Running Head: PAYOFF DOMINANCE

ANDREW M. COLMAN AND MICHAEL BACHARACH

PAYOFF DOMINANCE AND

THE STACKELBERG HEURISTIC

ABSTRACT. Payoff dominance, a criterion for choosing between equilibrium points in games, is intuitively compelling, especially in matching games and other games of common interests, but it has not been justified from standard game-theoretic rationality assumptions. A psychological explanation of it is offered in terms of a form of reasoning that we call the Stackelberg heuristic in which players assume that their strategic thinking will be anticipated by their co-player(s). Two-person games are called Stackelberg-soluble if the players' strategies that maximize against their co-players' best replies intersect in a Nash equilibrium. Proofs are given that every game of common interests is Stackelberg-soluble, that a Stackelberg solution is always a payoff-dominant outcome, and that in every game with multiple Nash equilibria a Stackelberg solution is a payoff-dominant equilibrium point. It is argued that the Stackelberg heuristic may be justified by evidentialist reasoning.

1. INTRODUCTION

The concept of *payoff dominance* is widely accepted in non-cooperative game theory as a criterion for choosing between Nash equilibria. A simple example in a 2×2 game is shown in Table 1.

Table 1 about here

Player 1 and Player 2 both prefer the pure-strategy equilibrium A_1A_2, which yields payoffs of (2, 2), to the other pure-strategy equilibrium B_1B_2, which yields payoffs of (1, 1). The mixed-strategy equilibrium, with probabilities of 1/3 assigned to A_1 and A_2 and 2/3 to B_1 and B_2, yields even lower payoffs of (2/3, 2/3). The equilibrium A_1A_2 is said to *payoff dominate* B_1B_2 because each player receives a greater payoff if the outcome is A_1A_2 than if it is B_1B_2. The matrix in Table 1 is an example of a *pure coordination game* (e.g., Colman, 1995, chap. 3; Schelling, 1960, chap. 3), or more specifically a *unanimity game* (Kalai and Samet, 1985) or a *matching game* (Bacharach and Bernasconi, 1994), because (a) the players have the same strategy sets, and (b) the payoffs are positive if the players choose the same strategy and zero otherwise. Games of this type are the simplest possible exemplars of *games of common interests* - the class of games in which one outcome Pareto-dominates all other outcomes (Aumann and Sorin, 1989).

Intuitively, it seems obvious in the simple example of Table 1 that rational players will choose A rather than B and that the payoff-dominant equilibrium point A_1A_2 will therefore be the outcome, and there is general agreement in the game theory literature with this intuition (Bacharach, 1993; Crawford and Haller, 1990; Farrell, 1988; Gauthier, 1975; Harsanyi and Selten, 1988; Lewis, 1969; Sugden, 1995). In fact, payoff dominance is so obvious in games of this type that no experimental tests of it appear to have been published.

Formally, in any game $\Gamma = \langle N, S_i, H_i \rangle$, where $N = \{1, 2, ..., n\}$ is a set of players, $n \geq 2$, S_i ($i \in N$) is Player i's strategy set, $|S_i| \geq 2$, and H_i ($i \in N$) is a real-valued payoff function defined on the set $S = S_1 \times S_2 \times ... \times S_n$, if e and f are two Nash equilibria, *e payoff-dominates*

f iff $H_i(e) > H_i(f)$ for every player $i \in N$.

The *payoff dominance principle* is the assumption that if one equilibrium point e payoff-dominates all others, then rational players will play their parts in e. Harsanyi and Selten (1988) incorporated this principle (together with a secondary criterion called *risk dominance*) into their general theory of equilibrium selection in games, though apparently only provisionally and with some reluctance (see their comments on pp. 355-363), and others have used it in diverse branches of game theory. For example, Bacharach's (1993) variable-frame theory, when applied to matching games, provides determinate solutions only with the help of payoff dominance.

Granted that, as Harsanyi and Selten (1988) claim, rational players know that they should play their parts in some equilibrium point, each player is still faced with a problem of equilibrium *selection*. A player has no reason to choose a payoff-dominant equilibrium point unless there is some reason to believe that the other player(s) will do likewise; but the other player(s) face exactly the same quandary. Indeed, it seems that the underlying difficulty in justifying the claim that rational players will play their part in *some* equilibrium point has raised its head again at the level of equilibrium selection. No one has provided a convincing justification of the principle of payoff dominance or even of the weaker principle that in a game of common interests players should play their parts in the Pareto-optimal profile. All plausible attempts in the literature to rationalize A_1A_2 in the game shown in Table 1 have involved essential changes in the specification of the game, introducing either repetitions (e.g., Aumann and Sorin, 1989) or a "cheap talk" stage in which one or more players can choose to make a costless announcement before choosing between A and B (e.g., Anderlini, 1990; Farrell, 1988).

It seems from the record of failure to tease a justification of A out of the classical principles of game theory that there is none to be found in that quarter. But there are three avenues that offer some hope of explaining A as rational, either in some nonstandard sense or in a closely related game. The first would involve bounding Player 2's rationality: for instance, even a tiny probability that Player 2 is a "level 0" player (Stahl and Wilson, 1994)

induces *A* choices in all "level *n*" and "level *n* worldly Nash" players for $n > 0$. Closely related to this rationalization are "equiprobability" arguments for *A* (Hurley, 1991). Second, *A* can be explained as the product of cooperative reasoning or team thinking (Bacharach, 1995; Hurley, 1991; Sugden, 1993).

In this article we offer a third possible explanation of the intuition that it is rational to choose *A*. We hypothesize that players are influenced by a form of reasoning that we call the *Stackelberg heuristic*. The basic idea is that the players believe that their co-players can "read their minds". We show first that the Stackelberg heuristic yields maximin choices in strictly competitive games and *A* choices in the game shown in Table 1 and throughout the family of games that includes it; second that it is in equilibrium with itself in these games; and third that it may be rationalized by "evidentialist" reasoning. The validity of evidentialist reasoning is controversial, and we shall remain neutral on the question of the rationality of the Stackelberg heuristic. But in view of the fact that it is undoubtedly appealing in other cases, such as Newcomb's problem (Nozick, 1969), we suggest that, whatever its validity, it may underlie our intuitions of the rationality of choosing *A*.

2. RATIONALITY ASSUMPTIONS

A proposition is *common knowledge* among a set of players if every player knows it to be true, knows that the other players know it to be true, knows that the other players know that the other players know it to be true, and so on. The standard knowledge assumptions of game theory are called *complete information*. Complete information comprises the following pair of assumptions (see, e.g., Sugden, 1991, p. 765):

1. The specification of the game, including the rules, the players' strategy sets and payoff functions, and every proposition that can be proved about the game by valid reasoning, is common knowledge;

2. The players are rational in the sense of acting to maximize their individual expected utilities, and this is common knowledge in the game.

Complete information logically implies a further characteristic of game-theoretic

reasoning that Bacharach (1987) has called the *transparency of reason*. This is the property that if a player has reached some conclusion on the basis of complete information, for example about which strategy it would be best to choose, then the fact that the player has reached it must be known to the other player(s) – in fact, the conclusion itself must be common knowledge.

3. STRICTLY COMPETITIVE GAMES

Perhaps surprisingly, the argument used by von Neumann and Morgenstern (1944, section 14.4.1, pp. 100-104) to provide an a priori rational justification for selecting maximin strategies in finite, strictly competitive (two-person, zero-sum) games can be adapted to provide a psychological explanation for selecting payoff-dominant Nash equilibria in matching games and other games of common interests.

Von Neumann and Morgenstern approached the solution of strictly competitive games obliquely via the construction of two auxiliary models, slightly different from the basic game, which were later to be called *metagames* (Howard, 1971, 1974, 1987). For our purposes – for the argument on payoff dominance that follows – only "strictly determined" games whose payoff matrices have saddle points need to be considered, although the argument generalizes to non-saddle-point games. According to von Neumann and Morgenstern, Player 1 chooses as if playing in a first-level metagame in which Player 2 chooses second with the benefit of knowing which strategy Player 1 has chosen. In the extensive form of this metagame, rational choice is well defined: Player 2, moving second with perfect information of Player 1's earlier move, faces a straightforward decision under certainty and, being rational, chooses a payoff-maximizing reply to any of Player 1's moves, and because the game is zero-sum, Player 2's reply always minimizes Player 1's payoff. Player 1, knowing that Player 2 is rational and will invariably respond in this way with a best reply, therefore also chooses under certainty in the extensive form of the metagame.

Back in the basic normal-form game, according to this argument, Player 1, as if moving first in the metagame, ignores all elements of the payoff matrix $[a_{ij}]$ except the row minima

$\min_j a_{ij}$ and chooses a row containing the maximum of these minima, namely $\max_i \min_j a_{ij}$ (a *maximin* row). Because the game is zero-sum, the payoffs in the matrix $[a_{ij}]$ represent Player 1's gains and Player 2's losses; therefore Player 2, choosing as if playing in a metagame in which Player 1 moves second with foreknowledge of Player 2's move, chooses a column containing the minimum of the column maxima, namely $\min_j \max_i a_{ij}$ (a *minimax* column). Von Neumann and Morgenstern proved that, in every strictly determined game, these maximin and minimax strategies necessarily intersect in an equilibrium point. This equilibrium point, if it is unique, is generally accepted as the rational solution of the game.

4. EXTENSION TO NON-ZERO-SUM GAMES

In von Neumann and Morgenstern's (1944) classic treatment of strictly competitive games, the players choose as their strategies those that would maximize their payoffs in hypothetical metagames in which their co-players are assumed to respond with payoff-minimizing counter-strategies. This assumption, perfectly reasonable in the metagames but apparently ultra-pessimistic in the basic game, is justified on the ground that what is good for Player 1 is correspondingly bad for Player 2, and vice versa. In every two-person or *n*-person game of common interests, on the other hand, if *i* and *j* are any two players, what is good for Player *i* is correspondingly *good* for Player *j*, and vice versa, so the players have an analogous justification for assuming that their co-players will invariably respond with payoff-*maximizing* counter-strategies. In fact, the justification is stronger in this case, because in strictly competitive games the best that a player can hope to achieve through metagame rationality is to limit the damage by ensuring the best of the worst possible outcomes, and this motive may clash with more positive ambitions of obtaining the best possible outcome, whereas in non-zero-sum games of common interests such a conflict of motives is absent.

　　We give the name "Stackelberg heuristic" to the general method of reasoning that von Neumann and Morgenstern proposed for the particular case of strictly competitive games. Players use the Stackelberg heuristic if they choose strategies that would maximize their individual payoffs in hypothetical metagames in which their co-players are assumed to

respond with best replies to any choices that they might make.

A *game in normal form* is a triple $\langle N, S_i, H_i \rangle$, where $N = \{1, 2, ..., n\}$ is a set of players, $n \geq 2$, S_i ($i \in N$) is Player i's strategy set, $|S_i| \geq 2$, and H_i ($i \in N$) is a real-valued payoff function defined on the set $S = S_1 \times S_2 \times ... \times S_n$. Howard (1971) defined a metagame formally as follows. If Γ is a game in normal form, and if k is a player in Γ, then the (first-level) metagame $k\Gamma$ is the normal-form game that would be played if Player k's strategy choice in Γ occurred after those of the other players and with knowledge of the other players' strategy choices in Γ. This concept generalizes by recursion: if j, k are players, then the second-level metagame $jk\Gamma$ is the game that would be played if j chose a strategy in the first-level metagame $k\Gamma$ with knowledge of the other players' strategies in $k\Gamma$; and this can be continued up to $k_1 \ldots k_n\Gamma$, where each $k_i \in N$ is a player and $n \geq 1$, but for our purposes only first-level metagames need to be considered.

Table 2 about here

Consider once again the game shown in Table 1 and compare it with the normal form of its 2Γ metagame shown in Table 2. In the 2Γ metagame, Player 2's strategies are replaced by the set of all functions $F\colon S_1 \to S_2$, where $f(A_1) = x, f(B_1) = y$ is written x/y, $x, y \in \{A_2, B_2\}$. The strategies A_2/A_2, A_2/B_2, B_2/A_2, and B_2/B_2 are Player 2's pure strategies in the normal-form metagame 2Γ derived from the basic matching game Γ shown in Table 1. Thus if Player 1 chooses B_1 and Player 2 chooses A_2/B_2, for example, this means that Player 1 chooses B_1 and Player 2 chooses "If Player 1 chooses A_1 choose A_2, and if Player 1 chooses B_1 choose B_2"; these choices could be handed before the game to a referee who would determine that the net effect in the basic game is that Player 1 chooses B_1 and Player 2 chooses B_2. (In this respect metagame strategies function like ordinary strategies in sequential games with more than one move per player.) In this metagame, it is clear that Player 2, who is assumed to be rational, will choose the strategy A_2/B_2, because it is a *weakly dominant strategy*. It is also clear that Player 1 will anticipate this and will therefore choose the payoff-maximizing counter-strategy

A_1. In other words, by iterated deletion of weakly dominated strategies, the solution of the metagame 2Γ is the outcome $(A_1, A_2/B_2)$, which corresponds to the payoff-dominant outcome (A_1, A_2) in the basic game Γ, with payoffs (2, 2).

THEOREM 1. *The first-level metagame $j\Gamma$ of every finite two-person game Γ is soluble by iterated deletion of weakly dominated strategies. If Γ is a game of common interests, the resulting solution corresponds to the Pareto-optimal outcome.*

   Proof. In the first-level metagame $j\Gamma$ of every finite two-person game Γ, Player j's strategy set includes a weakly dominant strategy s. This must be so, because Player j's strategy set is the set F of all functions $f: S_i \to S_j$ ($i \neq j$, $i, j \in \{1, 2\}$). Player j's strategy set in the $j\Gamma$ metagame thus includes a strategy f' such that for each strategy s_i of Player i, $f'(s_i)$ is a best reply of Player j. Thus

$$H_j(s_i, f') \geq H_j(s) \quad \forall\ s \in S.$$

Ignoring the degenerate case in which $H_j(r)$ is constant for all $r \in S_i \times F$, f' is a weakly dominant strategy. Player i, who knows that Player j has a weakly dominant strategy f', can choose a counter-strategy for which $\max_i H_i(s_i, f'(s_i))$ is attained. This shows that $j\Gamma$ is soluble by iterated deletion of weakly dominated strategies. If Γ is a game of common interests, the solution corresponds to the Pareto-optimal outcome, because the payoffs pairs of $j\Gamma$ are elements of the set of payoff pairs of Γ, and since the game is one of common interests,

$$\exists\ (s^*, s^*): H_i(s^*, s^*) > H_i(s_1, s_2) \quad \forall\ i \in \{1, 2\},\ \forall\ (s_1, s_2) \neq (s^*, s^*).$$

Therefore $f'(s^*) = s^*$, and the s_i for which $\max_i H_i(s_i, f'(s_i))$ is attained is s^*, which shows that the solution is the Pareto-optimal outcome.   □

5. TRANSPARENCY OF DELIBERATION

The logic of the Stackelberg heuristic is essentially as follows. Common knowledge of rationality implies that each player knows the reasoning of the other, hence Player 2 knows Player 1's reasoning and Player 1 knows that Player 2 knows it. Now let C_1s_1 denote "Player 1 chooses strategy s_1", and let us assume what we shall call the *transparency of deliberation*:

$$\forall s_1 \in S_1, \ C_1s_1 \Rightarrow \text{Player 2 knows that } C_1s_1,$$

and the analogous version, *mutatis mutandis*, with the players' roles reversed. The transparency of deliberation is a stronger assumption than the transparency of reason, which applies only to players' strictly rational reasoning processes, but it is grounded in the same basic hypothesis, namely that human decision makers tend to be like-minded. Psychological investigations of stereotypes (see, e.g., Mackie and Hamilton, 1993; Oakes, Haslam, and Turner, 1994) have revealed a remarkable degree of consensus in people's understanding of their social environment, and research into attribution processes and social cognition (see, e.g., Fiske and Taylor, 1991; Hewstone, 1989; Schneider, 1995) has shown that the same basic cognitive processes underlie people's predictions and explanations of their own behaviour and that of others. The transparency of deliberation may be thought of as a psychological counterpart of the purely logical transparency of reason.

Mental processes of this type, underlying the Stackelberg heuristic, belong to a broader class of *simulation heuristics*, first identified by Kahneman and Tversky (1982), whereby people answer questions of various kinds about events through an operation resembling the running of a simulation model. The ease with which a mental model reaches a particular state may help a decision maker to judge the propensity of the actual situation to reach that outcome. Kahneman and Tversky provided empirical evidence that human decision makers use this heuristic to predict the behaviour of others in given circumstances and to answer questions involving counterfactual propositions by mentally "undoing" events that have occurred and then running mental simulations of the events with the corresponding input parameters of the model altered. In hypothetical metagames simulation plays two possible roles: simulation carried out by Player 2 may be the cognitive route by which Player 1's

deliberation becomes transparent to Player 2; and Player 1 may also use simulation to predict what would transpire if – counterfactually – the game were one of perfect information.

When appended to the assumptions of complete information, the transparency of deliberation implies that, in the basic game Γ, Player 2 effectively chooses as though playing in the metagame 2Γ. Because Player 1 knows this, and knows that Player 2 is rational, and because Player 1 is rational, the outcome is the payoff-dominant equilibrium. By symmetry, Player 1 effectively chooses as though playing in the metagame 1Γ, and because Player 2 knows this, the outcome is once again the payoff-dominant equilibrium.

6. STACKELBERG SOLUBILITY

Consider a game in normal form. We define the following best reply mapping β, which assigns a set of strategies for Player i to each strategy of the co-player j:

$$\beta(s_j) = \text{argmax } (s_i) \, H_i(s_i, s_j).$$

That is, the members of $\beta(s_j)$ are the strategies that maximize Player i's payoff given that Player j chooses the strategy s_j. In general, the best reply set $\beta(\cdot)$ may not be a singleton. But in the rest of this article we shall assume that it is, as indeed it is in every matching game. Henceforth, then, β is a function.

Following the work of Heinrich von Stackelberg (1934) on asymmetric duopoly games, we define Player 1's *Stackelberg payoff* for s_1, symbolized by $h(s_1)$, as

$$h(s_1) = H_1(s_1, \beta(s_1)),$$

and Player 2's Stackelberg payoff $h(s_2)$ for s_2 as

$$h(s_2) = H_2(\beta(s_2), s_2).$$

We shall assume that there is a unique strategy s_i for Player i ($i = 1, 2$) that maximizes Player i's Stackelberg payoff. We call this strategy Player i's *Stackelberg strategy s*. That is,

$$h(s_i) \text{ has a unique maximizer } s.$$

We make this assumption even though it does not hold for all matching games, because it avoids complications that are inessential to our argument. If (s, s) is a Nash equilibrium in the basic game Γ, then Γ will be called *Stackelberg-soluble* (or *h-soluble*) and (s, s) will be

called its Stackelberg solution (or h solution). This will be the case if and only if iterated deletion of weakly dominated strategies in the $j\Gamma$ metagame yields an equilibrium point that corresponds to an equilibrium point in the basic game.

THEOREM 2. *Every game of common interests is h-soluble, and its Pareto-optimal outcome is its h solution.*

   Proof. Let the Pareto-optimal outcome be (s^*, s^*). Because it is Pareto-optimal, s^* is a best reply to s^*. It follows that $H_1(s^*, s^*) = H_1(s^*, \beta(s^*)) = h(s^*)$. But because (s^*, s^*) is Pareto-optimal, $H_1(s^*, s^*)$ is the greatest payoff to Player 1 over all pairs of strategies, so $h(s^*)$ is the maximum of $h(s_1)$ over all s_1; that is, s^* maximizes $h(s_1)$. Similarly, s^* maximizes $h(s_2)$. Therefore, in view of the fact that (s^*, s^*) is a Nash equilibrium, it is the h-solution (s, s) of the game.   □

THEOREM 3. *In every game with more than one Nash equilibrium, a Stackelberg solution is a payoff-dominant Nash equilibrium.*

   Proof. If (s_1, s_2) is any Nash equilibrium, then $s_2 = \beta(s_1)$, and therefore $H_1(s_1, s_2) = h(s_1)$, Player 1's Stackelberg payoff for s_1. Similarly, $H_2(s_1, s_2) = h(s_2)$, Player 2's Stackelberg payoff for s_2. Now suppose that (s, s) is a Stackelberg solution and (s_1, s_2) is any other Nash equilibrium. Because Player 1's Stackelberg strategy s uniquely maximizes $h(s_1)$, $h(s) > h(s_1)$. Similarly $h(s) > h(s_2)$. Therefore (s, s) payoff-dominates (s_1, s_2).   □

   Remark. Theorem 3 implies that the payoff-dominance principle of equilibrium selection is a corollary of the Stackelberg heuristic, restricted to games in which this heuristic is in equilibrium with itself. That is, in any game with multiple equilibria that is Stackelberg-soluble, players who follow the heuristic play their parts in the payoff-dominant equilibrium point.

   The next two theorems show that h solubility is a broader property than common interests and a narrower one than payoff dominance.

THEOREM 4. *There are h-soluble games that are not games of common interests.*

   Proof. The game shown in Table 3 has Stackelberg payoffs as follows: $h(A_1) = 2$, $h(B_1) = 1$, $h(A_2) = 2$, $h(B_2) = 0$. There is a Nash equilibrium at (A_1, A_2), and it is evidently a Stackelberg solution because if both players choose strategies that maximize the Stackelberg payoffs, giving max (s_1) $h(s_1)$ and max (s_2) $h(s_2)$, the outcome is (A_1, A_2), but the game is not one of common interests. $\square$

Table 3 about here

THEOREM 5. *There are games with payoff-dominant Nash equilibria that are not h solutions.*

   Proof. The game shown in Table 4 has the following Stackelberg payoffs: $h(A_1) = 2$, $h(B_1) = 3$, $h(C_1) = 1$, $h(A_2) = 2$, $h(B_2) = 0$, $h(C_2) = 1$. There are two pure-strategy Nash equilibria at (A_1, A_2) and (C_1, C_2). The equilibrium (A_1, A_2) payoff-dominates (C_1, C_2), but the Stackelberg strategies are $s = B_1$ and $s = A_2$, and $(s, s) = (B_1, A_2)$ is not a Nash equilibrium and therefore not a Stackelberg solution. $\square$

Table 4 about here

7. DISCUSSION

Payoff dominance seems a highly plausible criterion for equilibrium selection, especially in matching games and other games of common interests. In this article we have presented an explanation of its plausibility in terms of a form of reasoning that we have called the Stackelberg heuristic. The Stackelberg heuristic does not explain payoff dominance for games in general, or even for two-person games in general because, as we have shown, there are games with payoff-dominant Nash equilibria that are not Stackelberg-soluble. The

explanation applies primarily to games of common interests, although we have shown that there are other games with Stackelberg solutions apart from games of common interests.

Underlying the heart of the explanation is the assumption that Player j will try to predict Player i's strategy choice and that Player i will expect this to happen and will try to maximize utility in the light of this expectation. In effect, the players will use a type of simulation heuristic, which has been studied in a different context by Kahneman and Tversky (1982). It is widely accepted (see, e.g., Bacharach, 1987) that given certain standard game-theoretic rationality assumptions, if there is a logically valid argument for Player i to choose a strategy, then Player j will predict it and Player i will expect that to happen. This arises from the transparency of reason, which in turn derives from the fact that human decision makers share the same reason. But the argument for payoff dominance presented in this article rests on a stronger assumption, which we call the transparency of deliberation, namely that, whatever is Player i's reasoning path to a strategy choice, whether logically valid or not, Player j will "discover" it.

The full or at least partial transparency of deliberation seems a reasonable assumption, and it rests on a hypothesis similar to that of the transparency of reason: that human decision makers share largely the same underlying cognitive structures and dispositions. The evidence from research into attribution processes and social cognition has already been alluded to. In addition, empirical evidence from experiments on matching games has shown that people, to their mutual benefit, are able to coordinate their strategies remarkably easily in practice. For example, when pairs of experimental subjects were invited to choose "heads" or "tails" independently, knowing that they will both win only if they both choose "heads" or both choose "tails", Schelling (1960, chap. 3) found in the United States that 86 per cent chose "heads", and Mehta, Starmer, and Sugden (1994) found in England that 87 per cent chose "heads". Even more remarkably, Schelling reported that pairs of subjects who imagined that they had to meet each other at a particular place, but knew that neither had been given a time for the meeting, virtually all chose 12 noon, thus correctly anticipating each other's choices through the transparency of deliberation.

We have argued that the Stackelberg heuristic may explain payoff-dominant choices, but so far we have offered no serious defence for using it. We turn now to the question of whether it is justified.

At first sight it may seem not to be. Its application to strictly competitive games by von Neumann and Morgenstern (1944, section 14.4.1) met early criticism on the simple ground that Player *i*'s premiss that any strategy choice will be discovered by Player *j* – that is, that the game is one of perfect information – is false, by definition, in a simultaneous-play game (Ellsberg, 1956). But this objection is unsound, because all that the definition excludes is that the rules of the game prescribe that Player *i*'s choice will be conveyed to Player *j* before *j* chooses a strategy. The definition does not exclude the possibility that Player *j* might discover Player *i*'s choice by "mind-reading"; indeed it is an essential assumption of rational game theory that *valid* reasoning *is* transparent, so why should not all reasoning, whether valid or invalid, be similarly transparent?

A more serious objection is that Player *i* reasons that *any* choice will be discovered by Player *j*. In one sense this may be true: Player *i* may indeed justifiably believe, before deliberating, that "whatever I come up with by deliberating, Player *j* will have anticipated it". But it does not follow that whatever choice Player *i* were to make, Player *j* would in fact anticipate it. Player *j* would not, for example, anticipate a choice that was quite unreasonable or capricious. Yet Player *i*'s choice ranges over *all* strategies permitted by the rules.

The key question is whether the evidence that an arbitrary decision by Player *i* provides of how Player *j* will decide may be used by *i* to formulate a choice. There is no doubt, if we grant transparency and Player *j*'s best-reply rationality, that a decision by Player *i* to do s_i, arrived at by the sort of process that makes it transparent to *j*, is evidence for *j*'s choosing $\beta(s_i)$. Whether such evidence may legitimately be used by Player *i* is the issue that divides "evidentialists" and "causalists" in decision theory (Eells, 1985; Gibbard and Harper, 1978; Nozick, 1969).

Underlying the Stackelberg heuristic is evidentialist reasoning. Player *i* evaluates each act s_i according to conditional expected utility given the hypothesis that the choice is s_i. So

the evaluation of s_i is

$$E(u \mid s_i) = \Sigma(s_j) \Pr(s_j \mid s_i) H_i(s_i, s_j) = H_i(s_i, \beta(s_i)) = h(s_i),$$

because by transparency $\Pr(\beta(s_i) \mid s_i) = 1$. Against this way of evaluating actions, "causalists" claim that s_i should be judged by its probable consequences, and Player i's choice is causally independent of Player j's, so Player i cannot validly use a strategy choice s_i as a basis on which to reach any conclusion about the likely outcome of the game. It is interesting to note that this evidentialist justification of payoff dominance differs from another possible evidentialist justification, discussed by Lewis (1979), for playing cooperatively in the Prisoner's Dilemma game. In Lewis's argument, s_i provides evidence about s_j, and the evidence rests on the similarity of the way the two players reason and the symmetry of their joint situation. In the present argument the evidence rests on the transparency to Player j of Player i's reasoning. This argument is therefore of much wider scope, because it depends only on similarity of reasoning (the basis of transparency) and not on the symmetry of the situation.

Von Neumann and Morgenstern (1944, section 14.4.1) struggled to find a serious argument for their advocacy of the Stackelberg heuristic. We have suggested one. Valid or not, our evidentialist argument for the Stackelberg heuristic is what von Neumann and Morgenstern called a "direct argument". It is to be distinguished from the "indirect argument" that they used to show that the correct solution concept for a class of games must be an equilibrium (section 17.3.3). The indirect argument provides a test of the validity of a direct argument. Applied to the Stackelberg heuristic, the indirect argument shows that it provides a correct solution only in the subclass of games that are Stackelberg-soluble. It is precisely for that subclass that we hypothesize that it is used.

The Stackelberg heuristic is a form of the simulation heuristic in which people try to predict one another's behaviour in strategic interactions and to choose their own best strategies on the basis of these predictions. In these cases, they may appear to be acting as though their merely evidential actions were causal. Quattrone and Tversky (1984) showed experimentally that people tend to select actions that are associated with desired outcomes

even when they know that these actions are merely evidential or diagnostic of the outcomes and not causally connected with them. In the key experiment, subjects expressed a greater willingness to vote in an election when they believed that the outcome would depend on the proportion of like-minded voters who turned out on polling day rather than on the behaviour of non-aligned voters, even though the effect of one person's vote would be negligible (and equal) in both cases. The subjects thus behaved as though their own actions could somehow "induce" like-minded people to exercise their right to vote. Furthermore, subjects predicted that their preferred candidate would be significantly more likely to win the election if they themselves voted, and the strength of this perceived association correlated substantially ($r = .32$, $p < .001$) with their willingness to vote. These findings show that, in some circumstances at least, people behave as though their actions are causal even when they know them to be merely diagnostic or evidential.

Although it is clearly irrational to believe that an action that is merely evidential can be causal, it is not obviously foolish to behave in such a way that, if other people were to behave similarly, a mutually desirable outcome would result. In strategic interactions, a common-sense way of choosing a strategy often involves first predicting what the other players are likely to do, assuming that they will expect others to do likewise, and so forth, and then selecting the optimal reply in the light of these predictions and assumptions. This seems especially justifiable when other reasons for choice are lacking. In matching games and other games of common interests, at least, this approach may be implemented by choosing Stackelberg strategies. If players in such games do not use the Stackelberg heuristic, then it is not obvious what better method they have for choosing. We have proved that every Stackelberg solution is a payoff-dominant Nash equilibrium. Payoff-dominant Nash equilibria are intuitively obvious choices, and there must be some reason for this. Our suggestion of the Stackelberg heuristic is an attempt to clarify this essentially psychological phenomenon and to explain how experimental subjects manage to choose payoff-dominant equilibria in games in which there is no basis for rational choice according to standard game-theoretic assumptions.

ACKNOWLEDGEMENTS

REFERENCES

Anderlini, L.: 1990, `Communication, Computability and Common Interest Games', working paper, St John's College, Cambridge.

Aumann, R. J. and Sorin, S.: 1989, `Cooperation and Bounded Recall', *Games and Economic Behavior*, **1**, 5-39.

Bacharach, M.: 1987, `A Theory of Rational Decision in Games', *Erkenntnis*, **27**, 17-55.

Bacharach, M.: 1993, `Variable Universe Games', in K. Binmore, A. Kirman, and P. Tani (Eds), *Frontiers of Game Theory* (pp. 255-275), MIT Press, Cambridge, MA.

Bacharach, M.: 1995, `Cooperating Without Communicating', working paper, Institute of Economics and Statistics, University of Oxford.

Bacharach, M. and Bernasconi, M.: 1994, `An Experimental Study of the Variable Frame Theory of Focal Points', working paper, Institute of Economics and Statistics, University of Oxford.

Colman, A. M.: 1995, *Game Theory and its Applications in the Social and Biological Sciences*, Butterworth-Heinemann, Oxford.

Crawford, V. P. and Haller, H.: 1990, `Learning How to Cooperate: Optimal Play in Repeated Coordination Games', *Econometrica*, **58**, 571-595.

Eells, E.: 1985, `Causality, Decision, and Newcomb's Paradox, in R. Campbell and L. Sowden (Eds), *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem* (pp. 183-213), University of British Columbia Press, Vancouver.

Ellsberg, D.: 1956, `Theory of the Reluctant Duellist' *American Economic Review*, **46**, 909-923.

Farrell, J.: 1988, `Communication, Coordination and Nash Equilibrium', *Economics Letters*, **27**, 209-214.

Fiske, S. T. and Taylor, S. E.: 1991, *Social Cognition* (2nd ed.), McGraw-Hill, New York.

Gauthier, D.: 1975, `Coordination', *Dialogue*, **14**, 195-221.

Gibbard, A. and Harper, W. L.: 1978, `Counterfactuals and Two Kinds of Expected Utility', in C. Hooker, J. Leach, and E. McClennen (Eds), *Foundations and Applications of*

Decision Theory (Vol. 1, pp. 125-162), D. Reidel, Dordrecht-Holland.

Harsanyi, J. C. and Selten, R.: 1988, *A General Theory of Equilibrium Selection in Games*, MIT Press, Cambridge, MA.

Hewstone, M.: 1989, *Causal Attribution: From Cognitive Processes to Collective Beliefs*, Basil Blackwell, Oxford.

Howard, N.: 1971, *Paradoxes of Rationality: Theory of Metagames and Political Behavior*, MIT Press, Cambridge, MA.

Howard, N.: 1974, `"General" Metagames: An Extension of the Metagame Concept', In A. Rapoport (Ed.), *Game Theory as a Theory of Conflict Resolution* (pp. 261-283), D. Reidel, Dordrecht-Holland.

Howard, N.: 1987, `The Present and Future of Metagame Analysis', *European Journal of Operational Research*, **32**, 1-25.

Hurley, S. L.: 1991, `Newcomb's problem, Prisoner's Dilemma, and Collective Action', *Synthese*, **86**, 173-196.

Kahneman, D. and Tversky, A.: 1982, `The Simulation Heuristic', in D. Kahneman, P. Slovic, and A. Tversky (Eds), *Judgment Under Uncertainty: Heuristics and Biases* (pp. 201-208), Cambridge University Press, Cambridge, UK.

Kalai, E. and Samet, D.: 1985, `Unanimity Games and Pareto Optimality', *International Journal of Game Theory*, **14**, 41-50.

Lewis, D. K.: 1969, *Convention: A Philosophical Study*, Harvard University Press, Cambridge, MA.

Lewis, D. K.: 1979, `Prisoner's Dilemma is a Newcomb Problem', *Philosophy and Public Affairs*, **8**, 235-240.

Mackie, D. M. and Hamilton, D. L. (Eds): 1993, *Affect, Cognition, and Stereotyping: Interactive Processes in Group Perception*, Academic Press, San Diego, CA.

Mehta, J., Starmer, C., and Sugden, R.: 1994, `Focal Points in Pure Coordination Games: An Experimental Investigation', *Theory and Decision*, **36**, 163-185.
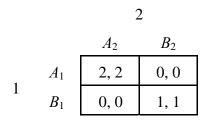
Neumann, J. von and Morgenstern, O.: 1944, *Theory of Games and Economic Behavior*,

Princeton University Press, Princeton, NJ. [2nd ed., 1947; 3rd ed., 1953]

Nozick, R.: 1969, `Newcomb's Problem and Two Principles of Choice', in N. Rescher (Ed.), *Essays in Honor of Carl G. Hempel: A Tribute on his Sixty-Fifth Birthday* (pp. 114-146), D. Reidel, Dordrecht-Holland.

Oakes, P. J., Haslam, S. A., and Turner, J. C.: 1994, *Stereotyping and Social Reality*, Blackwell, Oxford.

Quattrone, G. A. and Tversky, A.: 1984, `Causal Versus Diagnostic Contingencies: On Self-deception and the Voter's Illusion', *Journal of Personality and Social Psychology*, **46**, 237-248.

Schelling, T. C.: 1960, *The Strategy of Conflict*, Harvard University Press, Cambridge, MA.

Schneider, D. J.: 1995, `Attribution and Social Cognition', in M. Argyle and A. M. Colman (Eds.), *Social Psychology* (pp. 38-56), Longman, London.

Stackelberg, H. von: 1934, *Marktform und Gleichgewicht*, Julius Springer, Vienna and Berlin.

Stahl, D. O. and Wilson, P. W.: 1994, `Experimental Evidence on Players' Models of Other Players', *Journal of Economic Behavior & Organization*, **25**, 309-327.

Sugden, R.: 1991, `Rational Choice: A Survey of Contributions from Economics and Philosophy', *The Economic Journal*, **101**, 751-785.

Sugden, R.: 1993, `Thinking as a Team: Towards an Explanation of Nonselfish Behavior', *Social Philosophy & Policy*, **10**, 69-89.

Sugden, R.: 1995, `A Theory of Focal Points', *The Economic Journal*, **105**, 533-550.

Table 1

A Simple 2 × 2 Game Illustrating Payoff Dominance

		A_2	B_2
		2, 2	0, 0
	A_1		
	B_1	0, 0	1, 1

Table 1

A Simple 2 × 2 Game Illustrating Payoff Dominance

2

		A_2	B_2
1	A_1	2, 2	0, 0
	B_1	0, 0	1, 1

Table 2

The 2Γ Metagame of the Game in Table 1.

		2			
		A_2/A_2	A_2/B_2	B_2/A_2	B_2/B_2
1	A_1	2, 2	2, 2	0, 0	0, 0
	B_1	0, 0	1, 1	0, 0	1, 1

Note. Player 2 chooses after Player 1 with the benefit of knowing which strategy Player 1 has chosen. Player 2's conditional strategy x/y means "if Player 1 chooses A_1, choose x; if Player 1 chooses B_1, choose y", where $x, y \in \{A_2, B_2\}$.

Table 3

A Game that is h-Soluble but is Not a Game of Common Interests

	A_2	B_2
A_1	2, 2	3, 1
B_1	1, 3	4, 0

(Player 2 columns, Player 1 rows)

Table 4

A Game with a Payoff-Dominant Nash Equilibrium that is Not a Stackelberg

Solution

		2		
		A_2	B_2	C_2
	A_1	2, 2	1, 1	0, 0
1	B_1	1, -1	3, 0	0, -1
	C_1	1, 0	4, 0	1, 1

Andrew M. Colman,

Department of Psychology,

University of Leicester,

Leicester, LE1 7RH,

UNITED KINGDOM.

Michael Bacharach,

Institute of Economics and Statistics,

St Cross Building,

Manor Road,

OXFORD, OX1 3UL.