Cowles Foundation Paper 726

GAMES AND ECONOMIC BEHAVIOR **1**, 60–79 (1989)

Psychological Games and Sequential Rationality

JOHN GEANAKOPLOS AND DAVID PEARCE

Yale University, New Haven, Connecticut 06520

AND

ENNIO STACCHETTI

Stanford University, Stanford, California 94305

In psychological games the payoff to each player depends not only on what every player does but also on what he thinks every player believes, and on what he thinks they believe others believe, and so on. In equilibrium, beliefs are assumed to correspond to reality. Yet psychological games and psychological equilibria allow one to model belief-dependent emotions such as anger and surprise that are problematic for conventional game theory We are particularly interested in issues of sequential rationality for psychological games. We show that although backward induction cannot be applied, and "perfect" psychological equilibria may not exist, subgame perfect and sequential equilibria always do exist.

1. INTRODUCTION

Principles such as subgame perfection (Selten, 1965) and sequential rationality (Kreps and Wilson, 1982) serve, among other things, to restrict the implicit threats that a player can use to his advantage in an extensive form game. A threat that a certain choice will be made at some information set should be disregarded if it would not be in the interests of the player to make that choice were the information set actually reached. One sometimes hears this view challenged on the grounds that a player might carry out a threat (in violation of a sequential rationality constraint) as a matter of pride, or for the sheer joy of retaliation. The standard response is that if players are motivated by such emotional considerations, these

0899-8256/89 $3.00

should be reflected in the payoffs at the terminal nodes of the game tree. As long as the correct utility function is employed, issues of credibility are appropriately addressed by the usual solution concepts. While this answer is adequate in the simplest examples, beyond these it is misleading. A player's emotional reactions cannot in general be independent of his expectations and of his interpretation of what he learns in a play of a game. Hence, we argue that in many cases the psychological payoffs associated with a terminal node are endogenous, in the same sense as equilibrium strategies are. Indeed, in some examples, no single set of payoffs adequately summarizes the strategic situation.

Similar complications can arise whenever players' utilites depend not only on the physical outcome of a game but also on their *beliefs* before or during play (even in simultaneous games). Consequently, the traditional theory of games is not well suited to the analysis of such belief-dependent psychological considerations as surprise, confidence, gratitude, disappointment, embarrassment, and so on. The purpose of this paper is to develop a framework more general than an extensive form game, that, together with appropriately modified solution concepts, allows for a satisfactory treatment of a broad range of psychological phenomena. The principal distinguishing characteristic of what we call a *psychological game* is that the players' payoffs depend not only on what everybody does but also on what everybody thinks. More precisely, each player's payoff depends on his *hierarchy* of beliefs. A player's beliefs specify what he thinks will happen (that is, a probability measure over the product of others' strategy spaces), what he thinks each other player thinks will happen, and so on. Examples presented later show that it is natural for higher order beliefs to appear in the utility functions.

In equilibrium we shall require that beliefs correspond to reality. Nevertheless, their presence in the payoffs strictly enlarges the domain of game theory. Furthermore, it prevents the application of backward induction as a method for solving for credible equilibria in extensive form games. Trembling hand perfect equilibria (Selten, 1975) do not always exist for psychological games. Our main result, however, is that much of the rest of equilibrium theory *can* be maintained for psychological games. The analog of Nash equilibrium exists under the same kind of conditions as it does in conventional game theory. Despite the failure of backward induction, so do subgame perfect equilibria and sequential equilibria.

An important precedent for our work is provided by Gilboa and Schmeidler (1988). In their *information-dependent games,* a player's utility can depend upon his prior knowledge that the outcome will lie in a particular subset of outcome space. These sets parametrize utilities just as the players' belief hierarchies do in a psychological game. Gilboa and Schmeidler give examples that illustrate how emotions (including the ef-

fects of surprise, revenge, and fashion-consciousness) can be modeled in their framework. Their main concern, however, is to illuminate certain paradoxes of common knowledge, whereas we are interested in exploring the logic of sequential rationality in psychological games and proving existence of solutions for games whose payoffs depend on belief hierarchies.[1] Our solutions are analogous to traditional equilibrium notions. This is quite different from the approach Gilboa and Schmeidler take. Their idea of an "informationally consistent play" combines axioms concerning individual rationality with a fixed point requirement. Less closely related, but also treating emotional factors in strategic analysis, is a recent paper by Nalebuff and Shubik (1988).

We begin in Section 2 by studying normal form games only. This introduces some of the essential ideas associated with psychological games, while avoiding the complexities of the extensive form. Recall that in traditional equilibrium analysis, the equilibrium strategy profile is taken to be common knowledge among the players. If the same hypothesis is adopted for psychological games, a single profile generates all players' beliefs of all orders. Using this observation, we construct a "summary form" having dimensions drastically lower than those of the original psychological game, but capturing precisely the information needed to compute Nash equilibria or equilibrium refinements. Existence of equilibrium is established under fairly general conditions.

Several fully specified examples are presented in Section 2 using the framework and notation developed there, but even an impressionistic description of one psychological game may be helpful to readers at this point. Think of a two-person game in which only player 1 moves. Player 1 has two options: she can send player 2 flowers, or she can send chocolates. She knows that 2 likes either gift, but she enjoys surprising him. Consequently, if she thinks player 2 is expecting flowers (or that he thinks flowers more likely than chocolates), she sends chocolates, and vice versa. No equilibrium in pure strategies exists. In the unique mixed strategy equilibrium, player 1 sends each gift with equal probability. Note that in a traditional finite game with only one active player, there is always a pure strategy Nash equilibrium. That this is untrue in psychological games demonstrates the impossibility of analyzing such situations merely by modifying the payoffs associated with various outcomes: any modification will yield a game with at least one pure strategy equilibrium.

Section 3 introduces psychological games in extensive form. In this setting we can address "credible threat" questions of the kind mentioned

[1] Gilboa and Schmeidler note that a natural extension of their model is to consider knowledge about others' knowledge, and so on. Our research was conducted independently of theirs.

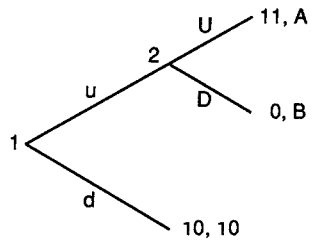

FIGURE 1

earlier. An elementary example is illustrated in Fig. 1. Player 1 cares only about physical outcomes, whereas 2's payoffs at two of the terminal nodes depend upon his initial expectations. If he is resigned at the beginning of the game to the idea that 1 will choose u, then $A = 5$ and $B = 1$, and 2 chooses U if reached (one could imagine that this corresponds to his choosing the greater monetary reward). Thus, in one credible equilibrium, the pair $(u; U)$ is played. But if 2 is confident ex ante that 1 will choose d, 2 would be bitterly disappointed if reached, and, in his fury, would choose D because it harms 1 (say, $A = 0$ and $B = 2$). One sees that this corresponds to another credible equilibrium, with choices $(d; D)$. In neither equilibrium is either player indifferent about his choice. In contrast, backward induction yields unique outcomes in traditional games except when there are ties in payoffs. The failure of backward induction in psychological games results from the fact that when a node is reached, it does not capture adequately the state of the game: the node identifies a history of play, but not the players' beliefs.

The inapplicability of backward induction implies that the usual proof of the existence of a subgame perfect equilibrium cannot be translated to psychological games. We show, however, that these games always have equilibria analogous to subgame perfect equilibria and sequential equilibria, respectively. Thus, there is no tension between the usual notions of sequential rationality and the presence of psychological influences on players' behavior.

Some examples are worked out at the end of Section 3; Section 4 contains brief concluding remarks.

2. NORMAL FORM PSYCHOLOGICAL GAMES

In this section we make no use of the extensive form of a game except insofar as it describes the strategies available to each player. Nevertheless, many of the novel features of psychological games appear even in

this normal form setting; examples are presented once the necessary definitions have been given. The natural analog to Nash equilibrium is shown to exist under relatively modest assumptions. We turn now to the formal definitions.

Let $N = \{1, \ldots, n\}$ be the set of players, and for each $i \in N$, let A_i be the nonempty, finite set of actions available to player i. For any set X (where the topology of X is understood), $\Delta(X)$ denotes the set of (Borel) probability measures on X. Thus, $\Sigma_i := \Delta(A_i)$ is the set of mixed strategies of player i. Let $\Sigma := \times_{i \in N} \Sigma_i$ and $\Sigma_{-i} := \times_{j \neq i} \Sigma_j$, $i \in N$. Each strategy profile $\sigma \in \Sigma$ induces a probability distribution P_σ over the outcome set (or set of pure strategy profiles) $A := \times_{i \in N} A_i$.

A first-order belief for a player is a probability measure over the product of the other players' mixed strategy sets. Thus, the set of first-order beliefs of player i is $B_i^1 := \Delta(\Sigma_{-i})$. Let $B_{-i}^1 := \times_{j \neq i} B_j^1$ and $B^1 := \times_{i \in N} B_i^1$. Endow B_i^1 with the weak topology. Since Σ_{-i} is a subset of a Euclidean space, it is a separable metric space, and therefore B_i^1 is a separable metric space. The sets of higher order beliefs are defined inductively for $k \geq 1$ by

$$B_i^{k+1} := \Delta(\Sigma_{-i} \times B_{-i}^1 \times \cdots \times B_{-i}^k),$$

$$B_{-i}^{k+1} := \times_{j \neq i} B_j^{k+1}, \qquad B^{k+1} := \times_{i \in N} B_i^k,$$

where for each k, B_i^{k+1} is endowed with the weak topology. The set of player i's beliefs is then $B_i := x_{k=1}^{\infty} B_i^k$; endow it with the product topology. This structure is familiar from the literature on games of incomplete information (see, for example, Harsanyi, 1967–1968; Mertens and Zamir, 1985; Brandenburger and Dekel, 1985).

Note that each piece of information appears many times in the belief hierarchy of player i. For example, a second-order belief is a probability measure over $\Sigma_{-i} \times B_{-i}^1$ (this allows for correlation between beliefs about others' actions and beliefs about others' first-order beliefs). Consequently, one can compute the marginal of the second-order beliefs with respect to Σ_{-i}; unless the beliefs are nonsensical, this should coincide with i's first-order beliefs. We say that beliefs satisfying the appropriate marginal restrictions are *coherent*. If X and Y are two spaces and $\delta \in \Delta(X \times Y)$, we denote by $\text{marg}(\delta; X)$ the marginal probability measure of δ on X: $\text{marg}(\delta; X)(E) := \delta(E \times Y)$ for each measurable set $E \subset X$.

DEFINITION. $b_i = (b_i^1, b_i^2, \ldots) \in \times_{k=1}^{\infty} B_i^k = B_i$ is *coherent* if for each $k \geq 1$, $\text{marg}(b_i^{k+1}, \Sigma_{-1} \times B_{-i}^1 \times \cdots \times B_{-i}^{k-1}) = b_i^k$. Denote by $\hat{B}_i(0)$ the set of player i's coherent beliefs.

Since each player knows that the other players are also rational, he should not believe that they may entertain beliefs that are not coherent, or

that they may believe that others may entertain beliefs that are not coherent, and so on. That is, coherency should be common knowledge. The set of beliefs for player i in which he is sure that it is common knowledge that beliefs are coherent is denoted $\bar{B}_i$; beliefs in $\bar{B}_i$ are said to be *collectively coherent.*[2] Let $\bar{B} := \times_{i \in N} \bar{B}_i$. For an extensive discussion on common knowledge of coherency, and alternative characterizations, the reader is directed to Brandenburger and Dekel (1985).

Player i's *utility function* $\bar{u}_i : \bar{B}_i \times A \to \mathbb{R}$ depends on the outcomes (as in the standard literature) and *also* on his beliefs. We assume that player i seeks to maximize the expected value of $\bar{u}_i$. Extend $\bar{u}_i$ to $u_i : \bar{B}_i \times \Sigma \to \mathbb{R}$ by

$$u_i(b_i, \sigma) := \Sigma_{t \in A} P_\sigma(t) \bar{u}_i(b_i, t) \qquad \text{for each } b_i \in \bar{B}_i \text{ and } \sigma \in \Sigma.$$

We interpret $u_i(b_i, \sigma)$ to be the payoff to player i if he believed b_i and then found out that σ was actually played. As in the standard theory of games, the payoffs for player i are defined first on the outcomes (given any belief profile b) and only afterward extended by taking expectations to all of the mixtures included in Σ. The reason is that player i is presumed not to observe the mixture used by any player $j \neq i$. On the other hand, there is no requirement that the payoffs be linear in the beliefs. In particular, player i may get the same payoff if he expects j to deliver chocolates, and she does, as he would if he expected her to deliver flowers and she did. Yet his payoff might be very different if he expected her to randomize and she did (perhaps because he might think she is indecisive).

DEFINITION. A *normal form psychological game* $G = (A_1, \ldots, A_n; u_1, \ldots, u_n)$ consists of an action set A_i and a utility function $u_i : \bar{B}_i \times \Sigma \to \mathbb{R}$ for each player i.

In general, players' beliefs may reflect their disagreement over various issues. But in equilibrium, all beliefs are assumed to conform to some commonly held view of reality. If σ is the equilibrium profile in question, each player i believes (with probability 1) that his opponents follow σ_{-i}, that each opponent $j \neq i$ believes that his opponents follow σ_{-j}, and so on. We denote this profile of beliefs by $\beta(\sigma) = (\beta_1(\sigma), \ldots, \beta_n(\sigma)) \in \bar{B}$.

[2] For each $j \in N$ and $k \geq 1$, let $Y_j^k := \times_{l=1}^k B_j^l$. Inductively define, for $\alpha = 0, 1, \ldots$, the sets

$X_j^k(\alpha) :=$ projection of $\hat{B}_j(\alpha)$ into Y_j^k, $j \in N$ and $k \geq 1$,

$X_{-i}^k(\alpha) := \times_{j \neq i} X_j^k(\alpha)$ (regard it as a subset of $\times_{l=1}^k B_{-i}^l$), and

$\hat{B}_i(\alpha + 1) := \{b_i \in \hat{B}_i(\alpha) |$ for every $k \geq 1$, $b_i^{k+1}(\Sigma_{-i} \times X_{-i}^k(\alpha)) = 1\}$.

Then $\bar{B}_i := \cap_{\alpha > 0} \hat{B}_i(\alpha)$ is the set of *collectively coherent* beliefs of player $i \in N$.

Definition. A *psychological Nash equilibrium* of a normal form psychological game G is a pair $(\hat{b}, \hat{\sigma}) \in \overline{B} \times \Sigma$ such that

(i) $\hat{b} = \beta(\hat{\sigma})$ and

(ii) for each $i \in N$ and $\sigma_i \in \Sigma_i$, $u_i(\hat{b}_i, (\sigma_i, \hat{\sigma}_{-i})) \leq u_i(\hat{b}_i, \hat{\sigma})$.

Examples

The Bravery Game

Player 1 must publicly take a decision, and is concerned with what his friends (player 2) will think about him. He can take a bold decision, which exposes him to the possibility of danger, or a timid, safe decision, so his action space is $A_1 = \{\text{bold, timid}\}$. His friends do not choose any action in the game, thus $A_2 = \{0\}$ (since their action space is a singleton, we omit mention of their strategy in what follows). Player 1 chooses bold with probability p and timid with probability $1 - p$. His payoff depends not only on what he does but also on what he thinks his friends think of his character (that is, on what he thinks they think he will do). In general their beliefs are given by a distribution on p, but to keep things simple, let us suppose that player 1 cares only about the mean $\tilde{q}$ of his beliefs about the mean of their beliefs. In particular, we suppose that player 1 would prefer to be timid rather than bold, unless he thinks that his friends expect him to be bold, in which case he prefers not to disappoint them.

Players 2 prefer to think of their friend as bold; in addition, it is good for them if he *is* bold. We let q represent their expectation of p (so $\tilde{q}$ is player 1's expectation of q). The game and payoffs are described in Fig. 2.

Note that unless $\tilde{q} \geq \frac{1}{2}$, player 1 prefers to be timid rather than bold. Players 2 always prefer that player 1 act boldly, especially if they expect him to. In equilibrium we must have $p = q = \tilde{q}$. It would be very nice for

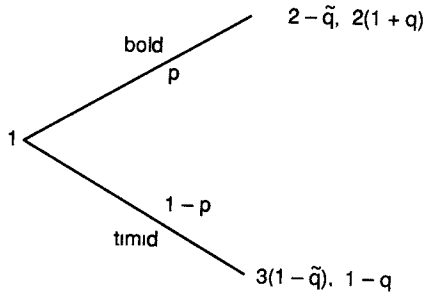

FIGURE 2

players 2 to believe that their friend is bold ($q = 1$), but if he is not, then in equilibrium they cannot hold these beliefs.

Equilibrium beliefs must correspond to equilibrium play, yet they can still exercise a decisive influence on what happens. In a traditional game with one active player there is either a unique equilibrium or a continuum of equilibria, all with the same payoff for the active player. Yet in the bravery game there are three equilibria.

In one equilibrium $p = q = \tilde{q} = 1$, and the payoffs are (1, 4). In a second equilibrium $p = q = \tilde{q} = 0$, and the payoffs are (3, 1). In the third equilibrium $p = q = \tilde{q} = \frac{1}{2}$ and payoffs are $(\frac{3}{2}, \frac{7}{4})$. Player 1 is best off when his friends expect little, but if their expectations are high he is trapped into meeting them, and then in a vicious cycle (for player 1) they are justified in holding such a lofty opinion of him.

A Confidence Game

Player 1 has invited a woman for a date, but he is not sure she will accept. He cannot tell whether she is player 2, who likes him, or player 3, who does not (nature chooses the woman's identity, with equal probabilities). Even if she likes him, it is not certain that she will accept. She will go out with him only if she thinks he is quite confident of himself, and that she measures by his probability assessment of being accepted. The payoff to player 1 depends not only on whether he is accepted or rejected but also on his expectations. If his invitation is accepted and he was pessimistic, he will be happy, but even more so when he expected it and was ready to be accepted. If he is rejected, he will be extremely depressed if he was optimistic, but his disappointment is moderated by initial pessimism.

The extensive form of the game and the payoffs are presented in Fig. 3. Note that q is 1's expectation of p, and s is 1's expectation of r. Moreover, $\tilde{q}$ represents player 2's expectation of player 1's expectation of her accepting him, and similarly $\tilde{s}$ is 2's expectation of 1's expectation of player 3 accepting the invitation. The first move is by nature (player 0). When

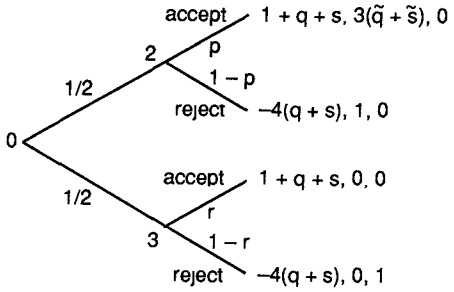

FIGURE 3

$q + s$ is higher, player 1 is happier to be accepted, but also more devastated by rejection. When $\tilde{q} + \tilde{s} > \frac{1}{3}$, player 2 prefers to accept 1's offer, but otherwise does not. Player 3 always prefers rejection.

We can easily calculate that there are again three equilibria; in the first, $p = q = \tilde{q} = 1$ and $r = s = \tilde{s} = 0$ and the payoffs are $(-1, \frac{3}{2}, \frac{1}{2})$ to the three players. In the second, $p = q = \tilde{q} = 0 = r = s = \tilde{s}$, and the payoffs are $(0, \frac{1}{2}, \frac{1}{2})$. In the last equlibrium, $p = q = \tilde{q} = \frac{1}{3}$, $r = s = \tilde{s} = 0$, and the payoffs are $(-\frac{8}{9}, \frac{1}{2}, \frac{1}{2})$. Player 1 has the best chance of an affirmative response in equilibrum 1, in which he is the most confident. But his disappointment is so great when he is rejected by player 3 that the most confident equilibrium is also the worst equilibrium for him.

Existence Theorem for Nash Equilibria of Normal Form Psychological Games

The preceding examples were relatively easily described, because each player cared about lower order beliefs in his belief hierarchy, and only about expectations. In general, a complete description would need to include payoffs associated with higher level beliefs, such as: "I, player 1, think that player 3 will choose 'heads', but I think that 2 thinks 3 will play 'tails', and I think 2 thinks I think 3 will play tails, ..." Fortunately, this kind of information is unnecessary for the purpose of *equilibrium* analysis. As noted before, any candidate equilibrium (b, σ) must involve belief hierarchies reflecting common knowledge of σ. We therefore define a "summary form" of a simultaneous psychological game, which contains just that information needed to compute the Nash equilibria of the game.

Consider a normal form psychological game $G = (A_1, \ldots, A_n; u_1, \ldots, u_n)$. *For* $\sigma, \tau \in \Sigma$ and $i \in N$, let $w_i(\sigma, \tau) := u_i(\beta_i(\sigma), \tau)$; $w_i : \Sigma \times \Sigma \to \mathbb{R}$ is called the *summary utility function* of player i, and $\hat{G} := (A_1, \ldots, A_n; w_1, \ldots, w_n)$ is the *summary form* of G.

LEMMA. *The function* $\beta_i : \Sigma \to \bar{B}_i$ *is continuous when* $\bar{B}_i$ *is given the product topology. Hence, if* $u_i : \bar{B}_i \times \Sigma \to \mathbb{R}$ *is continuous,* $w_i : \Sigma \times \Sigma \to \mathbb{R}$ *is continuous.*

The proof is straightforward and is omitted.

THEOREM. *Let* $G = (A_1, \ldots, A_n; u_1, \ldots, u_n)$ *be a normal form psychological game. Assume* $u_i : \bar{B}_i \times \Sigma \to \mathbb{R}$ *is continuous for each* $i \in N$. *Then G has a psychological Nash equilibrium.*

Proof. Let $BR_i : \Sigma \to \Sigma_i$ be player i's "best response" correspondence, defined for each $\sigma \in \Sigma$ by

$$BR_i(\sigma) := \{\hat{\tau}_i \in \Sigma_i \mid w_i(\sigma,(\hat{\tau}_i, \sigma_{-i})) \geq w_i(\sigma,(\tau_i, \sigma_{-i})) \quad \text{for all } \tau_i \in \Sigma_i\}.$$

Note that for each $b_i \in \bar{B}_i$, $\tau, \tau' \in \Sigma$, and $\lambda \in (0, 1)$,

$$u_i(b_i, \lambda\tau + (1 - \lambda)\tau') = \lambda u_i(b_i, \tau) + (1 - \lambda)u_i(b_i, \tau').$$

In particular this implies that $w_i(\sigma, \tau)$ is *concave* in τ_i. Therefore, the set $BR_i(\sigma)$ is convex. Since w_i is continuous, the Maximum theorem (see, for instance, Berge, 1963) implies that BR_i is upper semicontinuous and compact valued. Therefore, the correspondence $BR : \Sigma \to \Sigma$, defined by $BR(\sigma) := \times_{i \in N} BR_i(\sigma)$, $\sigma \in \Sigma$, admits a fixed point $\hat{\sigma}$. It is easy to verify that $(\beta(\hat{\sigma}), \hat{\sigma})$ is a Nash equilibrium of G. Q.E.D.

We conclude this section with a discussion of the continuity assumption in the existence theorem. Evidently, Theorem 1 is true if w_i is continuous, for which the continuity of u_i when B_i is given the product topology is sufficient but not necessary. As we show with an example, the continuity of u_i when B_i is endowed with the product topology is not completely innocuous. Consider the following situation: a man is deciding whether or not to give a woman flowers. The psychological phenomenon we wish to capture is the following. The woman may become unhappy for either of two reasons: she might expect flowers and not receive them, or she might conclude from his behavior that he is willing to disappoint her. Thus, even if she is not expecting flowers but believes that he thinks she is expecting flowers, she will be unhappy not to receive flowers, because this indicates his willingness to disappoint her. To carry the story one step further, if her fifth-order beliefs are that his fourth-order beliefs are that her third-order beliefs are those described in the previous sentence, then, regardless of her third-order beliefs, she would be unhappy not to receive flowers. As this reasoning can be continued indefinitely, one sees that unhappiness at not receiving flowers may be triggered by a belief of arbitrarily high order. Formally, we define her utility function as follows. Let p be the probability that he chooses to take her flowers, and q_1 be her expectation of p. Inductively define, for $k \geq 1$,

q_{2k} = his expectation of q_{2k-1} and

q_{2k+1} = her expectation of q_{2k}.

Her "level of disappointment" is a function of her hierarchy of beliefs, defined by $l(b_2) := \max\{q_{2k+1} | k \geq 0\}$. Her utility function is then specified by $\bar{u}_2(b_2, \sigma) := \sigma + (1 - \sigma)(1 - l(b_2))$, for each $b_2 \in \bar{B}_2$, where $\sigma = 1$ represents the outcome "he takes her flowers," and $\sigma = 0$ the outcome "he doesn't take her flowers." The reader may check that $u_2 : \bar{B}_2 \times \Sigma_1 \to \mathbb{R}$ is *not* continuous.

3. Extensive Psychological Games

This section restricts attention to a rather simple class of extensive psychological games. In principle, a player's utility might depend not only on his beliefs at the beginning of the game but also on his beliefs as play progresses (about others' beliefs at various junctures, as well as their strategies). Full-blown extensive belief hierarchies of this kind are not well understood and raise some difficult questions concerning coherency restrictions and equilibrium analysis. In particular it is not appropriate to suppose that all beliefs at every information set are generated by the hypothesis that the strategy profile σ is common knowledge. We plan to address these issues in subsequent work. Here we avoid them by considering utility functions that depend upon "reduced" belief hierarchies. Specifically, a player's utility depends only on the strategy profile played, his *initial* beliefs about what will be played, his initial beliefs about others' *initial* beliefs about what will be played, and so on. Thus, the belief hierarchies studied here have the same structure as those of Section 2.

We begin by developing the necessary notation for the extensive form, and then proceed to questions of perfection and sequential rationality.

The Game Form

A *game form* $F = (N, V, <, m, \rho, \Pi, A)$ consists of a set of players $N = \{1, \ldots, n\}$, a finite set of vertices V with a partial order $<$, a function m specifying for each nonterminal node which player is on the move, a system of probability distributions ρ specifying the moves of nature, an information partition Π of the nonterminal nodes, and an action correspondence A mapping each nonterminal node into a set of actions.

The partial order $<$ gives V the structure of a *tree* with root $v_0 \in V$ (i.e., $v_0 \leq v$ for all $v \in V$, where $x \leq y$ means $x < y$ or $x = y$) and terminal nodes T ($t \in T$ iff there is no $v \in V$ such that $t < v$). For each vertex v there is a unique path $v\downarrow := \{x \in V | x \leq v\}$ leading from v_0 to v. The set of immediate successors of a nonterminal node v is denoted by $s(v)$. For each nonterminal node v, $A(v)$ is the set of actions available at v. There is a one-to-one map between $s(v)$ and $A(v)$: for any $x \in s(v)$ there is a unique $a \in A(v)$ that leads from v to x.

The move function $m : V/T \to \{0, 1, \ldots, n\}$ specifies for each nonterminal vertex v what player chooses an action at v. Player 0 represents "nature": if $m(v) = 0$, then the action at v is chosen randomly according with the probability distribution $\rho(v)$. The system of probability distributions ρ is common knowledge among the players N.

If v is a nonterminal node and $m(v) = i \neq 0$, then $\Pi(v)$ represents the set of all vertices that player i cannot distinguish from v. If $m(v) = 0$, then

$\Pi(v) = \{v\}$. A player always knows when he is on the move and what actions are available. Therefore, we assume:

(I1) if $m(x) \neq 0$ and $y \in \Pi(x)$, then $m(y) = m(x)$ and $A(y) = A(x)$.

Let Π^i, $i = 0, \ldots, N$, denote the information sets where player i (or nature) is on the move: $\Pi^i := \{\Pi(x) | m(x) = i\}$. (I1) implies that $A(v)$ is constant on each $h \in \Pi^i$. Hence, for every $h \in \Pi^i$ and *any* $v \in h$, one can define $A_i(h) := A(v)$. Finally, we assume:

(I2) each player $i \in N$ has perfect recall (see Kuhn, 1953).

Initial Beliefs

DEFINITION. A *behavior strategy* for player i associates with every $h \in \Pi^i$ a probability distribution $\sigma_i(h)$ over $A_i(h)$. Let $\Sigma_i(h) := \Delta(A_i(h))$ and Σ_i denote the set of player i's behavior strategies, and define $\Sigma := \times_{i \in N} \Sigma_i$ and $\Sigma_{-i} := \times_{j \neq i} \Sigma_j$.

The set of *initial beliefs* for an extensive game form have the same structure as the sets of beliefs in normal form games. We retain the relevant notation from Section 2, along with the notions of coherency and collective coherency defined there. Thus, $\overline{B}_i$ denotes the set of player i's collectively coherent beliefs and $\overline{B} := \times_{i \in N} \overline{B}_i$.

Utility Functions

Each strategy profile $\sigma \in \Sigma$ (together with the system of probability distributions ρ) induces a probability distribution P_σ over the terminal nodes T. Player i's *utility function* $\bar{u}_i : \overline{B}_i \times T \to \mathbb{R}$ depends on his intial beliefs and on the outcome reached. We shall assume that player i only cares about the expected value of $\bar{u}_i$, and extend his utility function to $u_i : \overline{B}_i \times \Sigma \to \mathbb{R}$ by

$$u_i(b_i, \sigma) := \Sigma_{t \in T} P_\sigma(t) \bar{u}_i(b_i, t).$$

DEFINITION. An *extensive psychological game* $\Gamma := (F, (u_i)_{i \in N})$ consists of a game form F together with a utility function u_i for each player.

Subgame Perfection

Once we have identified the strategy spaces Σ_i, the hierarchies of initial beliefs $\overline{B}_i$, and the payoffs u_i, a psychological game in extensive form can also be thought of as a psychological game in normal form. We can accordingly define a Nash equilibrium of $\Gamma = (F, (u_i)_{i \in N})$ to be any pair $(\hat{b}, \hat{\sigma}) \in \overline{B} \times \Sigma$ such that

$$u_i(\hat{b}_i, (\sigma_i, \hat{\sigma}_{-i})) \leq u_i(\hat{b}_i, \hat{\sigma}) \qquad \text{for all } \sigma_i \in \Sigma_i,$$

and $\hat{b}_i = \beta_i(\hat{\sigma})$ for all $i \in N$, just as we did for psychological games in normal form. Recall that the map $\beta_i : \Sigma \to \overline{B}_i$ is the embedding of Σ in $\overline{B}_i$ according to which σ is "common knowledge."

But the extensive structure of F is essential for the natural extension of subgame perfection to psychological games. For any $b \in \times_{i \in N} \overline{B}_i$, let $\Gamma(b) = (F, (u_i(b_i, \cdot))_{i \in N})$ be the standard (nonpsychological) extensive form game whose payoffs are computed according to $u_i(b_i, \cdot)$, $i \in N$.

DEFINITION. The pair $(\hat{b}, \hat{\sigma}) \in \overline{B} \times \Sigma$ is a *subgame perfect* (respectively, trembling hand perfect) *psychological equilibrium* of $\Gamma = (F, (u_i)_{i \in N})$ if it is a psychological Nash equilibrium of Γ and $\hat{\sigma}$ is a subgame perfect (respectively, trembling hand perfect) equilibrium of $\Gamma(\hat{b})$, in the traditional sense.

In the standard game of perfect information shown in Fig. 4, the unique subgame perfect equilibrium can be found by backward induction: if reached, players 2 and 3, respectively, have every incentive to play $q = 1$ and $r = 1$. Knowing this, player 1 plays $p = 0$. The game summarized in Fig. 5 adds a psychological component to the payoffs of players 2 and 3: the satisfaction that each derives from hurting player 1 is proportional to the amount by which 1's action is perceived to have lowered the (standard) payoff of the player in question. The number $\tilde{q}$ represents player 3's expectation of the probability that 2 plays down. The number $\tilde{r}$ represents player 2's expectation of the probability that 3 would play down if reached. For example, if 1 chooses up and player 2 thinks that had down been chosen, player 3 would likely have played down, then 2 revels in hurting player 1 and chooses up. Note that here it is *not* possible to start at the end of the tree and work backward. What 2 wants to do depends on how much he thinks he lost by 1's failure to choose down; that loss

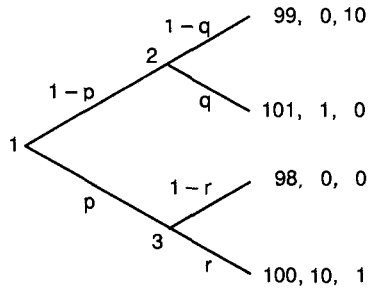

FIGURE 4

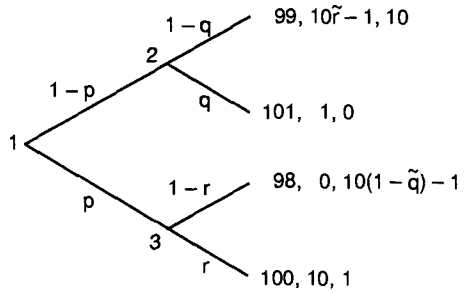

FIGURE 5

depends on what 3 would have done if reached. But 3's desired choice depends in turn on what he thinks he would have received in the upper half of the tree, and this is determined by 2's expected behavior. It is easy to check that there is no equilibrium in pure strategies. There is, however, a unique subgame perfect psychological equilibrium in mixed strategies: 2 and 3 each randomize in such a way as to make the other indifferent between his two pure strategies. In the solution, $\tilde{q} = q = \frac{4}{5}$, $\tilde{r} = r = \frac{1}{5}$, and player 1 chooses $p = 0$.

The next example shows that in psychological games, trembling hand perfect equilibria (Selten, 1975) need not exist. In Fig. 6, $\tilde{p}$ is player 1's expectation of player 2's expectation of the probability that 1 chooses up. In equilibrium $\tilde{p} = p$. All candidates for a trembling hand perfect psychological equilibrium involve 2 playing up with certainty ($q = 1$). But in the only psychological equilibrium in which 2 plays $q = 1$, player 1 chooses $p = 0$. This induces the standard game shown in Fig. 7. However, $(\alpha_2; \beta_1)$ is *not* trembling hand perfect in the induced game; the only such profile is $(\alpha_1; \beta_1)$.

This raises the question of whether subgame perfect equilibria and sequential equilibria always exist in psychological games. Fortunately,

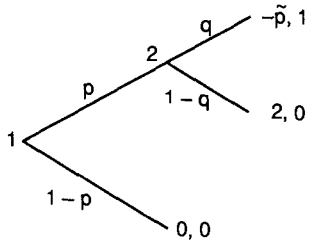

FIGURE 6

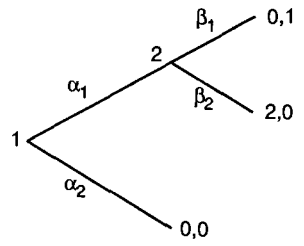

FIGURE 7

we can answer in the affirmative. For the next definition, recall from Kreps and Wilson (1982) that a belief system μ associates with each information set a probability distribution over the nodes in that set. Let M be the set of belief systems for an extensive psychological game Γ.

DEFINITION. The triple $(\hat{b}, \mu, \hat{\sigma}) \in \overline{B} \times M \times \Sigma$ is a *sequential psychological equilibrium* of $\Gamma = (F, (u_i)_{i \in N})$ if $(\hat{b}, \hat{\sigma})$ is a psychological Nash equilibrium of Γ and $(\mu, \hat{\sigma})$ is a sequential equilibrium of $\Gamma(\hat{b})$, in the traditional sense.

THEOREM. *Let $\Gamma = (F, (u_i)_{i \in N})$ be an extensive psychological game. Assume $u_i : \overline{B}_i \times \Sigma \to \mathbb{R}$ is continuous for each $i \in N$. Then Γ has a subgame perfect psychological equilibrium. Indeed, Γ has a sequential psychological equilibrium.*

Proof. Let $\varepsilon > 0$. For each $i \in N$ and $h \in \Pi^i$, let

$$\Sigma_i^\varepsilon(h) := \{\tau \in \Sigma_i(h) | \tau(a) \geq \varepsilon \text{ for every } a \in A_i(h)\}.$$

Γ^ε denotes the perturbed extensive psychological game obtained from Γ by restricting player i's strategies ($i \in N$) to the set

$$\Sigma_i^\varepsilon := \{\sigma_i \in \Sigma_i | \sigma_i(h) \in \Sigma_i^\varepsilon(h) \quad \text{for each } h \in \Pi^i\}.$$

Let $\Sigma^\varepsilon := \times_{i \in N} \Sigma_i^\varepsilon$. If $i \in N$, $\sigma_i \in \Sigma_i$, $h \in \Pi^i$, and $\tau_i \in \Sigma_i(h)$, then σ_i/τ_i will denote the behavior strategy for player i specifying the mixture τ_i at the information set h, and $\sigma_i(h')$ at every other information set $h' \in \Pi^i$. For $\sigma \in \Sigma$, $\varepsilon > 0$, $i \in N$, and $h \in \Pi^i$, define

$$BR_h^\varepsilon(\sigma) := \{\hat{\tau}_i \in \Sigma_i^\varepsilon(h) | w_i(\sigma, (\sigma_i/\hat{\tau}_i, \sigma_{-1})) \geq w_i(\sigma, (\sigma_i/\tau_i, \sigma_{-i})) \quad \text{for all } \tau_i \in \Sigma_i^\varepsilon(h)\}.$$

The correspondence $BR_h^\varepsilon : \Sigma \to \Sigma_i^\varepsilon(h)$ is u.s.c., and compact and convex valued. Therefore, the correspondence $BR^\varepsilon : \Sigma^\varepsilon \to \Sigma^\varepsilon$ (defined by the coordinate correspondences BR_h^ε, $h \in \Pi/\Pi^0$) admits a fixed point $\hat{\sigma}^\varepsilon$. The strategy profile $\hat{\sigma}^\varepsilon$ is a Nash equilibrium of $\Gamma^\varepsilon(\beta(\hat{\sigma}^\varepsilon))$.

Let $\{\varepsilon_r\}$ be a sequence of positive numbers converging to 0, and $\{\hat{\sigma}^r\}$ be a sequence of strategy profiles such that $\hat{\sigma}^r$ is a Nash equilibrium of $\Gamma^{\varepsilon_r}(\beta(\hat{\sigma}^r))$ for each $r \in \mathbb{N}$. Since Σ is compact, $\{\hat{\sigma}^r\}$ has an accumulation point $\hat{\sigma}$. Since β is continuous, it is not difficult to see that $(\beta(\hat{\sigma}), \hat{\sigma})$ is a Nash equilibrium of Γ (and that $\hat{\sigma}$ is a Nash equilibrium of $\Gamma(\beta(\hat{\sigma}))$). Moreover, for all $r \geq 0$, $i \in N$, $h \in \Pi^i$, and $\tau_i \in \Sigma_i^{\varepsilon_r}(h)$, we also have

$$\Sigma_{t \in T} P^h_{\hat{\sigma}^r}(t)\bar{u}_i(\beta_i(\hat{\sigma}^r), t) \geq \Sigma_{t \in T} P^h_{(\hat{\sigma}_i^r/\tau_i, \hat{\sigma}^r_{-i})}(t)\bar{u}_i(\beta_i(\hat{\sigma}^r), t),$$

where for any $\gamma \in \Sigma$ we mean by $P^h_\gamma(t)$ the probability that terminal node t is reached if play proceeds according to γ, conditional on h being reached. Observe that by passing to subsequences we may suppose that for all $h \in \Pi^i$ and $\tau_i \in \Sigma_i(h)$, $P^h_{\hat{\sigma}^r}$ and $P^h_{(\hat{\sigma}_i^r/\tau_i, \hat{\sigma}^r_{-i})}$ converge to what we shall call $\tilde{P}^h_{\hat{\sigma}}$ and $\tilde{P}^h_{(\hat{\sigma}_i/\tau_i\ \hat{\sigma}_{-i})}$, respectively. (If $\hat{\sigma}$ reaches h with positive probability, then $\tilde{P}^h_{\hat{\sigma}} = P^h_{\hat{\sigma}}$ and $\tilde{P}^h_{(\hat{\sigma}_i/\tau_i, \hat{\sigma}_{-i})} = P^h_{(\hat{\sigma}_i/\tau_i, \hat{\sigma}_{-i})}$.) Since u_i and β_i are continuous, in the limit

$$\Sigma_{t \in T} \tilde{P}^h_{\hat{\sigma}}(t)\bar{u}_i(\beta_i(\hat{\sigma}), t) \geq \Sigma_{t \in T} \tilde{P}^h_{(\hat{\sigma}_i/\tau_i, \hat{\sigma}_{-1})}(t)\bar{u}_1(\beta_i(\hat{\sigma}), t).$$

Hence $(\beta(\hat{\sigma}), \hat{\sigma})$ is a subgame perfect psychological equilibrium of Γ. For each $i \in N$ and $h \in \Pi^i$, one can immediately compute conditional probabilities $\mu(h)$ to verify that $(\beta(\hat{\sigma}), \mu, \hat{\sigma})$ is a sequential equilibrium of the psychological game Γ. Q.E.D.

We now present an example concerning the evolution of sympathy between players. Its purpose is twofold. First, it demonstrates the reasoning involved in analyzing a psychological game with substantial dynamic structure. Second, it illustrates the structural features of equilibria of games in which a player builds sympathy not by being generous, but by being more generous than expected. Such considerations are frequently significant in actual strategic situations, and their implications, as we show, can be somewhat counterintuitive.

EXAMPLE. *A dynamic prisoner's dilemma game.* Consider a finite horizon dynamic prisoner's dilemma game in which player 2 becomes sympathetic to player 1 if 1's play in early periods is unexpectedly cooperative. Specifically, player 2's payoffs in every period are modified by a "sympathy factor" α, hereafter called the *stock*. The stock α is increased (or decreased) at the end of every period by an amount proportional to the

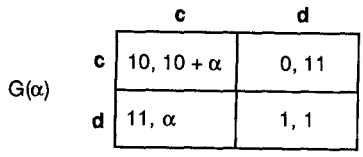

FIGURE 8

difference between player 2's expectation of 1's choice and player 1's actual choice. Suppose the current stock is α. Then the payoffs for the present period are given by the matrix in Fig. 8.

That is, player 2's payoff from cooperation is augmented by α, the extent to which he is sympathetic to 1. If 2 expects 1 to choose c with probability p, then next period the stock will be $\alpha + k(1 - p)$ if 1 chooses c, and $\alpha - kp$ if 1 chooses d, where the "sympathy coefficient" $k \geq 0$ is a given parameter. Thus, the stock α is increased whenever 2 is pleasantly surprised by 1's action, and decreased otherwise.

Denote by $G_k^T(\alpha)$ the T-period game with initial stock α and sympathy coefficient k. Periods in the game $G_k^T(\alpha)$ are numbered in decreasing order, so, for example, period T is chronologically the first period of $G_k^T(\alpha)$, and period 1 is the last.

Recall that in the traditional finitely repeated prisoner's dilemma game $G_0^T(0)$, the only equilibrium path involves the outcome (d, d) in every period, and the only equilibrium payoff is (T, T). We are interested in whether 1's ability to manipulate 2's emotions in the psychological game puts 1 at a relative advantage, and whether his payoff exceeds that in the traditional prisoner's dilemma game.

For concreteness, we first discuss the specific game $G_2^3(0)$, that is, the three-period game in which the initial stock of sympathy is 0, and the coefficient of sympathy k is 2. Here the term "equilibrium" will mean "subgame perfect psychological equilibrium." The game has no equilibrium in pure strategies. If player 1 were expected to play d in period 3, he could increase α to 2 by surprising 2 with the choice c. It is easy to check that all equilibrium payoffs of $G_2^2(2)$ are so much better for 1 than his equilibrium payoff in $G_2^2(0)$ (the subgame that would result if 1 played d in period 3) that he would want to play c in period 3, a contradiction. On the other hand, if 1 were expected to play c, doing so would leave the stock at 0. But in the two-period continuation game, a stock of 0 turns out to be no better for 1 than a negative stock, so he would defect to d in period 3. Thus, in any equilibrium, 1 must randomize in period 3.

Of the two equilibria of $G_2^3(0)$, the simpler is as follows. In period 3,

player 2 plays d and 1 chooses c with probability $\frac{1}{2}$. Consequently, the stock in period 2 is either -1 or 1. In the former case, no further cooperation occurs on either side, and each player receives a payoff of 2 in the two-period subgame (not counting the payoff from period 3). In order that 1 be willing to randomize in period 3, his payoff in the subgame reached if he cooperates must be 3 (balancing the one unit cost of cooperation in period 3). The equilibrium achieving this in the subgame involves 1 playing c in period 2 (leaving him no opportunity to raise α above the critical level of 1) and d in period 1, while 2 plays d in period 2 and cooperates with probability $\frac{2}{10}$ in period 1. Player 2 is willing to play d in period 2 and to randomize in the final period because when $\alpha = 1$, he is indifferent. Whether 1 plays c or d in period 3, his supergame payoff is 3, exactly what it would be in a standard prisoner's dilemma game with an unsympathetic opponent. Ironically, player 2's expected payoff is 13. In this case, it is better to be manipulated than to be the manipulator.

In the second equilibrium, player 1 again cooperates with probability $\frac{1}{2}$ in period 3, but player 2 cooperates with certainty. Player 2's incentives for doing so come from the (subgame perfect) "threat" that if the stock in the two-period subgame is 1 and he has cooperated, the resulting equilibrium will be the same as that described in the preceding paragraph. If he has *not* cooperated, a much less attractive equilibrium for him results: in both remaining periods, player 1 chooses d and player 2 chooses c. In the event that 1's choice in period 3 is d, there is no further cooperation, regardless of 2's play in period 3. The payoff pair in this second equilibrium is (13, 12).

Note that the first of the two equilibria of $G_2^3(0)$ could be called a state-space equilibrium, because behavior depends only on the period and the current stock. The second could not: although player 2's action in period 3 cannot influence the state in period 2, his period 3 choice does affect the equilibrium in the subgame $G_2^2(1)$. In the games $G_2^T(0)$ with $T > 3$, there continue to be multiple equilibria because a variety of behaviors on 2's part can be enforced by credible threats as before. But no matter how long the game, there is always a state-space equilibrium in which player 1 receives expected payoff T (as though he were in a standard prisoner's dilemma game). In this equilibrium, he cooperates with probability $\frac{1}{2}$ in period T while 2 plays d. If the stock in period $T - 1$ is -1, there is no further cooperation on the equilibrium path. If instead the stock is 1, player 1 cooperates in all subsequent periods except the last. Player 2 cooperates toward the end of the game. For example, if $T = 25$, in the subgame $G_2^{24}(1)$ he plays d in periods 24 to 4, inclusive, and then cooperates with probability $\frac{4}{10}$ in period 3 and with certainty in periods 2 and 1. (There is a variety of other endgame behaviors by 2 that are payoff-equivalent and satisfy the relevant incentive constraints.) This amount of

cooperation is just enough to compensate player 1 for having cooperated 24 times.

Solutions for other positive values of the sympathy coefficient k are similar. The state-space equilibrium for $k \in [0, 1)$ exhibits no cooperation on the equilibrium path.

4. Conclusion

Emotional reactions often depend on expectations. An event might provoke in someone feelings of disappointment or relief, gratitude or anger, or pride or embarrassment depending on what the individual expected ex ante, or thought others expected, and so on. Psychological games provide a framework for the formal analysis of strategic settings in which expectations and emotions play a role. Our work is complementary to that of Gilboa and Schmeidler (1988) whose information-dependent games are a closely related alternative means of extending traditional game theory. We formulate the natural analogs of Nash equilibrium and several of its refinements in psychological games, and show by a series of examples that they exhibit novel properties. These include multiplicity of subgame perfect equilibria in some perfect information games (without ties in payoffs) and the necessity for randomization in certain games with only one active player.

Equilibrium analysis in psychological games suffers from an extra source of simultaneity: payoffs associated with a given strategy profile are generated endogenously and differ across equilibria of the game. Consequently, backward induction no longer affords a proof of the existence of subgame perfect equilibria. Indeed it turns out that the simplest analog of trembling hand perfect equilibrium need not exist in psychological games. On the other hand, we prove the existence of subgame perfect and sequential psychological equilibria. Thus, to return to the discussion that opened the Introduction, there is no fundamental tension between the presence of emotional factors in decision-making and the imposition of sequential rationality.

This paper restricts attention to a relatively simple class of extensive psychological games in which only initial beliefs enter the utility function. More generally, a player's payoff might depend, for example, upon his opinion at some information set that he did not expect to reach, about what some other player will conclude subsequently about the progress of play. Extending the domains of the utility functions raises complicated questions concerning the appropriate formulation of the extensive belief hierarchies, the degree of common knowledge that should be imposed by

an equilibrium theory on agents off the equilibrium path, and a number of related issues. We hope to study these problems in future work.

ACKNOWLEDGMENTS

We are grateful to Faruk Gul for important suggestions. We also thank an anonymous referee for his comments, and the National Science Foundation for its financial support.

REFERENCES

BERG, C. (1963) *Topological Spaces*. New York, NY: Macmillan.

BRANDENBURGER, A , AND DEKEL, E. (1985) "Hierarchies of Beliefs and Common Knowledge," mimeo, Harvard University

GILBOA, I , AND SCHMEIDLER, D. (1988). "Information Dependent Games: Can Common Sense be Common Knowledge?" *Econ. Lett.* **27,** 215–221

HARSANYI, J. (1967–1968) "Games with Incomplete Information Played by Bayesian Players," Parts I, II, and III, *Manage Sci.* **14,** 159–182, 320–334, 486–502

KREPS, D., AND WILSON, R. (1982). "Sequential Equilibria," *Econometrica* **50,** 863–894.

KUHN, H. (1953). "Extensive Games and the Problem of Information," in *Contributions to the Theory of Games* (H. Kuhn and A. Tucker, Eds), Vol. 2. Princeton, NJ. Princeton Univ. Press

MERTENS, J.-F , AND ZAMIR, S. (1985). "Formulation of Bayesian Analysis for Games with Incomplete Information," *Int. J. Game Theory* **14,** 1–29.

NALEBUFF, B., AND SHUBIK, M. (1988). "Revenge and Rational Play," Discussion Paper No. 138, Woodrow Wilson School

SELTEN, R. (1965). "Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragetragheit," *Staatswiss.* **121,** 301–324

SELTEN, R. (1975). "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *Int J. Game Theory* **4,** 25–55.