Homo Oeconomicus 28(3): 265–296 • (2011)

www.accedoverlag.de

The Revealed-Preference Interpretation of Payoffs in Game Theory

Aki Lehtinen
Social and Moral Philosophy Unit, Department of Political and Economic Studies, University of Helsinki, Finland
(eMail: aki.lehtinen@helsinki.fi)

Abstract There are two different ways of interpreting the idea that payoffs in games are revealed preferences. One could argue, first, that choices define payoffs in the sense that preferences and choices are conceptually tied to each other, and secondly, that there is an experimental procedure that could be used for finding out players' payoffs. It will be shown that, in game theory, the first interpretation is conceptually impossible and the second is misleading. It is thus argued that payoffs in games should be interpreted as representations of players' preferences rather than revealed preferences. It is also shown that, whereas it is reasonable to assume that players choose the non-cooperative strategy in a Prisoner's Dilemma game because the argument for choosing strongly dominant strategies is compelling, a revealed-preference argument does not support such a choice.

Keywords game theory, revealed preferences, analysing and modelling games, preference elicitation, Binmore

1. Introduction

Some economists argue that payoffs in games are to be interpreted as revealed preferences. A similar idea is expressed in the notion that the payoffs are what they are, in other words they are given in the sense that it would be methodologically illegitimate to tamper with them while analysing the game. The methodological point is that the analysis of the game and the construction of the payoff matrix should be kept separate. Ken Binmore presents both arguments in *Playing Fair* (1994).

ISBN 978-3-89265-077-5 ISSN 0943-0180

I argue in this paper that payoffs cannot be interpreted as revealed preferences in games. I will focus mainly on Binmore's views because he has explicitly and repeatedly declared that he entertains a revealed-preference interpretation of payoffs. I also argue that Binmore violates his own injunction to keep the game fixed while analysing the Prisoner's Dilemma, and that his contention that players choose the non-cooperative strategy as a matter of revealed preference is faulty.

Daniel Hausman (2000) suggests two ways of interpreting the notion of revealed preferences that could be employed for interpreting payoffs in games. The first is that the formal apparatus makes it possible for economists to *dispense with the notion of preference* altogether. If *choices define preferences* it is not possible to make a conceptual distinction between the two (see also Sen, 1973, 1987), which is why dispensing with the notion altogether is sometimes equated with this idea. One could thus argue that choices define payoffs in the sense that preferences and choices are conceptually tied to each other. The second interpretation is that the theory shows how *choices reveal preferences* and how to test *claims about preferences* (Grüne, 2004, see also Hausman, 2005b). This could be taken to mean that revealed preferences provide an empirically respectable way of finding out players' preferences. The idea is that it is possible, at least in principle, to set up a preference-elicitation procedure (the reference lottery technique or similar) for finding out players' payoffs. If it were possible to define preferences with choices only, it would also be possible to do without any psychological assumptions. This third interpretation motivates Binmore too.

However, players cannot define their preferences by making strategy choices in a game. The reason, in short, is that they are not given the opportunity to make pair-wise choices from among all the outcomes while they are playing, and we can only observe their equilibrium choices and the associated outcomes. Furthermore, I argue that the second interpretation is misleading because payoffs are not defined by the players' choices in an elicitation procedure.

This paper is structured as follows. Section 2 focuses on the distinction between modelling and analysing games. I discuss Binmore's account of payoffs in game theory in Section 3, and suggest that contrary to his claim, although players should play non-cooperatively in the Prisoner's Dilemma game, this conclusion cannot reasonably be based on a revealed-preference argument. Section 4 describes how a revealed-preference interpretation of payoffs meshes with the practice of testing game theory. The discussion in Section 5 concerns the role of revealed-preference arguments and conditions in an analysis of Sen's famous example of apple-picking, which Binmore

criticises. In Section 6, I consider what denying a revealed-preference interpretation in game theory entails by comparing it to the alternative of viewing payoffs as underlying choices. Section 7 concludes the paper.

2. Modelling and Analysing Games

Binmore (1994, pp. 27, 161-2, 169) distinguishes between *modelling* and *analysing* a game in the same book in which he started using the revealed-preference argument. Modelling a game means constructing a payoff structure, and analysing it means deriving a prediction of the players' choices on the basis of their given preferences and a solution concept. Modelling thus involves specifying the players' payoffs, the strategies available to them, and their information sets, whereas analysing entails deriving equilibrium strategies on the basis of a solution concept – without this nothing would be left for analysis.

One reason for distinguishing between modelling and analysing is to force the game theorist to include all possible motivating factors in the payoffs. If the payoffs contain all this information, it precludes the introduction of new psychological variables during the analysis. Binmore (1994, pp. 161-2) argues, for example, that Sen's reasoning on sympathy and commitment should be written into the payoffs of a game, in other words it should be taken into account in the modelling.

The point with the distinction is thus that game-theoretical analyses should not be criticised by invoking issues that properly belong to modelling. This is clearly a reasonable requirement. Indeed, Binmore's point is not new in the discussion. Game theorists and decision theorists 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, Rubinstein, 1991).

The argument that all motivating factors should be incorporated into the payoffs is related to revealed preferences as follows: if payoffs were obtained by observing the players' choices, they would incorporate information about all the motivating factors because, *by assumption*, all these factors affect their choices. Thus, if payoffs were actually thus obtained, there would no longer be any need to invoke the distinction between modelling and analysing. Given that Binmore seems to emphasise revealed preferences rather than the modelling-analysing distinction in his recent writings, he may well have drawn precisely this conclusion. On the other hand, there is another interpretation of the distinction: it could be argued that it does not imply that payoffs are revealed by players' choices. The idea is rather that,

however they are obtained, once they are fixed in modelling they are not to be tampered with when the game is being analysed.

It seems obvious that a revealed-preference interpretation must concern modelling rather than analysing. If such an argument is used in analysing a game, and if it is successful, it makes the very notion of analysing the game redundant. If payoffs were to be defined by the players' choices in the game there would be no need to analyse what they would choose in equilibrium. The game theorist would already have analysed the game by modelling it (see also Hausman, 2000). Thus, if the revealed-preference interpretation is to have some role in the distinction between analysing and modelling, it would have to imply that games are to be modelled by finding out the players' preferences through observing their choices.

Binmore and other prominent game theorists seem to believe that modelling is difficult. Rubinstein, for example, states that determining which factors to include in a game-theoretical model requires 'intuition, common sense and empirical data' (Rubinstein, 1991, p. 919). It is not very surprising, then, that the practice of game theory is not, in fact, to elicit the preferences (Weibull, 2004).[1] Game theorists usually simply postulate a payoff structure. Binmore's point, however, is not to change this state of affairs by arguing that one should actually elicit the preferences for games, but rather to regiment the way in which game-theoretical analyses are conducted. This supports his argument (Binmore, 1994, pp. 98, 165) that a revealed-preference interpretation of payoffs makes it very difficult to know how to choose payoffs to represent some real-world situation. The implication is that Binmore's revealed-preference interpretation is motivated by the same kind of methodological considerations as the distinction between analysing and modelling games rather than by a desire to change the way in which games are currently modelled or to encourage their empirical analysis (cf. Sugden, 2001).

3. Binmore's Account of Revealed Preferences in Games

This is how Binmore conceives of the revealed-preference interpretation:

> Modern utility theory makes tautology of the fact that action B will be chosen rather than A when the former yields a higher payoff by *defining* the payoff of B to be larger than the payoff of A if B is chosen when A is available (Binmore, 1994, p. 169).

[1] I know of a paper entitled 'Elicitation for games' (Kadane et al., 1992), but it concerns the elicitation of beliefs rather than preferences.

Table 1 — The Prisoner's Dilemma game

		Eve	
		Dove	*Hawk*
Adam	*Dove*	(3,3)	(1,4)
	Hawk	(4,1)	(2,2)

The idea is that if an action B is chosen over an action A, B must be more preferred than A and thereby it must have higher utility *by definition*. Note that this is different from saying that if given the choice, a rational agent is *assumed* to choose a more preferred action rather than a less preferred one.

As Binmore explains, there are three sets of concepts to consider:

> ... a set A of actions, a set B of possible states of the world, and a set C of final consequences. These are connected by a function f: $A \times B \rightarrow C$ that describes the consequence c = f(a,b) of taking action a when the state of the world is b. Orthodox *revealed preference* theory then provides consistency conditions for Eve's behavior in A to be described by saying that she chooses *as though* maximizing the expected value of a utility function defined on C, relative to a subjective probability distribution defined on B. (Binmore, 1998, pp. 360-1) (my emphasis)[2]

The traditional view of payoffs in game theory is that they are von Neumann-Morgenstern (vNM) utilities,[3] Binmore follows this tradition with the Savagean twist that beliefs are subjective rather than objective. He does not seem to differentiate between revealed preference theory in circumstances of certainty (e.g., Samuelson, 1938; Houthakker, 1950; Arrow, 1959) and in circumstances of uncertainty.[4]

Binmore is particularly concerned to show that players cannot cooperate in a Prisoner's Dilemma (PD) because of the way their preferences are defined (e.g., Binmore, 2007b, pp. 13–15, 2009, pp. 26–29). Table 1 shows this game.

[2] See also Binmore (1994, p. 97).

[3] But see Mariotti (1995, 1996, 1997), and the discussion in Battigalli (1996) and Hammond (1996).

[4] The Samuelsonian revealed preference theory concerning choices under certainty has been interpreted in various ways. Stanley Wong (1978) argued that Samuelson gave three different interpretations: the purpose of the Samuelson (1938) paper was to derive the results of ordinal utility theory (demand theory) without recourse to unobservables; revealed preference theory was used in Samuelson (1948) as a solution to the problem of constructing an individual's indifference map; and in Samuelson (1950) the purpose was to 'find the full empirical implications of ordinal utility theory' (p. 369). See Hands (2012) for an account of how current revealed preference theorists differ from earlier ones.

I take Binmore to be claiming that since payoffs are to be interpreted as revealed preferences, arguing that players might choose the cooperative strategy (*Dove*) in a one-shot PD game is tantamount to changing the game, and thereby not a proper analysis of it. Note how the idea that the payoffs must be fixed in analysing the game is intertwined with revealed-preference arguments.

> In the one-shot Prisoner's Dilemma, Eve's utility for the outcome (*Dove*, *Hawk*) is made larger than her utility for the outcome (*Dove*, *Dove*) because we are given that she would choose *Hawk* if she knew that Adam were sure to choose *Dove* ... Once this is understood, it becomes obvious that all the endless disputation over the standard game-theoretic analysis of the Prisoner's Dilemma is based on the simplest of misunderstandings. (Binmore, 1998, p. 360 fn.)

The idea is that if both players were presented with a choice between *Hawk* and *Dove*, each would choose *Hawk* when the other player chose *Hawk*, and also when the other player chose *Dove*. There are two ways of formulating this argument.

1) Players choose *Hawk* because their payoffs are defined to be higher if they choose *Hawk* than if they choose *Dove* irrespective of what the other player chooses, and they are assumed to maximise their payoffs.

2) Since choices define preferences according to revealed preference theory, the payoffs for *Hawk* are larger than those for *Dove* irrespective of what the other player chooses *because the players choose Hawk* if they are in a Prisoner's Dilemma.

Unlike the first argument, the second one appeals to revealed preferences. The difference between the two is that in the former the players *choose Hawk because* their *payoffs are defined* in such a way that this is what they will do if they are rational, but in the latter their *payoffs are defined* to be higher for *Hawk because* they *choose Hawk* in the game.

Binmore (2007b, p. 14, 2009, p. 29) also argues that the players could not have been in a Prisoner's Dilemma if they chose *Dove because* in that case they would have chosen *Hawk*. There are two ways of formulating this argument, which are closely related to 1) and 2) above.

3) Players choose *Hawk* because their payoffs are defined in such a way that they choose to play *Hawk* if they are in a PD. Hence, if they choose *Dove*, they could not have been in a PD.

4) Their payoffs in a PD are defined to be higher for *Hawk because* they *choose Hawk* if they are in a PD. Hence, their choice of *Dove* means that they are not in a PD and thus defines the game as something different.

Again, if the players' payoffs for playing *Hawk* are defined to be higher than for playing *Dove because* they *choose Hawk*, as in 4), a revealed-preference argument is being used.

Binmore insists that only versions 2) and 4) of these arguments are tenable, whereas 1) and 3) are 'nonsense':

> In game theory, we are usually interested in deducing how rational people will play games by observing their behavior when making decisions in one-person decision problems. In the Prisoners' Dilemma, we therefore begin by asking what decision Adam would make if he knew in advance that Eve had chosen *Dove*. If Adam would choose *Hawk*, we would write a larger payoff in the bottom-left cell of his payoff matrix than in the top-left cell. These payoffs may be identified with Adam's utilities for the outcomes (*Dove*, *Hawk*) and (*Dove*, *Dove*),[5] but notice that our story makes it nonsense to say that Adam chooses the former because its utility is greater. The reverse is true. We made the utility of (*Dove*, *Hawk*) greater than the utility of (*Dove*, *Dove*) because we were told that Adam would choose the former. In opting for (*Dove*, *Hawk*) when (*Dove*, *Dove*) is available, we say that Adam reveals a preference for (*Dove*, *Hawk*), which we indicate by assigning it a larger utility than (*Dove*, *Dove*). We next ask what decision Adam would make if he knew in advance that Eve had chosen *Hawk*. If Adam again chooses *Hawk*, we write a larger payoff in the bottomright cell of his payoff matrix than in the top-right cell...
>
> Our data says that Adam will choose *Hawk* if he learns that Eve is to play *Dove* and that he will also choose *Hawk* if he learns that she is to play *Hawk*. He thereby reveals that his choice doesn't depend on what he knows about Eve's choice. If he is consistent, he will therefore play *Hawk* whatever he guesses Eve's choice will be. In other words, a consistent player must choose a strongly dominant strategy. (Binmore, 2007b, p. 13–14)

What Binmore means in stating that 'our story makes it *nonsense* to say that Adam chooses the former *because* its utility is greater', is that saying so would be committing the 'causal utility fallacy' (Binmore, 2009, pp. 19–21). In other words, utility does not provide any reasons for choosing one way or the other but rather merely represents an individual's preferences that are supposedly based on choices. I fully agree that utility does not

[5] Binmore apparently turned Adam into the column player and Eve into the row player between 1998 and 2007. What he states in the text is applicable to the game shown in this paper if the outcomes are (*Hawk*, *Dove*) and (*Dove*, *Dove*) when Eve chooses *Dove*.

provide reasons, and that it does not commit a theorist to any particular psychological assumptions, but I do not believe that the causal utility fallacy must be formulated in terms of choices, because payoffs in games cannot be based on choices. If Adam were to choose between the consequences associated with (*Dove*, *Hawk*) and (*Dove*, *Dove*) outside the game, he would have to choose the former because he prefers it to the latter. It is indeed nonsense to say that he makes this choice because his utility for the former is higher if the 'because' in the sentence is understood in a reason-giving way. However, if we interpret the sentence 'Adam chooses the former *because* its utility is greater' as a description of his preferences, it is not nonsensical. The intended meaning would then be that Adam chooses the former over the latter because he prefers it, and because he is already assumed to choose the most preferred alternatives.

I will now argue that only 1) and 3) are acceptable. Contrary to what Binmore seems to be suggesting here, payoffs cannot be constructed from players' choices in the game. I should first emphasise that defection is what we should expect from the players if they really are in a one-shot Prisoner's Dilemma game. I also agree that they would choose *Hawk* because their payoffs are defined to be higher, and they are already assumed to maximise their expected payoff. Finally, I also agree that if someone chose the cooperative strategy in a game that was supposed to be a PD, then a PD could not have been a complete description of the players' motivations and constraints (see also Blackburn, 1995). I believe that the methodological argument against cooperating in a PD is correct – it is just that *revealed preferences* cannot be used to sustain the argument for playing *Hawk*. The *players' choices in a game cannot* be used for defining the payoffs *in that same game*.[6] There are two reasons why this is conceptually impossible. First, we can only observe equilibrium play. Secondly, players' choices in the game could not define their payoffs because they are not given the opportunity to choose from among all outcomes.

Let us start with the first argument. Game-theoretical models predict and explain by specifying an outcome as an equilibrium. By assumption, it is not possible to observe disequilibrium outcomes because the equilibrium specifies what the theory predicts or explains. Assuming that a game describes some real-world phenomenon correctly, we can observe the players' choices in its equilibria, but we cannot observe all the choices that would derive from non-equilibrium strategies. Hence, we are not able to observe preferences for all other possible outcomes. It is thus not possible, even in

[6] This is how Binmore's point is commonly perceived. Ross (2005, p. 357), for example, argues that 'intentional-stance ascription infers beliefs and desires from strategic play at equilibrium'.

principle, to find out or define the players' preferences by observing their choices in the particular situation(s) the game is supposed to model. Payoffs from using various different strategies cannot be defined by observing what strategies the players use.

The confusion arises from Binmore's assertion that, 'We are given that she would choose *Hawk* if she knew that Adam were sure to choose *Dove*', and that it was a tautology that 'action B [i.e. *Hawk*] will be chosen rather than A [i.e. *Dove*] when the former yields a higher payoff by defining the payoff of B to be larger than the payoff of A if B is chosen when A is available' (see also Guala, 2006). These claims would seem to lead to the argument that the players must choose *Hawk* because their utility for *Hawk* is higher than that for *Dove*, whatever the other player chooses, and because it is tautological that they choose the outcome with the higher utility. It does not follow, however, that they choose the *Hawk* strategy as a matter of tautology, or as a consequence of the revealed-preference argument that choices define preferences.

To see why this is so, let us consider Binmore's phrase 'if B is chosen when A is available' in his argument that the payoff of B is defined to be higher than that of A. It may mean that B is chosen when A is available under circumstances of *certainty*. However, the choice the players are presented with, *when they are actually in the game*, is not between two sure outcomes, but rather between the *strategies Hawk* and *Dove* when the other player is playing *Hawk* or *Dove*, and this choice is a matter of derivation from a solution concept.

It has been argued that the players may choose from among all possible consequences only if, in the case of two players, one of them is removed and the game is effectively turned into a decision-theoretic situation.[7] Note that Binmore is discussing such choices in the long quotation above. Is he not, then, guilty of remodelling the game while analysing it? If 'asking what decision Adam would make if he knew in advance that Eve had chosen *Hawk*' (or *Dove*) is used for making a claim about how the game will be played, this is precisely what Binmore is doing. However, the players in the PD are not facing a situation in which they already know what the other player has chosen. Considering a situation in which the other player has already made a choice thus amounts to changing the game.

Furthermore, if in analysing the game we are allowed to consider choice situations the players do not face while playing it, we have to admit that if the players were given the opportunity to choose between (*Hawk*, *Hawk*) and (*Dove*, *Dove*), they would choose the latter. It will not do to retort that

[7] Rubinstein and Salant (2008) make a similar point about a coordination game.

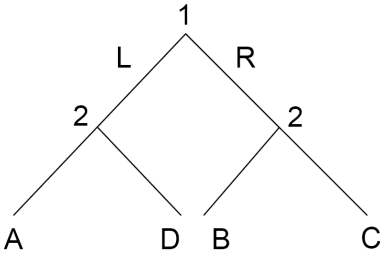
this counterfactual choice does not matter in analysing the game because Adam cannot choose between (*Hawk*, *Hawk*) and (*Dove*, *Dove*) in it: neither could he choose between (*Hawk*, *Dove*) and (*Dove*, *Dove*), or between (*Hawk*, *Hawk*) and (*Dove*, *Hawk*). It is rather that he chooses between two strategies that have lotteries as outcomes. Let A be the consequence associated with (*Hawk*, *Dove*), B with (*Dove*, *Dove*), C with (*Hawk*, *Hawk*) and D with (*Dove*, *Hawk*). Adam chooses between *Hawk*: (p, 1-p; A, C) and *Dove*: (p, 1-p; B, D) when p describes the probability that Eve will choose *Dove*. My point, of course, is not that we should consider the revealed preference between (*Hawk*, *Hawk*) and (*Dove*, *Dove*) when we analyse the Prisoner's Dilemma. The point of this reduction ad absurdum is rather to show that we should abstain altogether from considering what the players would choose in such counterfactual situations when we analyse the PD game because doing so changes it into something else. The PD is not correctly described by two simultaneous decision-theoretic choices. If it were, there would be no need for game theory. Although Binmore does not tamper with the payoffs, he does tamper with the structure of the game because he changes the description of the choices available to the players.

Let us now consider another example. If the payoff for consequence B is higher than that for A, this does not yet imply that the player will always choose B over A in a game because he or she may not get to choose B over A under certainty. If a player gets to choose between two lotteries R = (p_1,1-p_1;B,C) and L = (p_2,1-p_2;A,D) that also involve outcomes C and D, he or she may choose L over R if p_2 is sufficiently higher than p_1 or if the utility for C is low enough as compared to D. Suppose, for example, that the payoffs are given by A = (2,4), B = (3,2), and C = (0,3) = D,[8] and that these outcomes may be obtained in the complete knowledge game between players 1 and 2 as defined in Figure 1.

Applying subgame perfection, we can derive the prediction that player 1 will choose L. Player 2 then chooses A over D, and A emerges as the outcome. However, player 1 strictly prefers B over A. He or she had a good reason not to choose the option that contained his or her best outcome B (as one possible prize in the lottery between B and C) because player 2 is sure to choose C over B if he or she gets to make the choice. Note that if player 1's choice of L were to define his or her utility for A as greater than that for B, it would provide false information about his or her preferences. The theory of revealed preferences was originally developed for choice under

[8] The latter equality sign (=) means that the payoffs are the same under C and D while the description of the outcomes may be different.

Fig. 1 — a simple game



certainty, and this example illustrates why applying it under uncertainty is indeed questionable. (See also Mas-Colell, 1982.)

This simple game also illustrates another reason why players' choices in games cannot define their preferences (cf. Grüne-Yanoff and Lehtinen forthcoming). Player 1 only makes a choice between L and R, and his or her choices thus cannot be used for defining any of the payoffs for the final outcomes. Furthermore, player 2 never faces a choice between A and B, and his or her choices could in principle only define his or her preferences between A and D or between B and C, but not both. This simple game shows that the plausibility of Binmore's strategy of defining payoffs with players' strategy choices vanishes as soon as games that are even a little more complicated than the PD are analysed. If the revealed-preference interpretation really means that *all* choices define preferences, then it is inconsistent in game theory because it is conceptually impossible for strategy choices to define preferences.

Hausman (2005a, 2011) provides an even more extreme example in which the description of the consequences is such as to preclude the possibility that the players could make choices from among all of them *even outside the game*. Suppose that Darcy first chooses whether or not to propose marriage to Elizabeth. If he does not propose, the game ends and the payoffs are A = (2,2). If he proposes, Elizabeth then accepts B = (3,1) or turns down C = (1,3) the proposal. Darcy's preferences between B and C cannot be defined by choices because he cannot control Elizabeth's choices. It is not a choice that he could make, even in principle. This example reveals the fundamental reason why payoffs cannot be revealed by choices in a game. A game is a situation in which, by very definition, players cannot make choices on behalf of other players.

Let us now return to the Prisoner's Dilemma. If we have a game before us that we claim to be a PD, and we observe that a player chooses *Dove*, it is not

a tautology that the players are not in a PD because the player might choose *Dove* if he or she were using a solution concept that conflicted with using dominant strategies. However, not choosing *Dove* would be a tautology if using a dominant strategy were a tautology, but it is not, as I will now show.

One of the reasons for assuming that payoffs are vNM utilities is that it guarantees that the utilities are linear with respect to the probabilities, and that the players are assumed to be individually rational in a decision-theoretical sense. I also subscribe to this interpretation if adopting it is not taken to be a tautology, and if the reason for adopting it is not the epistemic security that it allegedly provides. This assumption has the pragmatic benefit of guaranteeing that we can compute expected utilities in games by combining beliefs and payoffs. However, whereas Binmore interprets the claim that payoffs are vNM utilities to mean that they are revealed preferences, I interpret it to mean merely that they are linear with respect to the probabilities and that the players are rational. It follows from the latter that players must choose the strategy that maximises their expected payoffs. Adam will thus choose *Hawk* if the expected utility for *Hawk* is higher than that for *Dove*. Since E(*Hawk*) = p4+(1-p)2 = 2p+2 and E(*Dove*) = p3+(1-p)1 = 2p+1, E(*Hawk*) > E(*Dove*) irrespective of the value of p. Given that the expected payoffs are always higher for *Hawk* in a PD, Adam will choose *Hawk*.

Bernheim (1984) and Pearce (1984) showed that using non-dominated (rationalisable, to be exact) strategies is a consequence of individual maximisation. Whatever the strategic considerations in a game, the players will never choose strictly dominated strategies if they are individually rational in the sense of single-person decision theory. The Prisoner's Dilemma is thus really a special case in that it is possible to use a solution concept specifying behaviour that is already implied by individual maximisation in a decision-theoretic sense. This explains why Binmore was able to say that 'a consistent player must choose a strongly dominant strategy' without contradicting himself. He also stated that it was a tautology that players acted as though they were maximising expected payoffs. They choose expected utility-maximising strategies because the payoffs are already assumed to be defined on the basis of expected utility-maximising choices, in other words in such a way that the players are assumed to be rational in the sense of satisfying vNM postulates. If the payoffs are indeed assumed to be vNM utilities, the players must use strongly dominant strategies because it would be inconsistent to assume that they are rational and irrational at the same time. The use of strongly dominant strategies is a matter of tautology if assuming that the payoffs are vNM utilities is a tautology. It is not, however, because *assuming* that the payoffs are vNM is a *methodological choice*.

Note that even if Binmore's argument that playing strongly dominant strategies is a tautology were valid, it would not follow that using any other solution concept was a tautology. If the tautology argument were to be applied consistently to all solution concepts in game theory, it would follow that using a solution concept would always be a tautology. However, there are games for which several different solution concepts can be defined, and which provide conflicting behavioural advice. For example, it cannot be tautological for players to use forward induction in some game if backward induction also applies (see Grüne-Yanoff and Lehtinen forthcoming for an example). However, Binmore apparently does not think that revealed preference arguments need or can be used for all solution concepts:

> But the literature on backward induction does not adopt a strict revealed preference approach because it then becomes a tautology that backward induction holds. Whatever behavior we might observe in a finite game of perfect information is compatible with backward induction if we are allowed to fill in the payoffs after the event. (Binmore, 1996)

Why is it acceptable to use a revealed preference argument for defending dominant strategies in PD games but not backward induction? Given that Binmore (1996, 1997, 2007a, p. 109) argues against the rationality of backward induction, it seems that he restricts using revealed preference arguments to those solution concepts that he thinks rational agents should employ.

The choice of a strictly dominated strategy is not ruled out because of revealed preferences, or because it would be a tautology that they are ruled out, but rather because the solution concept is so compelling or because the payoffs are already assumed to be vNM utilities. These two arguments for choosing the *Hawk* strategy are independent of each other. If one is convinced of the plausibility of choosing non-dominated strategies one does not need to assume that payoffs are vNM, and conversely, if one already assumes that the payoffs are vNM, it is not necessary to appeal to dominant strategies in order to justify the choice of the *Hawk* strategy.

Despite the existence of two independent arguments for the *Hawk* strategy, it is not *conceptually* impossible for a player to choose *Dove* even in a real PD game: such a choice would require the player to violate the dominant-strategy solution concept. Furthermore, if we wish to interpret the choices of such players as violations of the dominant strategy solution concept rather than as deriving from a different game, we must admit that the players are not individually rational either. However compelling this solution concept is – and it is very compelling – the players cannot be assumed

to use it as a matter of tautology. It would be a tautology only if inconsistent players could not end up playing Prisoner's Dilemmas as a matter of tautology.

This, in a sense, reflects Binmore' argument: if a person has inconsistent preferences and thus violates vNM postulates, we could not elicit them in a reliable way because consistency is a precondition for successful elicitation, and representing preferences with utilities presupposes that the preferences have been successfully elicited. Binmore takes this to mean that once a game theorist writes down a payoff matrix, he or she is thereby committed to the consistency of the players as a matter of tautology. In other words although inconsistent people could play Prisoner's Dilemmas, game theory would refuse to analyse such PDs on the grounds that the payoffs could not be interpreted as vNM utilities and the game could thus not have been written down correctly. However, if my argument against interpreting payoffs as revealed preferences holds, we are never in a position to guarantee that the payoffs have been written down correctly. Interpreting payoffs as vNM utilities thus cannot be justified on epistemic grounds because we could not define them in terms of actual choices even if the players were rational. The justification rather has to be based on pragmatic considerations. Whether or not we would care to adopt this interpretation depends on what we wanted to do with game theory.

I will now turn to the second, and in my view the more plausible, interpretation of what Binmore meant when he stated that the payoff of B would be defined as larger than that of A if B was chosen when A was available: B would be chosen if A and B were available in a *counterfactual* choice situation.[9] In interpreting payoffs as revealed preferences, Binmore probably meant that they described behaviour that would ensue if the outcomes were presented to the players in a counterfactual choice situation because we are *given* what the players *would* choose.[10] The payoff associated with a pair of strategies (*Dove*, *Hawk*), for example, could refer to some consequence A, which may be a physical object, a state of affairs or an event, but it could also be a combination of mental and physical objects, events, and so on. The payoff for (*Hawk*, *Hawk*) refers to another consequence, B. Now, if the game specifies that the payoff associated with strategies (*Dove*, *Hawk*) is smaller than that associated with strategies (*Hawk*, *Hawk*), it means that the player prefers B to A, and that he or she would choose B over A in a counterfactual choice situation between them.

[9] See Skyrms (1998) for an account of counterfactuals in games. Skyrms interprets revealed preferences in games in terms of 'dispositions to choose'.

[10] Note that his interpretation is taken to refer to *dispositions* (e.g., Ross 2006).

It is perfectly sensible to say that the payoffs are defined in such a way that playing *Hawk* yields a higher payoff for the row player than playing *Dove* if the column player were to play *Hawk* (or *Dove*). The counterfactual interpretation of payoffs is thus acceptable. However, even this does not imply that players choose *Hawk* as a matter of tautology because the counterfactual choices are really different from the choices in the game. The counterfactual choices are assumed to be independent of the strategic considerations that are present in the game. One still needs the solution concept for deriving the (*Hawk*, *Hawk*) outcome because the counterfactual choices do not define the choices in the game.

Furthermore, since the players cannot make their payoff-defining choices while playing the game, in other words while they are choosing their strategies, the counterfactual revealed-preference interpretation of payoffs must be *intrinsically counterfactual*. What I mean by this is that if we already have a payoff structure for the whole game before us, the only way in which we can appeal to revealed preferences is counterfactual. The problem with such an argument is that it does not provide any actual epistemic surplus; it does not deliver a way of finding out the actual preferences in any situation. Daniel Hausman (2008, p. 137) expresses this as follows: 'In switching from actual to hypothetical choice, one has abandoned the empiricist ideal of avoiding references to and reliance on anything that is not observable.'

4. Testing Game Theory and the Revealed-Preference Interpretation

If the revealed-preference interpretation of payoffs is to confer scientific respectability on game theory, it might be taken to imply that modelling games should be done by actually eliciting the players' preferences. However, Binmore apparently does not consider it worth arguing that payoffs for games are or should be modelled by means of revealing preferences through some elicitation procedure. He assumes that the elicitation has already taken place and then ignores the issue:

> Although the revealed-preference interpretation of utility is maintained throughout this book, half of the labour of constructing a utility function from an agent's choice behaviour is skipped. It is assumed that a preference relation has already been constructed and that it remains only to show that it can be represented using an appropriate utility function. (Binmore, 1994, p. 268)

There have been some recent efforts to provide revealed-preference conditions under which the players' choices rationalise various solution con-

cepts.[11] These accounts are completely different from Binmore's. They 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*. They vividly demonstrate that what is needed for their verification or falsification is the possibility of observing the players' choices in each case in which at least one strategy is deleted from one player.

Binmore's comment above gives the impression that we can safely assume that we have all this information. He also states that modern utility theory 'assumes that we already know what people choose in some situations, and uses this data to deduce what they will choose in other situations' (Binmore, 2009, pp. 8–9). Note that he does not claim that we actually know what people chose in some situation and then use these data to deduce what they will choose in other situations. He makes the much weaker claim that we *assume* that we know what they will choose.

Binmore also argues the following:

> Being able to fit a utility function only tells us that the behavior is consistent – it doesn't tell us why the behavior is consistent. For example, one way of explaining the behavior of that half of the population of inexperienced subjects who cooperate in the one-shot Prisoners' Dilemma is to say that they are optimizing a social utility function whose arguments include the welfare of others. Another is to attribute any consistency in their behavior to the fact that they are unconsciously operating a social norm better adapted to repeated situations. Both explanations fit the data equally well, but the former explanation is easier to criticize. What is the point of insisting that players have other-regarding utility functions built into their brains if doing so doesn't allow predictions to be made about how they will play in future, or in other games? But we know that the behavior of subjects in the one-shot Prisoners' Dilemma changes markedly over time as they pick up experience. A social utility function fitted to the behavior of an inexperienced subject will therefore fail to predict how he or she will behave when experienced-let alone when they play other games in other contexts. (Binmore, 2007a, p. 18)

He is here alluding to the dispute over whether cooperation in a PD is to be interpreted in terms of other-regarding preferences due to inequality aversion (Fehr and Schmidt, 1999) or an unconsciously adopted social norm. He also suggests that attributing other-regarding preferences to the subjects is illegitimate because it is not consistent with the revealed preference perspective.

---

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

> The theory of revealed preference tells us that we can describe the behavior of agents who choose consistently as optimization relative to some utility function. However, economists who take the orthodox neoclassical position seriously are very careful not to deduce that the observed behavior was generated by the agent actually maximizing whatever utility function best fits the data. This would be to attribute the kind of psychological foundations to neoclassical theory that its founders invented the theory to escape. (ibid.)

I confess that I have a hard time interpreting what Binmore tries to argue here. On the one hand, being able to fit a utility function does not discriminate between the inequality aversion hypothesis and the hypothesis of an unconsciously operating social norm. On the other hand, the hypothesis of an unconsciously operating social norm is just as psychological as the inequality aversion hypothesis. Even if it is not, it is clear that accounting for cooperation in terms of an unconscious norm is epistemically just as problematic as using inequality aversion. Thus, if Binmore intended to argue for the unconscious norm hypothesis on the grounds of revealed preferences, it would be a fallacious argument. This leaves us with the claim that the inequality aversion hypothesis does not predict well when people learn how to play the PD or when they play other games. Binmore has repeatedly argued for such 'portability' of experimental results (beginning in Binmore et al., 2002).[12]

In his comment on the experimental results obtained by Henrich et al. (2005), Binmore argues that they do not show that the tools of game theory are inadequate because game theory does not need to assume self-interest on the part of players. This is correct in the sense that game theory does not need to assume self-interest. However, as many authors note, it is necessary to assume something about the motivations of people in order to make the theory testable.[13] Binmore is not suggesting that game theory is analytic and thus beyond empirical evaluation because he is willing to discuss the interpretation of the experimental results in the first place (see, e.g., Binmore, 1999), and he has conducted various experiments himself (Binmore, 2007a). Since the players in the Henrich et al. experiments failed to act in the way the theory predicts, something must yield: either the payoffs in the model were incorrect, or the solution concept was faulty, or the players were not rational.

---

[12] Binmore and Shaked (2010) provide an extensive methodological critique of Fehr and Schmidt's inequality aversion hypothesis in which non-portability features prominently.

[13] See Guala (2006) for a recent discussion.

There is an alternative interpretation of the results of Henrich et al.[14] The point of the experiment would be to measure people's preferences by way of assuming that the solution concept is correct. The deviation from equilibrium play would then provide information about how exactly the payoffs differ from what they would have been if the ultimatum game had described the preferences. The point of the experiment would thus be to reveal preferences. Although this is a possible interpretation, it is clear that one cannot use such an experimental procedure to simultaneously test game theory (i.e. the solution concept) because the possibility of revealing preferences correctly presupposes the validity of the solution concept.[15] Note also that it does not provide a case of eliciting preferences *for* games because the agents are not asked to play another game with the elicited preferences.

It may be worthwhile to point out a possible misunderstanding. The aforementioned interpretation of Henrich et al. is not really one based on revealed preferences because presupposing the validity of the solution concept implies assuming something about the psychology of the subjects.[16] A proper revealed-preference interpretation supposedly works without any such assumptions. This is also why denying the revealed-preference interpretation of payoffs in games does not imply the claim that one cannot obtain reasonably reliable information on individual preferences in game-theoretical experiments. Of course, it is possible, but the question about whether this can be understood from the point of view of revealed preferences concerns not whether such information can be attained, but rather whether it can be done without any psychological assumptions. Since solution concepts do incorporate such assumptions, *revealed preference theory* means extrapolating behaviour in one set of circumstances on the basis of choice data in another set of circumstances only if the data come from the reference lottery but not if they come from a game-theoretical experiment.

Being able to use the results from experiments in other circumstances (i.e. replicability) is a criterion any sensible scientist should endorse. I do not need to take a position on how successful Binmore's use of this criterion

---

[14] Till Grüne-Yanoff suggested this interpretation to me.

[15] It seems that this is not Binmore's interpretation. The reason is that since Binmore does not believe that backward induction is rational to begin with, it would be odd if he were to presuppose its validity in the experiments. Furthermore, given that he has defended the evolutionary justification of solution concepts (Binmore, 1987, and also 2007a, pp. 4, 28), he seems more willing to argue that the test subjects had not yet learned to be rational.

[16] Binmore (2007a, p. 312) seems to agree on this: 'There is no need for game theorists to seek to insulate themselves from the criticism of experimentalists by claiming that their theorems have no relevance to how real people behave.'

is in his dispute with Fehr and Schmidt.[17] The question, for my purposes, is whether the criterion is inherently based on revealed preferences. I do not think it is, but Binmore gives the impression that he thinks otherwise. The criterion implies that there must be some fixed preferences which can be transported to another context. He says that stability of preferences is an implicitly understood assumption in revealed preference theory (2009, p. 9). Indeed, as Wong (1978) noted long ago, if preferences are not stable, there is reason to think that choices will not satisfy any consistency requirements. However, there are arguments for the fixity of preferences that have nothing to do with revealed preferences. For example, the idea that explaining something with a change of preferences is *ad hoc* (see, e.g., Stigler and Becker, 1977), such arguments are not intrinsically related to them.

Inequality aversion is a particular psychological hypothesis. According to Binmore, this hypothesis is not valid in the relevant cases because something else is:

> Only after the learning phase is over can we expect to find subjects at a Nash equilibrium, each behaving as though trying to maximize his or her own utility function given the behavior of the other subjects. But do we then not find them simply maximizing money? (Binmore, 2007a, p. 4)

Although maximizing money is not as straightforwardly psychological as inequality aversion, it is surely based on some psychological facts or theories. If it isn't, why would people need to learn to act in this way? But then, if it is psychological, Binmore is here proposing one psychological hypothesis rather than another.

5. Descriptions of Consequences

The idea behind elicitation is that if players' preferences satisfy a set of axioms, it is possible to construct a choice experiment in such a way that the utility function can be defined (see e.g., Hirshleifer and Riley, 1992). As mentioned above, game theorists typically do not model games by conducting such elicitation exercises. The point of the revealed-preference argument is merely to show that it *would be possible in principle* to elicit the preferences. Thus, in claiming that we are given what the players would choose in a counterfactual choice situation, Binmore could be referring to the choices they would make in such elicitation procedures. I will now put forward an

[17] More generally, I am not qualified to evaluate the debate. Although I am criticising Binmore's use of the revealed-preference argument in it, what I write in this section should not be taken as an argument for (or against) the inequality-aversion hypothesis.

argument suggesting that although eliciting preferences for games is indeed possible in principle, doing so successfully presupposes that the game theorist knows a wide variety of psychological and contextual facts about the players. Given that the point of the revealed-preference approach is precisely to do without 'peeping into the subjects' heads', the mere possibility that the theorist does not know about these mental matters undermines the rationale of this interpretation.

Modelling payoffs in games by means of elicitation methods is particularly prone to the problem of state-dependence because the context of the elicitation of utilities *cannot be the game itself* (Guala, 2006). Let us take the example from Sen (1993) to which Binmore (1998) refers.

> Modern theory of revealed preference ... recognizes that Eve's actions may depend on states of the world which she has not observed or which have yet to occur...
>
> Sen (1993) tells us that people never take the last apple from a bowl, and hence are inconsistent when they reveal a preference for no apples over one apple when offered a bowl containing only one apple but reverse this preference when offered a bowl containing two apples ... Once we have set her problem in an appropriately wide context [one in which the consequence space includes the effects of choosing the last apple in a society where this kind of an act provokes moral disapprobation], Eve's apparent violation of the consistency postulates of revealed preference theory disappears. She likes apples enough to take one when no breach of etiquette is involved, but not otherwise.
>
> The lesson to be learned from Sen's example is not that the consistency requirements of revealed preference theory are too strong to be useful, but that the first thing to do when they seem to fail is to ask whether the choice problem has been adequately modeled. (Binmore, 1998, pp. 360-2)

Standard utility theory assumes context-free preferences, but the problem of picking an apple provides an example of the general problem of state-dependence. (See Drèze and Rustichini, 2004, Karni, 2009, for overviews.) Preferences are said to be state-dependent when the prevailing state of nature is of direct concern to the decision maker. Binmore criticises Sen for not describing the consequences correctly. (See also Binmore, 2007b, pp. 392-3, 2009, p. 9.)[18] They should be defined in such a way that the state-dependence is already taken into account. It is clear that the Weak Axiom of Revealed Preference (WARP) is not violated if the alternatives are thus re-described, but as many authors have remarked, if we are always allowed to make such re-descriptions of the choice alternatives, the consistency axioms

[18] Dowding (2002) puts forward the same argument. This example is also discussed in Pettit (1991), Baron (2000, p. 235), Chapman (2003) and Ross (2005, pp. 133–140).

will be vacuously fulfilled. The main concern, however, is not vacuity in that the example reveals an epistemological issue that the revealed-preference interpretation was supposed to resolve. The problem is that mere choices are not sufficient for finding out or defining a person's preferences.

A game theorist can take state-dependent preferences into account and re-describe the choice options in an appropriate manner only if he or she *knows* the exact manner in which the players' preferences are state-dependent. Assume now, for the sake of argument, that the game theorist is able to take state-dependence into account. We could then ask how he or she has arrived at the right characterisation of the consequences. In particular, is it possible that this could have been done by merely observing the players' choices as the revealed-preference perspective requires? If the game theorist can take state-dependence into account in an appropriate way, he or she can do so only by making some assumptions about the mental factors underlying the choices. If all we are given about the situation is that the person first does not choose the apple, and then does choose it under another set of circumstances, we have to admit that there is inconsistency *if the person's preferences were as we first supposed them to be*. Of course, if we observed such inconsistency in terms of the preferences we first ascribed to our subject, we would begin to doubt whether this first ascription was the correct one. The only way in which we can align our conception of the choice alternatives such that it matches the conception of the choosing subject is by making psychological assumptions.[19] It would not be possible to re-describe the choice options in this example unless we already had some background knowledge about etiquette. This is a very basic and well-known criticism of behaviourist psychology upon which revealed preference theory is based: mere choices do not provide us with sufficient information about the players' preferences. Sen (particularly in 1993, 1995, 1997) expressed this idea in stating that internal consistency of choice does not make any sense (see also Gaertner and Xu, 1999). It is necessary to refer to something external to choice in order to make sense of consistency axioms. The theory does not allow us to do without psychological assumptions when we attribute preferences to individuals. It is then an empirical question how often and to what degree game theorists and other social scientists successfully describe what agents take to be the real choice alternatives.

It has been established by various contributors that if preferences are state-dependent, they cannot be reliably elicited (e.g., Karni, 1999). The reason for this is that when they are state-dependent the description of the consequences depends on the circumstances in which some object of desire

[19] See Wong (1978), Sen (1973, 1977, 1993, 1995) and Sugden (1985).

is enjoyed. Successful elicitation presupposes that the subjects describe the consequences in the same way as the modeller (cf. Rubinstein and Salant, 2008). Under state-dependence this means that the modeller has to know exactly the manner in which preferences are state-dependent. Binmore's (2007b, p. 394) response is to argue that one can identify the set of consequences with a subject's states of mind rather than with physical objects. For example, the relevant consequences are the *states of mind* that accompany having an umbrella-on-a-sunny-day or having an umbrella-on-a-wet-day rather than just having an umbrella (Binmore, 2009, p. 7). This very example of having an umbrella in different weather conditions is used in many illustrations of the problem of state-dependence. It is a good example only in that it is easily understood. The theorist is probably able to differentiate the value of having an umbrella in different weather conditions, but if the player cares about whether or not a given object was attained through a non-cooperative choice by another player, it is not so evident that the theorist's psychological acuity suffices.

Binmore's response amounts to arguing that such epistemic problems have *always* already been solved. Given that he also welcomes the idea that this move makes the theory tautological, it might be useful to reformulate the criticism. As Karni (2009, p. 227) argues, 'the state-independent utility function ... is observationally equivalent to the state-dependent utility function ... Hence the validity of the state-independent utility convention is not subject to refutation within the framework of the methodology of revealed preference.' Thus, in general, mere choices will never tell us whether or not preferences are state-dependent. The main problem is not vacuity or tautology, however, but is rather that one cannot just assume away epistemic problems. Furthermore, given that Binmore does so by assuming that the game theorist *always has the relevant psychological information*, his argument is a masterpiece of question-begging. Camerer (2008, p. 58) expresses this pithily: 'The revealed preference approach solves the problem of figuring out when choices betray true preferences by assuming it never happens. This is like an ostrich solving the problem of escaping danger by sticking its head in the sand.'

The danger is that if it were to become generally believed that revealed preference theory frees game theory, or economics, from psychological assumptions, theorists would be interested in redefining the choice options only when the subject's behaviour seemed to violate the WARP or some other consistency condition. As Sen (1973) noted long ago, the same choice may derive from various kinds of considerations. If the consideration of contextual matters and psychological assumptions starts only after WARP has been violated, one could mis-describe the underlying preferences even when

WARP was not violated. Let us modify Sen's example a little. Suppose that a person takes two glasses of milk when offered them in one situation, and then takes one glass but refuses another in another situation. He thus first chooses two glasses from {2 glasses, 1 glass, no milk}, and then one glass from the same set of alternatives. If we describe the options more coarsely by saying that the choice set is rather {milk, no milk} the choices are consistent with WARP. Suppose, however, that the person loathes milk, but he has been told that when invited to dinner with a particular group of Arab Bedouins it is extremely impolite not to drink at least two glasses, and he is not too disgusted to drink any. After the first occasion, however, he is no longer able to finish two glasses and can only manage one. These coarsely described choices do not violate WARP but they do give a misleading view of the person's preferences: if we bring him back to his home in Germany he will refuse milk on all occasions even though the implicit (but unobservable) assumption in revealed preference theory is that preferences do not change.

Although Sen gives examples in which consistency axioms may reasonably be violated, his main point is not so much that they are or should be violated, but rather that the theory does not do the job it is supposed to do: it does not provide a theory based merely on observable behaviour.[20] His point is that fulfilling consistency conditions does not provide us with sufficient information about a person's preferences, and that some psychological assumption is necessary even when revealed preference theory is applied.[21] The necessity of appealing to some psychological factors does not, of course, mean that one has to subscribe to some *particular* psychological theory such as the assumption of self-interest.

6. The 'Standard Position' Versus Revealed Preferences

What difference would it make to accept or deny that payoffs are revealed preferences in games? Mongin and d'Aspremont (1998, p. 386) discuss what they call a 'standard position,' which they contrast to revealed preferences by saying that payoffs describe preferences that *underlie* the player's choices in games. According to this interpretation, the players' choices re-

[20] What I suggest here is fully consistent with considering choice-consistency axioms a starting point for finding out what the preferences are (cf. Dowding, 2002). Indeed, assuming rationality on the part of the subjects in question is the starting point for interpreting any behaviour, as Donald Davidson emphasises.

[21] Even Samuelson did not seem to think that the theory of revealed preference made any sense unless it was assumed that the choices derived from a conscious mind that had understandable goals: 'While I can see why a man with a mind should exercise it consistently, I fail to see why a beast with no mind should satisfy the Weak Axiom or even consistency of demand choices.' (Samuelson, 1963, p. 235)

flect their payoffs in the sense that they are assumed to choose a strategy that yields them the highest expected payoff. Utility (or payoff) itself, however, is viewed as underlying their choices rather than being identical to them.

Binmore (2006) approvingly cites Don Ross thus: 'The theory of revealed preference tells us that any consistent behaviour can be described by saying that the decision maker is behaving as though maximizing a utility function' (Ross, 2006). He thus equates (expected) utility theory with revealed preference theory. If the theory of revealed preference meant nothing more than the idea of describing consistent behaviour such that the decision maker acts as if he or she were maximising a utility function, then I cannot see why anybody would wish to disagree. After all, this has been demonstrated in representation theorems (e.g., von Neumann and Morgenstern, 1947, Savage, 1954), the mathematical validity of which is beyond doubt. If this is what a revealed-preference interpretation of payoffs means in game theory as well, it is correct, but utterly trivial in that the 'standard position' also accepts the claim that consistent behaviour (and preferences) can be represented. The idea that consistent preferences can also be represented as payoffs in game-theoretic analyses is not objectionable. I am labouring this point in order to make it clear that one cannot subscribe to the revealed-preference interpretation without making any philosophical commitment. As Hausman (2000) notes, the theory of revealed preference does not help in solving theoretical or empirical puzzles (see also Wong, 1978, p. 51). It does not offer anything substantially new over and above the standard utility-based account. Indeed, it was adopted only because it was thought to be methodologically superior to the preference-based account in providing a scientifically respectable theory based only on observable choices.[22] Another reason for its adoption was to provide an account that does not make any particular assumptions about the psychological causes of choice behaviour. Given Binmore's assertion that modern decision theory 'makes a virtue of assuming nothing whatever about the psychological causes of our choice behavior' (Binmore, 2009, pp. 8–9), the latter reason also is his motivation.

If Binmore and others who subscribe to the revealed-preference interpretation ultimately retreat to the position that revealed preferences just mean representing preferences with utilities, they denounce the philosophical commitment that the theory has had in the history of economics. This is one reason why philosophers cannot understand why one must use misleading revealed-preference rhetoric. Given that philosophers are not necessarily

[22] The development of current mainstream theories of utility and revealed preference has been driven by a concern to provide a theory that refers only to observable variables. See e.g., Mandler (1999) and Giocoli (2003) for historical reviews.

interested in changing the way in which game-theoretical analyses are conducted, but are rather just intent on changing this rhetoric such that the use of psychological assumptions is acknowledged, economists and game theorists do not have much to fear from admitting that payoffs are not reasonably interpreted as revealed preferences.

There are two major differences between the 'standard position' and the revealed-preference interpretation. The first is that the former refutes the idea that the latter provides a method for finding out individual preferences in a way that does not require any mental attributions or assumptions. The revealed-preference interpretation is supposed to provide a theory that does not require reference to mental states. Since the *only* reason for adopting such an interpretation is that it might provide an epistemologically respectable theory that avoids such reference, if it fails in this respect it cannot be anything but misleading.

Stanley Wong and Amartya Sen argue that the theory fails precisely in this respect because it does not allow dispensing with 'a peep into an agent's head'. The fact of this failure, however, does not imply that game theory should be criticised on the grounds that it makes unrealistic psychological assumptions. Indeed, just as Binmore has emphasised, (payoffs in) game-theoretical analyses need not be based on any *particular* psychological assumptions. Denying a revealed-preference interpretation thus merely means denying that there is an epistemologically foolproof way of defining players' payoffs for games without invoking psychological assumptions.

The second difference is that, according to the 'standard position', although it is possible to define utilities on the basis of choices in some circumstances, there are *some* choices that should not be taken as utility-defining. Given that players' choices *in* games cannot be used for defining their preferences, a conceptual distinction between choices and preferences must be made. To say that preferences 'underlie' choices is a way of expressing the idea that there is a major conceptual distinction between the two.

Adopting the 'standard position' thus does not amount to very much. In addition to going along with the idea that consistent preferences can be represented, it is fully consistent with the claim that choices can sometimes be used for defining preferences, and that revealed-preference axioms can be interpreted as conditions for finding out information about individual preferences.

From the point of view of practising economists, it may look as though abandoning the revealed-preference interpretation would imply admitting that the methods they use in their daily practice are faulty. Indeed, at times they say that criticising revealed preferences amounts to criticising standard

economics (e.g., Gul and Pesendorfer, 2008). Some recent contributions view the theory of revealed preference merely as a way of extrapolating the behaviour of individuals in some set of circumstances on the basis of choice data in another set of circumstances (e.g., Bernheim and Rangel, 2008, 2009). Those who subscribe to such an interpretation, and actually use it to calibrate parameters concerning individuals' preferences, may wonder what could be wrong with their methods. There is nothing wrong with these methods, in fact. Critics of revealed preferences have typically not argued against the attempts to reveal preferences in research practice. If individual choices are used to construct preferences by way of using the WARP, for example, I do not see why anyone would object to the research practice. It also involves an attempt to obtain information on preferences with very thin psychological assumptions. The moot point is whether or not such methods can be taken to *guarantee* correct information on individual preferences.

7. Conclusions

Given that players do not have the opportunity to choose from among all outcomes when they are actually making choices in a game, and that we can only observe the equilibrium choices, the first revealed-preference interpretation is conceptually impossible if it is taken to imply that the players' choices may be used for defining their preferences. Players' choices in a game do not and could not define their preferences in that game. Thus, the argument for choosing *Hawk* in a Prisoner's Dilemma is based on the plausibility of the dominant strategies, or the assumption that players are already assumed to satisfy vNM postulates rather than revealed preferences. If the revealed-preference interpretation is to be an illuminating account of game-theoretical practice, only the second interpretation could be invoked, and it would inevitably imply that games are modelled by actually eliciting payoffs. In that it is not the way in which games are currently modelled, this is a misleading interpretation.

Because the game and the elicitation procedures are two different contexts, there is always the possibility that the transfer of payoffs from the latter to the former fails due to the state-dependence of preferences. Simply assuming that the game theorist is always able to diagnose the reason why a player violates a choice-consistency condition in an elicitation procedure does not solve this epistemic problem. Even if the game theorist were successful in such a diagnosis, it would be only because he or she used psychological resources that went beyond the players' choices. In other words, because the players' choices are not sufficient in themselves to produce a correct characterisation of the choice options, theorists who really want to

know what is the right game, must invoke psychological assumptions about the players.

The extent to which it is possible for a game theorist to know the players' subjective conceptions of the alternatives is a matter that cannot be decided by means of philosophical argument. Nevertheless, the mere possibility that they cannot be taken adequately into account undermines the idea that the revealed-preference interpretation makes reference to psychological factors unnecessary. It is true that most game-theoretical work requires only elementary psychological knowledge, and that game theory is not dependent on any particular psychological doctrine, but nevertheless the revealed-preference interpretation gives a misleading impression that game theorists can conduct their modelling and analysis without the need to consider psychological issues at all.

What Binmore tries to argue through the revealed-preference interpretation is better expressed by saying that payoffs really are what they are once the payoff matrix has been written, and that they should not be modified while the game is being analysed.[23] Given that Binmore himself violates the methodological injunction to keep the game fixed in his analysis of the Prisoner's Dilemma in terms of revealed preferences, and given that he no longer emphasises the distinction in his later writings (2007b, 2009), he may now think that the same methodological point is better expressed in terms of revealed preferences. However, this argument can and should be made without invoking revealed preferences. Viewing payoffs as *underlying* choices is fully consistent with the requirement of separating modelling and analysis.

Perhaps Binmore's main point is that game theory should not be used in such a way that the theorist postulates a preference structure for a game and then uses this structure to tell a story about some real-world phenomenon. We might as well tell these plausible but un-testable stories without recourse to game theory because it is not doing any real job in such accounts. This is a sensible argument. If game theory is to be useful in some way, its usefulness must derive from surprising or interesting *analyses of strategies*. The revealed-preference interpretation of payoffs is not necessary in this case, however, because the same argument applies to viewing payoffs as underlying preferences.

---

[23] Some game theorists use the term 'revealed preference' in informal discussions to refer to the idea that we do not know what the preferences are, but that they are what they are. This usage borders on the paradoxical given that the only reason why the theory of revealed preferences was presented in the first place was to provide a way of knowing what the preferences were.

Acknowledgements

I am grateful to Till Grüne-Yanoff, Jack Vromen, various anonymous reviewers, and audiences at seminars in the Erasmus Institute for Philosophy and Economics and the University of Helsinki (philosophy of science group) for their useful comments. The usual disclaimer applies. Funding from the Academy of Finland is gratefully acknowledged.

References

Arrow, K.J. (1959), Rational choice functions and orderings, *Economica* 26: 121–127.

Baron, J. (2000), *Thinking and Deciding*, 3rd ed., Cambridge University Press: Cambridge, UK; New York.

Battigalli, P. (1996), The decision-theoretic foundations of game theory: Comment, in: K.J. Arrow, C. Schmidt and M. Perlman (eds.), *The Rational Foundations of Economic Behaviour: Proceedings of the IEA Conference held in Turin*, Macmillan Press: Hampshire, pp. 149–154.

Bernheim, D. (1984), Rationalizable strategic behavior, *Econometrica* 52: 1007–1028.

Bernheim, D. and Rangel, A. (2009), Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics, *Quarterly Journal of Economics*, 124: 51–104.

Bernheim, D. and Rangel, A. (2008), Choice-theoretic foundations for behavioral welfare economics, in: A. Caplin and A. Schotter (eds.), *The Foundations of Positive and Normative Economics: A Handbook*, Oxford University Press: Oxford, pp. 155–192.

Binmore, K. (2009), *Rational Decisions*, Princeton University Press: Princeton and Oxford.

Binmore, K. (2007a), *Does Game Theory Work? The Bargaining Challenge*, MIT Press: Cambridge, Mass.; London.

Binmore, K. (2007b), *Playing for Real*, Oxford University Press: New York.

Binmore, K. (2006), Why do people cooperate? *Politics, Philosophy & Economics* 5: 81–96.

Binmore, K. (1999), Why experiment in economics? *Economic Journal* 109: 16–24.

Binmore, K. (1998), *Game Theory and the Social Contract: Just Playing*, The MIT Press: London.

Binmore, K. (1997), Rationality and backward induction, *Journal of Economic Methodology* 4: 23–41.

Binmore, K. (1996), A note on backward induction, *Games and Economic Behavior* 17: 135–137.

Binmore, K. (1994), *Game Theory and the Social Contract: Playing Fair*, The MIT Press: London.

Binmore, K. (1987), Modeling rational players: Part 1, *Economics and Philosophy* 3: 179–214.

Binmore, K., McCarthy, J., Ponti, G., Samuelson, L. and Shaked, A. (2002), An experiment on backward induction, *Journal of Economic Theory* 104: 48–88.

Binmore, K. and Shaked, A. (2010), Experimental economics: Where next? *Journal of Economic Behavior & Organization* 73: 87–100.

Blackburn, S. (1995), Practical tortoise raising, *Mind* 104: 695–711.

Camerer, C.F. (2008), The case for mindful economics, in: A. Caplin and A. Schotter (eds.), *The Foundations of Positive and Normative Economics; A Handbook*, Oxford University Press: New York, pp. 44–69.

Carvajal, A., Ray, I. and Snyder, S. (2004), Equilibrium behavior in markets and games: Testable restrictions and identification, *Journal of Mathematical Economics* 40: 1–40.

Chapman, B. (2003), Rational choice and categorical reason, *University of Pennsylvania Law Review* 151: 1169–1210.

Dowding, K.M. (2002), Revealed preference and external reference, *Rationality and Society* 14: 259–284.

Drèze, J.H. and Rustichini, A. (2004), State-dependent utility and decision theory, in: S. Barberà, P.J. Hammond and C. Seidl (eds.) *Handbook of Utility Theory*, Kluwer Academic Publishers: Boston, pp. 839–892.

Fehr, E. and Schmidt, K.M. (1999), A theory of fairness, competition, and cooperation, *Quarterly Journal of Economics* 114: 817–868.

Gaertner, W. and Xu, Y.S. (1999), Rationality and external reference, *Rationality and Society* 11: 169–185.

Giocoli, N. (2003), *Modeling Rational Agents: From Interwar Economics to Early Modern Game Theory*, Edward Elgar: Cheltenham.

Grüne, T. (2004), The problem of testing preference axioms with revealed preference theory, *Analyse & Kritik* 26: 382–397.

Grüne-Yanoff, T. and Lehtinen, A. (forthcoming), Philosophy of game theory, in: U. Mäki (ed.) *Handbook of the philosophy of economics*, Elsevier: Amsterdam.

Guala, F. (2006), Has game theory been refuted? *Journal of Philosophy* 103: 239–263.

Gul, F. and Pesendorfer, W. (2008), The case for mindless economics, in: A. Caplin and A. Schotter (eds.), *The Foundations of Positive and Normative Economics; A Handbook*, Oxford University Press: New York, pp. 3–39.

Hammond, Peter J. (1996): Consequentialism, structural rationality and game theory, in: K.J. Arrow, E. Colombato and M. Perlman (eds.), *The Rational Foundations of Economic Behaviour: Proceedings of the IEA Conference held in Turin, Italy*, St. Martin's Press; Macmillan Press in association with the International Economic Association, New York; London, pp. 25–42.

Hands, D.W. (2012), Realism, commonsensibles, and economics: The case of contemporary revealed preference theory, in: A. Lehtinen, J. Kuorikoski and P. Ylikoski (eds.) *Economics for Real*, Routledge: London, pp. 156–178.

Hausman, D.M. (2011), Mistakes about preferences in the social sciences, *Philosophy of the Social Sciences* 41: 1–25.

Hausman, D.M. (2008), Mindless or mindful economics: A methodological evaluation, in: A. Caplin and A. Schotter (eds.) *The Foundations of Positive and Normative Economics; A Handbook*, Oxford University Press: New York, pp. 125–151.

Hausman, D.M. (2005a), Sympathy, commitment, and preference, *Economics and Philosophy* 21: 33–50.

Hausman, D.M. (2005b), 'Testing' game theory, *Journal of Economic Methodology* 12: 211–223.

Hausman, D.M. (2000), Revealed preference, belief, and game theory, *Economics and Philosophy* 16: 99–115.

Henrich, J., Boyd, R., Bowles, S., Camerer, C.F., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N.S., Hill, K., Gil-White, F., Gurven, M., Marlowe, F.W., Patton, J.Q. and Tracer, D. (2005), Economic man in cross-cultural perspective: Behavioral experiments in 15 small-scale societies, *Behavioral and Brain Sciences* 28: 795–815.

Hirshleifer, J. and Riley, J.G. (1992), *The Analytics of Uncertainty and Information*, Cambridge University Press: Cambridge.

Houthakker, H. (1950), Revealed preference and the utility function, *Economica* 17: 159–165.

Kadane, J.B., Levi, I. and Seidenfeld, T. (1992), Elicitation for games, in: C. Bicchieri and M.L. Dalla Chiara (eds.), *Knowledge, Belief and Strategic Interaction*, Cambridge University Press: Cambridge, pp. 21–26.

Karni, E. (2009), State-dependent preferences, in: P. Anand, P.K. Pattanaik and C. Puppe (eds.) *The Handbook of Rational & Social Choice*, Oxford University Press: New York, pp. 222–238.

Karni, E. (1999), Elicitation of subjective probabilities when preferences are state-dependent, *International Economic Review* 40: 479–486.

Kohlberg, E. and Mertens, J. (1986), On the strategic stability of equilibria, *Econometrica* 54: 1003–1037.

Mandler, M. (1999), *Dilemmas in Economic Theory: Persisting Foundational Problems of Microeconomics*, Oxford University Press: New York.

Mariotti, M. (1997), Decisions in games: Why there should be a special exemption from bayesian rationality, *Journal of Economic Methodology* 4: 43–60.

Mariotti, M. (1996), The decision-theoretic foundations of game theory, in: K.J. Arrow, C. Schmidt and M. Perlman (eds.), *The Rational Foundations of Economic Behaviour: Proceedings of the IEA Conference held in Turin*, Macmillan Press: Hampshire, pp. 133–148.

Mariotti, M. (1995), Is Bayesian rationality compatible with strategic rationality? *Economic Journal* 105: 1099–1109.

Mas-Colell, A. (1982), Revealed preference theory after Samuelson, in: G.R. Feiwel (ed.), *Samuelson and the Modern Economic Theory*, Kluwer: London, pp. 73–81.

Mongin, P. and d'Aspremont, C. (1998), Utility theory and ethics, in: S. Barberà, P.J. Hammond and C. Seidl (eds.), *Handbook of Utility Theory*, Kluwer Academic Publishers: Dordrecht, pp. 371–481.

Pearce, D.G. (1984), Rationalizable strategic behavior and the problem of perfection, *Econometrica* 52: 1029–1050.

Pettit, P. (1991), Decision theory and folk psychology, in: M. Bacharach and S. Hurley (eds.), *Decision Theory, Issues and Advances*, Blackwell: Oxford, pp. 147–175.

Ray, I. and Zhou, L. (2001), Game theory via revealed preferences, *Games and Economic Behavior* 37: 415–24.

Ross, D. (2006), Evolutionary game theory and the normative theory of institutional design: Binmore and behavioral economics, *Politics, Philosophy & Economics* 5: 51–79.

Ross, D. (2005), *Economic Theory and Cognitive Science: Microexplanation*, The MIT Press: Cambridge Mass.

Rubinstein, A. (1991), Comments on the interpretation of game theory, *Econometrica* 59: 909–924.

Rubinstein, A. and Salant, Y. (2008), Some thoughts on the principle of revealed preference, in: A. Caplin and A. Schotter (eds.), *The Foundations of Positive and Normative Economics; A Handbook*, Oxford University Press: New York, pp. 116–124.

Samuelson, P.A. (1963), Discussion, *American Economic Review* 53: 231–236.

Samuelson, P.A. (1950), The problem of integrability in utility theory, *Economica* 17: 355–385.

Samuelson, P.A. (1948), Consumption theory in terms of revealed preference, *Economica* 15: 243–253.

Samuelson, P.A. (1938), The empirical implications of utility analysis, *Econometrica* 6: 344–356.

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

Sen, A.K. (1997), Maximization and the act of choice, *Econometrica* 65: 745–779.

Sen, A.K. (1995): Is the idea of purely internal consistency of choice bizarre? in: J. Altham (ed.), *World, Mind, and Ethics*, Cambridge University Press, Cambridge, pp. 19–31.

Sen, A.K. (1993), Internal consistency of choice, *Econometrica* 61: 495–521.

Sen, A.K. (1987), Rational behavior, in: S. Eatwell, M. Milgate and P. Newman (eds.), *New Palgrave dictionary of economics*, Macmillan: New York, pp. 68–76.

Sen, A.K. (1977), Rational fools: A critique of the behavioral foundations of economic theory, *Philosophy & Public Affairs* 6: 317–344.

Sen, A.K. (1973), Behaviour and the concept of preference, *Economica* 40: 241–259.

Skyrms, B. (1998), Subjunctive conditionals and revealed preference, *Philosophy of Science* 65: 545–574.

Sprumont, Y. (2000), On the testable implications of collective choice theories, *Journal of Economic Theory* 93: 205–232.

Stigler, G.J. and Becker, G.S. (1977), De gustibus non est disputandum, *American Economic Review* 67: 76–90.

Sugden, R. (2001), Ken Binmore's evolutionary social theory, *Economic Journal* 111: 213–243.

Sugden, R. (1985), Why be consistent? A critical analysis of consistency requirements in choice theory, *Economica* 52: 167–183.

Von Neumann, J. and Morgenstern, O. (1947), *Theory of Games and Economic Behavior*, 2nd ed., Princeton University Press: Princeton.

Weibull, J.W. (2004), Testing game theory, in: S. Huck (ed.) *Advances in Understanding Strategic Behavior*, Palgrave: New York, pp. 85–104.

Wong, S. (1978), *The Foundations of Paul Samuelson's Revealed Preference Theory: A Study by the Method of Rational Reconstruction*, Routledge & Kegan Paul: London.