This is just a speculation. As a matter of fact most experimental game theorists (including Cox) agree that to assume selfish preferences over money only is too restrictive. There is more in the argument of our utility functions, and this extra content should be accommodated in the framework of GT.

This leads to the second way in which GT could turn out to be refuted. Suppose that subjects cared about something else besides their own monetary gain, but that GT wasn't flexible enough to accommodate this 'something else' within its framework. Were the bridge principles unable to account for the observations, in particular, the only option left would be to modify the theory of rational play (the 'hard core' of GT).

This possibility is rarely discussed, perhaps because of a taken-for-granted analogy between folk psychology and formal rational choice theory. According to our everyday understanding, *any* observed pattern of behaviour can in principle be 'rationalised' by imputing to the individual an appropriate – perhaps highly sophisticated – structure of beliefs and desires. The intertemporal *coherence* of beliefs and desires can also always be restored by simply modifying their contents – no matter what the individual does or is going to do next. Alexander Rosenberg is the philosopher who has most forcefully argued that "the fundamental explanatory strategy of economic theory is of a piece with that of our ordinary explanations of human action" (1992, p. 118). And the view that "no intentional action can be internally irrational" (in the above sense), has been dubbed the *Plato Principle* by Donald Davidson (1982, p. 294).

The Plato Principle however is false when applied to GT, because the formal theory of rational choice is less flexible than folk psychology, and as a consequence has more limited explanatory resources. This applies to the theory both in its normative and in its descriptive version, but given the focus of this paper, let us phrase it in a descriptive idiom. There is a tension between certain refinements of the consequences and the Savage measurement procedure. Consider the sequential prisoner's dilemma game in Figure 1 (borrowed from Hausman 2000). The first monetary prize between brackets at each end-node is **A**'s monetary payoff, the second one is **B**'s monetary payoff; the letters *f, g,* etc. are arbitrary labels assigned to the consequences. Clark and Sefton (2001) report experimental data where 40% of first-movers (**A**) play a *trust* move (Up) in a game of this kind, and second-movers (**B**) respond Left 35% of the time.

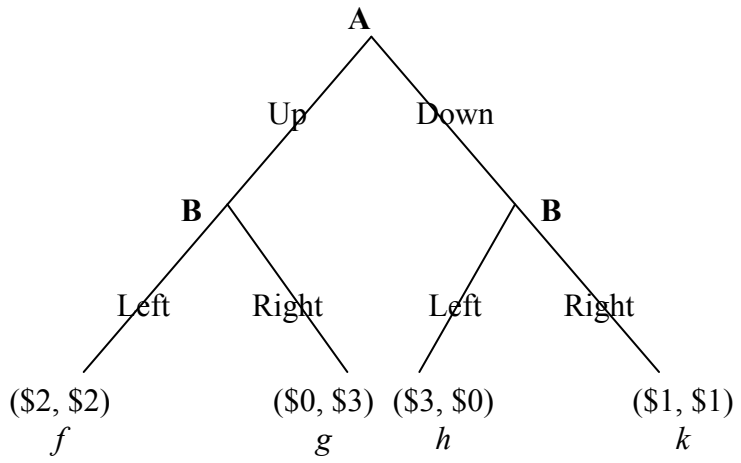

Figure 1

Can we 'refine' **A** and **B**'s preferences in order to account for this behaviour? It depends. Suppose **A** and **B** are motivated by two sets of considerations. They prefer more money to less, but also have a taste for fairness. In equation (2) deviations from fairness are defined with respect to a standard π^f:

(1) $V_i = \pi_i - \alpha_i(\left|\pi_i^f - \pi_i\right| - \sum_{j \neq i}^{n} \beta_i \left|\pi_j^f - \pi_j\right|)$.

Player's *i*'s payoffs are denoted by π_i, and the parameter α introduces *i*'s concern for the pattern of distribution of the payoffs (as opposed to his own payoffs). When α=0, the player is purely selfish. Everything on the right-hand side of α represents fairness considerations. When the standard of fairness is equality of payoffs ($\pi_i^f = \pi_j^f$), this becomes a model of inequality aversion (Levine 1998, Fehr and Schmidt 1999, Bolton and Ockenfels 2000). The parameter β weighs the impact of the first set of fairness considerations (inequality in one's payoffs) against that of the second one (inequality in other players' payoffs).

This is just one example, representative of a wider class of models of fairness preferences. Two points are worth noticing. First, these are not alternatives to the theory of rational play at the core of GT. They are attempts to pin down the bridge principles connecting GT with observed behaviour (or better: with observable features of the outcomes, like the distribution of monetary gains π_i). Such models play in the hands of the refiners, and their empirical successes in predicting experimental data are not failures of GT.

Secondly, these models add to the refiners' weaponry because they are *genuinely* consequentialist in spirit. The extensive form of the game in Figure 1 is quite redundant. The theory in (1) can be applied simply by looking at the normal form of the game (Table

2). In principle, this is fully compatible with the orthodox Savage measurement procedure. Indeed, equation (1) could be read as a generalisation of some class of preference structures observed by means of the Savage procedure.

How would this measurement work, in practice? All we have to do is to present player **B** with the choices in Figure 2:

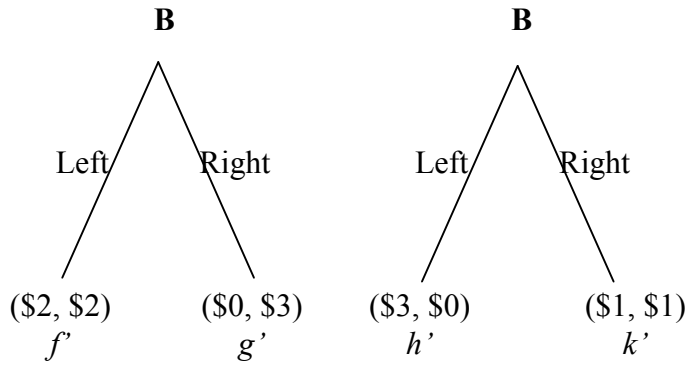

Figure 2

B's preferences, as revealed in these binary-choice tasks, will suffice to predict her behaviour in the sequential PD game, *if* they conform to the preference structure of (1). In this scenario, *f=f'*, *g=g'*, *h=h'*, and *k=k'*.

But let's suppose instead that **B** is not primarily motivated by distributional concerns. The reason she plays the equal payoff could be that she is *reciprocating* **A**'s first move. She chooses Left in response to a 'trust' move because she wants to be kind towards **A**, but has no reason to do so when **A** has not played at all. In this scenario, looking at **B**'s choices in Figure 2 will be useless (see also Hausman 2000, p. 108, for this point).

Modelling reciprocity

Reciprocity is a tricky notion. Consider the *basic reciprocity formula*:

(2) $V_i = \pi_i + \rho[\kappa(.)\sigma(.)]$.

Here ρ is a parameter weighing the impact of reciprocity considerations against the usual self-interested preference for one's own monetary gains (π_i). κ is a *kindness function*, representing how 'nice' or 'offensive' *j*'s behaviour is towards *i*. The other function, σ, captures *i*'s reaction to *j*'s behaviour. A *negative* kindness combined with a *positive* reaction (or vice-versa) results in a reduction of overall utility (V_i). Reacting negatively to negative kindness, and positively to positive kindness, leads to an increase of overall utility.

At first sight, it may look as if the arguments of these two functions could be filled in consistently with the spirit of genuine consequentialism. But this is unlikely. Consider the reaction function first: a negative reaction is one that *hurts j*, typically by reducing her payoff with respect to a given standard or reference point.The standard is most naturally interpreted as what *j would have* achieved, had *i* played a different strategy. We shall represent this possible payoff with π'_j . Similarly, *j*'s kindness towards *i* depends not only on *i*'s relative payoff with respect to some absolute standard (say, equality, as in most models of 'distributional' preferences), but on what *i could* have achieved, had *j* played differently. Again, we will use π'_i to represent such counterfactual payoff, where the counterfactual mark (′) refers to the fact that it is a *possible* payoff – an outcome that could in principle be achieved in *this* game.[11]

(3) $$V_i = \pi_i + \rho[\kappa(\pi_i - \pi'_i)\sigma(\pi_j - \pi'_j)]\,.$$

The most advanced models of reciprocity in this spirit make use of the tools provided by *psychological game theory* (first proposed by Geannakoplos, Pearce and Stacchetti 1989). In psychological game theory the payoffs depend not just on the material structure of the consequences, but also on players' *beliefs*. Following Matthew Rabin's (1993) seminal work, reciprocity can be represented as involving a calculation of what player *i* believes that *j* believes that *i* will play. Player *j*'s intention is perceived as 'nasty' when *i* believes that *j* believes that *i* will make a 'nice' move, and yet *j* chooses a strategy that reduces *i*'s monetary payoffs with respect to the reference point.[12]

A discussion of psychological GT is besides the scope of this paper. How counterfactual considerations are modelled is not crucial: perhaps intentions are not even relevant *all* the time; perhaps in some cases **B** reacts to **A**'s 'nasty' moves only to signal that she's making a mistake, or in order to push the outcome towards a Pareto-superior solution (**B** might be driven by selfish motives, that is, and might have a crude stimulus-response model of **A**'s behaviour).[13] What really matters is the role of counterfactual considerations themselves. The identification of the preferences associated with each consequence requires an examination not just of what will happen, but also of *what might have happened*. To put it another way, the utilities are path-dependent: they do not just depend on the outcomes taken in isolation, but on the *whole structure of the game*.

Reciprocity cannot be 'refined'

Remember Hammond: "Consequentialism requires everything which should be allowed to affect decisions to count as a relevant consequence"; if reciprocity affects decisions, it

[11] I'm simplifying a lot here. The counterfactual 'reference point' in $k(\pi_i - \pi'_i)$ is in many occasions better represented as π_i^f, where f stands for the 'fairest' payoff that could have been achieved in the game. It is likely that π_i^f be intendified in each game by comparing *i* and *j*'s various possible payoffs.

[12] Falk and Fischbacher (2000) and Dufwemberg and Kirchsteiger (2003) are the two prominent models in this tradition.

[13] As a matter of fact, in some dilemma experiments some subjects 'reciprocate' even when they are playing against a computer (Houser and Kurzban 2002).

must be accounted for at the moment of describing the consequences, according to consequentialism.

This turns out to be problematic in the Savage framework, however. The problem has been around for a long time. In a letter from Robert Aumann to Leo Savage, in 1971 (cf. Dreze 1987), the issue is first raised of how to distinguish *acts* (say, getting an umbrella) from *consequences* (e.g. getting wet). Aumann points out that the distinction is fuzzy. Savage agrees but says in a very pragmatic vein that this may not matter as long as "I seem to be able to couch my decision problems in terms of them"; for "... a consequence is in the last analysis an experience" (Dreze 1987, p. 79).

But this is overoptimistic. Savage (1954) defines *acts* as functions from *states of the world* to *consequences*. (For example: the act 'taking the umbrella' assigns the consequence 'staying dry' both to the state 'it rains' and to the state 'it does not rain'; 'not taking the umbrella' assigns the consequence 'getting wet' to the state 'it rains' and the consequence 'staying dry' to the state 'it does not rain'.) All the existing versions of expected utility theory prove the existence of the EU function (representation theorem) from the assumption that decision makers form preferences over *every* act that can possibly be constructed by combining consequences with states of the world. John Broome (1991) calls this feature the *rectangular field assumption* (see also Sugden 1991).

The rectangular field assumption is not *implied* by expected utility theory, from a logical point of view. But the existence of the expected utility function can be derived from the rectangular field assumption and other fundamental principles (Broome 1991, p. 117). So a violation of the rectangular field assumption does not logically contradict expected utility theory (although we still lack an alternative representation theorem that does without that assumption). It does, however, seriously invalidate the Savage measurement procedure.

For measurement purposes it is crucial that acts can be constructed arbitrarily. Take any state of the world, and combine it with any consequence in our theoretical universe: whatever we obtain by these arbitrary combinations must make sense, and must be something the agent can form a preference about. Now suppose that the consequence *g* of player **B** in the sequential PD game of Figure 1 is defined as 'getting $3 but being sorry for player **A**' (who had offered to cooperate). This is surely meaningful in the context of a game like the one above. But does it make sense in a different context, where that consequence does *not* follow from the act of choosing Right in the above game (after **A** has chosen Up)?

We must distinguish two cases (Broome 1991, p. 116):

(1) The (refined) consequence is *causally implausible* outside the specific game; for example, why should **B** feel sorry for **A**, if they are not playing the game of Figure 1?
(2) The (refined) consequence is *logically impossible* outside the specific game; for example, imagine we redefined *g* as 'getting $3 after player **A** has chosen Up': clearly this is inconsistent with the game that **B** is playing in, say, Figure 2.

To sum up: when too much is included in the description of outcomes, the consequence itself remains tied to the specific game and cannot be used to construct arbitrarily other acts (or functions from states of the world to consequences) (Sugden 1991; see also Munier 1996, and Verbeek 2001). Refining options to include reciprocity considerations clearly raises problems of the second (logical) kind. If players care about 'what might have been', then this must be included in the definition of a consequence. Indeed the consequences must be refined so as to include entire branches of the game. But then one cannot use such consequences to create arbitrarily other acts, if such acts are to make sense at all. The Savage measurement procedure is not flexible enough to neutralise reciprocity counterexamples.[14]

Evidence of reciprocity

Reciprocity has been discussed by experimental game theorists for more than two decades now. Most of this discussion however has taken place *before* the existence of reciprocal motivations had been established empirically. Consider social dilemma games again. As already mentioned, many subjects seem willing to cooperate in a one-shot sequential PD game. But this behaviour of course does not constitute direct evidence in favour of reciprocity, because the mutual cooperation outcome provides an *equal* distribution of the material payoffs (see Figure 1 above) and therefore might be preferred by players that are averse to inequality.

A crucial test for reciprocity must implement some variant of the Savage measurement procedure. Players' preferences over a set of outcomes must be observed *first* 'in isolation', and *then* in the context of a game that is likely to trigger reciprocity considerations (or more generally a concern for 'what might have been'). The comparison between choices in these two contexts will provide direct evidence for reciprocity as well as a proper test of GT. Several experimenters have done precisely this in recent years, although experimental game theorists themselves do not seem to have noticed the general implications of these results.

The experiments usually start with so-called Dictator's Games. A Dictator's Game (Forsythe, Horowitz, Savin and Sefton 1994) is not even a strategic game, strictly speaking, for it involves a straight choice between different allocations. The outcome is not determined by an interaction between the players: one player chooses, and the other takes whatever has been chosen. A Dictator's Game then, like a stripped-down version of the Savage measurement procedure, simply elicits the (ordinal) structure of preferences of the experimental subjects. Since the ordering, rather than the cardinality, of preferences is what matters in many simple experimental games, the full-blown Savage method is unnecessary.

[14] Regret is another motivation that can lead to problems with the rectangular field assumption (see Sugden 1986), also outside the realm of game theory, in individual decision making and rational choice theory in general. Unlike reciprocity, however, we don't have very good evidence about regret right now.

Once preferences have been elicited in the Dictator's Game, it is then possible to check whether players behave rationally in more complex strategic situations. Take the sequential Prisoner's Dilemma game. Jim Cox (2004) uses a slightly more complex version of this setting, called the 'investment game'. Player **A** here chooses how much of a given endowment to pass over to **B**. Whatever **A** chooses, the sum will be tripled by the experimenter. Player **B** then decides how much of this 'investment' to return to **A**.

After the game is played, Cox offers some subjects exactly the same sums which have been selected by players of type **A** in the first experiment – except that type-**B** players this time do not know that it's an 'investment' from another player (the sum is presented as a 'pie from the sky'). The experimenter can therefore check whether the behaviour of **B** players in this Dictator's Game is radically different from their behaviour in the investment context. As a matter of fact it is: in the investment task the average amount returned by **B** players is more than twice the amount allocated to **A** in the Dictator's task ($4.94 vs. $2.06).

Gary Charness and Matthew Rabin (2002) use even simpler settings (Figure 3). The first part of the experiment (on the left) is again a Dictator's Game. The first payoff between brackets is **A**'s monetary gain and the second one is **B**'s, as usual. The figures under each outcome represent the observed distribution of player **B**s' choices. A majority of subjects goes for the Pareto-superior outcome, but there is significant evidence of inequality aversion too.

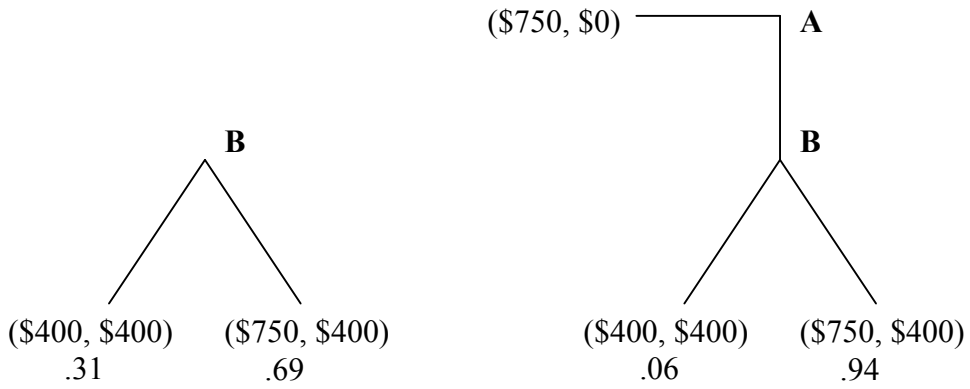

Figure 3

The same (sub-)game is then transferred in a different context, where **B** is asked to make the same choice *after* player **A** has already made a move (Figure 3, on the right). As we can see, choices in this second setting can differ radically from those in the simple Dictator's Game. Concerns for equality in this specific case seem to be largely swept away by **B**'s desire to respond to **A**'s 'kindness' (**A** could have chosen unilaterally to give **B** absolutely nothing, after all).

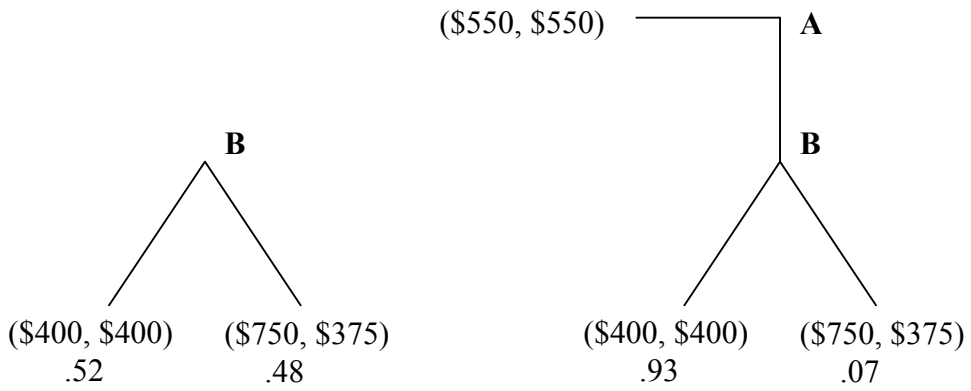

Figure 4

Figure 7 presents another example of the importance of 'what might have been'. In the dictator's game (left), **B** players display a considerable willingness to sacrifice a small part of their material payoff to enhance **A**'s income. This willingness disappears in the enhanced game (right), where **B** moves knowing that **A** has refused the (550, 550) split (a phenomenon that Charness and Rabin dub 'concern withdrawal').

These results are confirmed by experiments conducted by Armin Falk, Ernst Fehr, and Urs Fischbacher (2003). Here, one 'branch' of a mini-ultimatum game is kept constant, while the alternatives open to **A** (the first-mover) are varied. In two of the experimental tasks (represented in Figure 5) the difference in **B**'s responses is remarkable: the rejection rate of the ($8, $2) offer jumps from .46 to .91 depending on whether the alternative possible offers are perceived as 'nice' ($5, $5) or 'nasty' ($10, $0).[15]

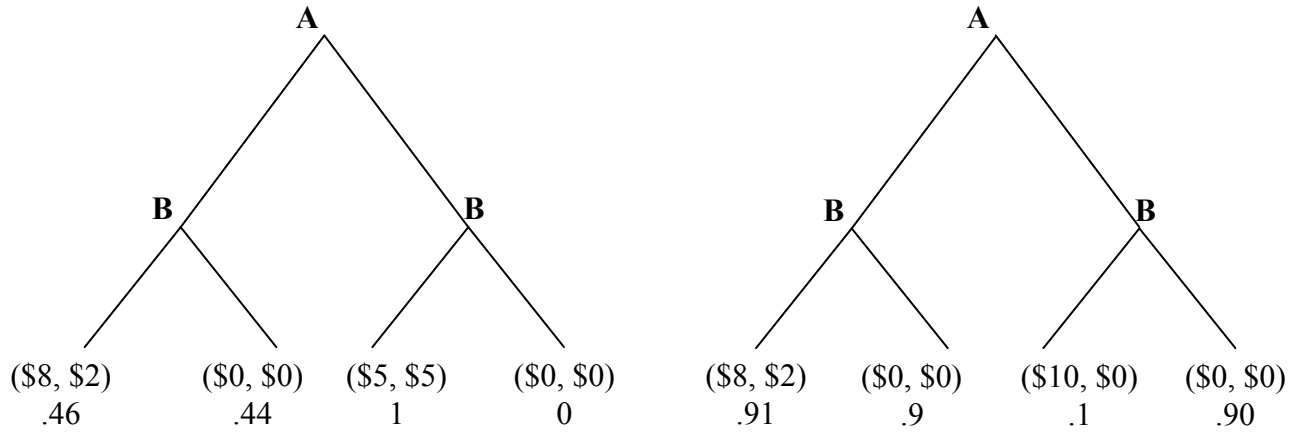

Figure 5

[15] The '$' sign here is used simply to mark that the payoffs are to be read as monetary gains, not as utilitities. Some experiments, in fact, used other currencies than dollars.

It's important to stress that these patterns can be generalised only with much difficulty. Reciprocity, if this is really the motivation behind these observations, is an erratic phenomenon. Charness (1996) reports no significant difference in a gift-exchange game played with intentional offers vs. a gift-exchange with randomly generated offers. People are as willing to 'return' the gift in the random game as in the 'intentional' context. Andreoni and Miller (2000) show that choices in a Dictator's Game context can be used to predict with a reasonable approximation the behaviour of subjects playing a Public Goods' experiment. Bolton, Brandts and Ockenfels (1998) find no difference between behaviour in social dilemmas with 'forced' first moves vs. behaviour in the same setting with voluntary moves. People's cooperative attitude seems to be the same regardless of the intentionality of the first mover's choice. McCabe, Rigdon and Smith (2003) however report the opposite result in a similar experiment.[16]

These results seem to be very sensitive to the details of the design, and reciprocity is an elusive and complex phenomenon. But minimally, it is possible to say that in a number of settings *the structure of the game matters*. Subjects do not just care about the material consequences but also about the alternative moves that are available and the process that leads to each consequence. They care about what might have been and how we got where we are now.

Concluding remarks

GT is the piece of contemporary economics which has had the most dramatic impact outside the boundaries of the discipline. Sixty years from its inception, GT is widely used in fields as different as political science, sociology, biology, psychology, law, military strategy and, last but not least, philosophy. Its empirical status, then, is an issue of wide scientific and philosophical relevance.

For many years a considerable degree of confusion about the status of GT was created by the proliferation of anomalous but indecisive evidence. Now that experiments of the right kind have been done – experiments that really test GT – it has become impossible to argue that the anomalies can be digested as usual. In order to see that, it is necessary to appreciate the crucial distinction between the theory of rational play and the bridge principles that connect the former with measurable properties of the real world. These two elements are sometimes conflated, sometimes treated as if they were completely independent. This probably leads to the (mistaken) belief that GT is as flexible an explanatory tool as the folk-psychological theory of action of which it constitutes a formal refinement.

The Savage procedure (and the rectangular field assumption upon which it is based) imposes some limits on the amount of information that can be incorporated in the

[16] Bicchieri (2005) argues, convincingly in my view, that these 'contradictory' results can be explained by the fact that social norms (like reciprocity) are highly context-sensitive and usually must be 'triggered' or 'cued' by some factor telling us that 'this is a context where the norm applies', i.e. where it is generally expected that people behave according to the norm.

description of the outcomes of a game. These technical limitations make sure that the spirit of genuine consequentialism (players are forward looking, subscribe to sub-game perfect rationality, and so on) is preserved.

It's important to appreciate the precise implications of what just said: nothing prevents us from applying the tools of GT and EUT across a limited range of experimental games, assuming the subjects have consistent preferences over a limited number of highly specific (or highly refined) options, once the outcomes have been (re-)described so as to include the overall structure of the game.[17] Such preference information might be useless outside the specific game-situation, of course, but could be used to predict and summarise choices in a limited set of games. What can't be done, if consequences become overly sophisticated, is to have a unique utility function across the set of *all* the games that can be constructed using those (and other) consequences.

From this respect, the 'naive' and often ridiculed convention of using monetary payoffs as initial conditions or proxies of players' utilities, has much to recommend. It is well known among historians and philosophers of science that the success of scientific theories is often heavily dependent on the creation of instruments of observation that allow adequate measurements of the key variables. Consider Galileo's telescope, but also the microscope for the atomistic research programme, the steam engine for thermodynamics, and so forth. Monetary payoffs are objective and observable, and provided we have the right theory at hands, can be used readily to generalise to other situations that are similar (in the monetary payoffs structure) to those that were tested in the laboratory. If this route is followed of course the 'core' of GT must be modified to account for the existing evidence. The partnership between the standard theory of rational play and the Savage procedure has proven to be unsustainable (given the evidence). We can keep the core of GT and reject the bridge principles (the Savage procedure), but face difficult problems of applicability. The other viable alternative is to keep the bridge and change the theory of play.

Why did it take such a long time to design and run the 'right' experiments to test GT? One reason is that, surprisingly, the Savage procedure was never part of the 'official' toolkit of the experimental game theorist. This is the result of complicated historical factors, in particular the fact that the methodology of experimental economics (especially so-called 'Induced Value Theory' – cf. Smith 1982) was developed mainly in the context of market experiments and was aimed at the active *inducement*, rather than the passive observation of preferences. But a proper discussion of these issues would require a long

[17] The Savage procedure prescribes *first* to measure the utility of the outcomes of a game in isolation and *then* to see how the subjects behave in more complex strategic situations (games). The rectangular field assumption is required if one wants to elicit utility measures for a wide range of consequences, or in other words if one is trying to specify the fine-grained structure of an individual's utility function. But the application of GT in each specific instance of course does not require this full-blown measurement to be carried out in practice. In fact, as we have seen, the simple games that are customarily tested in the laboratory often require only a simplified or stripped-down version of the Savage procedure – because the mere ordinal ranking of a handful of outcomes is what really matters.

detour into the history and methodology of experimental game theory, which cannot be attempted here.[18]

Interestingly, the refutation of GT took place simultaneously with the theory's most remarkable successes in policy-making. The reform of inefficient (usually centralised) methods of allocation and the design of new markets (for electricity, telecommunication, health services, etc.) is now increasingly informed by the principles of GT. These principles moreover are often subjected to preliminary testing in controlled conditions before being applied to real-world situations. These tests, however, are usually aimed at spotting and avoiding precisely the sort of problematic situations that the theory is unable to deal with. Other-regarding preferences for instance can simply be ruled out by design, by making sure that the available information and the structure of the game prevent such concerns from arising in the first place. This is economic engineering, as opposed to theory-testing: the world is modified so as to 'fit' the conditions that allow the theory to work. But since engineering successes often *do* drive science, we should not rush to any conclusion about the future of GT. Its refutation and its remarkable success can (and will) coexist in the minds of many social scientists for whom the theory is, and is likely to remain, the best game in town.

[18] See Guala (2005), however, especially Ch. 11.

References

Anand, Paul. 1993. *Foundations of Rational Choice under Risk*. Oxford: Oxford University Press.

Andreoni, James and Miller, John. 2000. "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism". *Econometrica* 70: 737-53.

Bicchieri, Cristina. 2005. *The Grammar of Society*. Cambridge: Cambridge University Press.

Binmore, Ken. 1992 *Fun and Games: A Text on Game Theory*. Lexington Mass.: D.C. Heat & Co.

Binmore, Ken. 1994. *Game Theory and the Social Contract. Vol. 1: Playing Fair.* Cambridge, Mass: MIT Press.

Binmore, Ken. 1999. "Why Experiment in Economics?" *Economic Journal* 109: F16--F24.

Bacharach, Michael. 1988. "Preferenze razionali e descrizioni", in M.C. Galavotti and G. Gambetta (eds.) *Epistemologia ed economia*. Bologna: CLUEB.

Bolton, Gary E. and Ockenfels, Axel. 2000. "ERC: A Theory of Equity, Reciprocity and Cooperation". *American Economic Review*. 90: 166-93.

Bolton, Gary E., Brandts, Jordi and Ockenfels, Axel. 1998. "Measuring Motivations for the Reciprocal Responses Observed in a Simple Dilemma Game". *Experimental Economics* 1: 207-19.

Broome, John. 1991. *Weighing Goods*. London: Blackwell.

Charness, Gary. 1996. "Attribution and Reciprocity in an Experimental Labour Market". Unpublished Paper.

Charness, Gary and Rabin, Matthew. 2002. "Understanding Social Preferences with Simple Tests". *Quarterly Journal of Economics*

Clark, Kenneth and Sefton, Martin. 2001. "The Sequential Prisoner's Dilemma: Evidence on Reciprocation", *Economic Journal* 111: 51-68.

Cox, James C. 2004. "How to Identify Trust and Reciprocity." *Games and Economic Behavior* 46: 260-81.

Davidson, Donald. 1982. "Paradoxes of Irrationality", in R. Wollheim and J. Hopkins (eds.) *Philosophical Essays on Freud*. Cambridge: Cambridge University Press.

Dowding, Keith. 2002. "Revealed Preference and External Reference". *Rationality and Society* 14: 259-84.

Dreze, Jacques. 1987. *Essays on Economic Decisions Under Uncertainty*. Cambridge: Cambridge University Press.

Dufwenberg, Martin and Kirchsteiger, Georg. 2003."A Theory of Sequential Reciprocity". *Games and Economic Behavior* 47: 268-98.

Falk, Armin, Fehr, Ernst, and Fischbacher, Urs. 2003. "On the Nature of Fair Behaviour", *Economic Inquiry* 41: 20-6.

Falk, Armin and Fischbacher, Urs. 2000. "A Theory of Reciprocity". Working Paper no. 6. Institute for Empirical Research in Economics, University of Zurich.

Fehr, Ernst and Schmidt, Klaus M. 1999. "A Theory of Fairness, Competition and Cooperation". *Quarterly Journal of Economics*

Forsythe R., Horowitz J.L., Savin N.E. and Sefton M. 1994. "Fairness in Simple Bargaining Experiments". *Games and Economic Behavior* 6: 347-69.

Geannakoplos, John, Pearce, David and Stacchetti, Ennio. 1989. "Psychological Games and Sequential Rationality". *Games and Economic Behavior* 1: 60-79.
Giere, Ronald. 1988. *Explaining Science.* Chicago: University of Chicago Press.
Guala, Francesco. 2000. "The Logic of Normative Falsification: Rationality and Experiments in Decision Theory". *Journal of Economic Methodology* 7: 59-93.
Guala, Francesco. 2005. *The Methodology of Experimental Economics*. Cambridge: Cambridge University Press.
Hammond, Peter. 1988. "Consequentialist Foundations for Expected Utility Theory". *Theory and Decision* 25: 25-78.
Hausman, Daniel M. 2000. "Revealed Preference, Belief, and Game Theory." *Economics and Philosophy* 16: 99-115.
Houser, Daniel and Kurzban, Robert. 2002. "Revisiting Kindness and Confusion in Public Goods Experiments". *American Economic Review* 92: 1062-9.
Hutchison, Terence W. 1938. *The Significance and Basic Postulates of Economic Theory*. New York: Kelley.
Lakatos, Imre. 1970. "Falsificationism and the Methodology of Scientific Research Programmes". In *Philosophical Papers*, *Vol. 1*. Cambridge: Cambridge University Press.
Levine, David K. 1998. "Modeling Altruism and Spitefulness in Experiments". *Review of Economic Dynamics* 1: 593-622.
McCabe, Kevin A., Rigdon, Mary L. and Smith, Vernon L. 2003. "Positive Reciprocity and Intentions in Trust Games." *Journal of Economic Behavior and Organization* 52: 267-75.
Machina, Mark. 1989. "Dynamic Consistency and Non-Expected Utility Models of Choice Under Risk". *Journal of Economic Literature* 27: 1622-68.
Munier, Bertrand. 1996. "Comment", in K.J. Arrow, E. Colombatto, M. Perlman, and C. Schmidt (eds.) *The Rational Foundations of Economic Behavior*. London: Macmillan.
Osborne, Martin J. and Rubinstein, Ariel. 1994. *A Course in Game Theory*. Cambridge, Mass: MIT Press.
Putnam, Hilary. 1974. "The 'Corroboration' of Theories". In P.A. Schilpp (ed.) *The Philosophy of Karl Popper*. La Salle, Ill.: Open Court.
Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics". *American Economic Review*. 83: 1281-302.
Ramsey, Frank. 1926. "Truth and Probability". In R.B. Braithwaite (ed.) *The Foundations of Mathematics and Other Logical Essays*. London: Routledge, 1931.
Rapoport, A. and Chammah, A.M. 1965. *Prisoner's Dilemma: A Study in Conflict and Cooperation*. Ann Arbour: University of Michigan Press.
Rosenberg, Alexander. 1992. *Economics – Mathematical Politics or Science of Diminishing Returns?* Chicago: University of Chicago Press.
Samuelson, Paul A. 1938. "A Note on the Pure Theory of Consumer's Behavior" .*Economica* 5: 61–71
Samuelson, Paul A. 1952. "Utility, Preference and Probability". In J. Stigliz (ed.) *The Collected Papers of Paul A. Samuelson, Vol. 1*. Cambridge: MIT Press.
Savage, Leonard. 1954. *The Foundations of Statistics*. New York: Wiley.

Sen, Amartya K. 1977. “Rational Fools: A Critique of the Behavioural Foundations of Economic Theory”, *Philosophy and Public Affairs* 6: 317-44.
Sen, Amartya K. 1985. “Rationality and Uncertainty”. *Theory and Decision* 18: 109-27.
Sen, Amartya K. 1993. “Internal Consistency of Choice”. *Econometrica* 61: 495–521.
Smith, Vernon L. 1982. “Microeconomic Systems as an Experimental Science”. *American Economic Review* 72: 923–55.
Sober, Elliott. 1993. *Philosophy of Biology*. Oxford: Oxford University Press.
Starmer, Chris. 2000. “Developments in Non-expected Utility Theory: The Hunt for a Descriptive Theory of Choice under Risk”. *Journal of Economic Literature* 38: 332-82.
Sugden, Robert. 1986. “Regret, Recrimination and Rationality”. *Theory and Decision* 19: 77-99.
Sugden, Robert. 1991. “Rational Choice: A Survey of Contributions from Economics and Philosophy.” *Economic Journal* 101: 751-785
Tversky, Amos. 1975. “A Critique of Expected Utility Theory: Descriptive and Normative Issues”. *Erkenntnis* 9: 163-73.
Verbeek, Bruno. 2001. “Consequentialism, Rationality, and the Relevant Description of Outcomes”. *Economics and Philosophy* 17:181-205.
von Neumann, John and Morgenstern, Oskar. 1944. *The Theory of Games and Economic Behavior*. Princeton: Princeton University Press; 2nd edition 1947.