Equilibrium in Justifiable Strategies: A Model of Reason-Based Choice in Extensive-Form Games

Ran Spiegler

The Review of Economic Studies, Vol. 69, No. 3. (Jul., 2002), pp. 691-706.

Stable URL:
http://links.jstor.org/sici?sici=0034-6527%28200207%2969%3A3%3C691%3AEIJSAM%3E2.0.CO%3B2-8

The Review of Economic Studies is currently published by The Review of Economic Studies Ltd..

Review of Economic Studies (2002) **69**, 691–706
0034-6527/02/00270691$02.00

Equilibrium in Justifiable Strategies: A Model of Reason-based Choice in Extensive-form Games

RAN SPIEGLER
Tel Aviv University

First version received January 2000; *final version accepted November* 2001 (*Eds.*)

I explore the idea that people care about the justifiability of their decisions in the context of two-person extensive games. Each player justifies his strategy s with a belief b of the opponent's strategy, which is consistent with the play path and maximally plausible (according to some exogenous criterion). We say that s is justifiable if against the *ex post* criticism that some other strategy s' outperforms s against b, the player can argue that playing s' would have exposed him to similar criticism in the opposite direction. Under a simplicity-based plausibility criterion, this concept implies systematic departures from maximizing behaviour in familiar games.

1. INTRODUCTION

While economic theory equates rationality with utility maximization, our everyday thinking accommodates additional notions of rational behaviour. One example is "*reason-based choice*". According to this notion, rational decisions can be defended by good reasons. Psychologists have argued that certain aspects of economic behaviour that are hard to reconcile with utility maximization (*e.g.* preference reversals) may be consequences of reason-based choice procedures (*e.g.* Shafir *et al.* (1993), Tversky and Shafir (1992)). The problem is that in contrast to maximizing behaviour, the notion of reason-based choice has not been subjected to formalization. Consequently, it is hard to assess its relevance to economic modelling.

This paper attempts to formalize reason-based choice procedures and embed them in economic models. As a first step in this direction, I find it instructive to examine environments, in which reason-based choice is particularly relevant—*i.e.* where agents need to *justify* their decisions *ex post*. Justifiability is an accountability procedure, which takes into account the *reasons* behind the agent's decisions, in addition to their consequences. Justifiability considerations are abundant in organizational and political decision-making, where choices have to be communicated to principals, electorates and other audiences. (See Tetlock and Boettger (1989, 1994) for experimental evidence.)

I study *ex post* justification procedures in the context of complete information, two-person extensive-form games. After a strategy profile (s_1, s_2) is realized, each player has to justify his strategy to a "critic", who can be interpreted as a principal. (Each player faces his own critic). I view the justification procedure as a debate, in which the player and critic exchange arguments and counter-arguments. (This debate is *not* modelled as a game.)

Player 1, say, justifies his strategy by an *ex post* belief $b \in S_2$ of the opponent's pure strategy. This belief need not be correct; player 1's critic knows s_1 and the play path induced by (s_1, s_2), denoted $h(s_1, s_2)$, but he does not necessarily know s_2. However, b cannot be arbitrary; it must satisfy two conditions: (1) *consistency* with $h(s_1, s_2)$; (2) *maximal plausibility*, according to some exogenous criterion. Formally, the plausibility criterion is a weak order R on S_2; b must

satisfy bRb', for every other belief b' that is consistent with $h(s_1, s_2)$. A belief that meets these conditions is called a *most plausible consistent belief* (MPCB), given $h(s_1, s_2)$.

Since a play path in an extensive game is usually consistent with multiple strategy profiles, s_2 cannot be pinned down on the basis of s_1 and $h(s_1, s_2)$. This is the source of the players' justifiability problem: optimizing against a consistent belief does not ensure justifiability because the belief may fail to satisfy maximal plausibility.

If s_1 is a best reply to some MPCB given $h(s_1, s_2)$, it is obviously justifiable, for neither the player's strategy nor the belief are objectionable. However, as we shall see, it may more often be the case that given s_2, there exist no $s_1 \in S_1$ and $b \in S_2$, such that b is an MPCB given $h(s_1, s_2)$ and s_1 is a best-reply to b. That is, whatever player 1 does, either the belief he uses to justify his strategy is inadmissible, or his strategy is sub-optimal. Is there a sense, in which player 1 can behave justifiably under such circumstances? In other words, can player 1 survive his critic's *ex post* criticism that s_1 failed to optimize against an MPCB $b \in S_2$ given $h(s_1, s_2)$?

The "solution" that I propose for this problem is motivated by the debate-like form of the justification procedure. In debates, an argument can be rebutted by a "smashing" counter-argument. Similarly in our model, player 1's strategy s_1 is **justifiable**, given $h(s_1, s_2)$, if there exists an MPCB $b \in S_2$ given $h(s_1, s_2)$, such that for every $s_1' \in S_1$ for which $u(s_1', b) > u(s_1, b)$, it is the case that $u(s_1, b') > u(s_1', b')$ for any MPCB $b' \in S_2$, given $h(s_1', b)$. This definition captures the following exchange of arguments between player 1 and his critic in their post-game debate:

Player 1's critic (argument). "s_1' would have outperformed s_1 against b, the belief that you [*i.e.* player 1] are using to justify s_1".

Player 1 (counter-argument). "But if I had played s_1', I could have been criticized *on the same grounds*, for having failed to play s_1, the very strategy that you [*i.e.* the critic] are now criticizing."

If player 1 can counter-argue in this way against the critic's argument, he wins the debate and his strategy s_1 is justifiable, given $h(s_1, s_2)$. This is the notion of "reason-based choice" that is captured in this paper: having a good reason for one's decision means that one is armed with "smashing" counter-arguments against *ex post* criticisms.

This is a powerful way of silencing criticism. It accepts the plausibility criterion R that underlies the critic's argument against s_1; it shows that following the critic's *ex post* recommendation would have exposed the player to a similar criticism, based on the same plausibility criterion, which would have recommended s_1 itself. Of course, this is not the only way to handle the critic's argument. For instance, the player could claim that R is an inappropriate criterion to begin with. However, the above counter-argument is especially effective because it is based on the rhetorical device of *erecting the counter-argument on the argument's premise* (see Walton (1989)).

To illustrate the justifiability concept, consider the centipede game represented by Figure 1. Suppose that the strategy "always C" is considered more plausible according to R than any strategy of the form "stop at the kth node" (denoted by $S(k)$), $k > 2$. This is a simplicity criterion: if a player takes the same action along the play path, the simplest belief is that he would have continued to do so off-path.[1]

Consider player 2's justification procedure, given that his strategy is $S(8)$ and the realized play path is marked by the bold crooked arrow.

1. Other plausibility criteria (*e.g.* rationality) may imply a different ranking, but this is irrelevant because justifiability is defined for a *given* criterion.

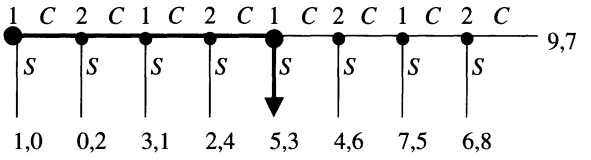


FIGURE 1

If optimizing against an MPCB were the only way to attain justifiability, then player 2's strategy would not be justifiable—it is only a *second*-best reply to $S(5)$, which is the unique MPCB of player 1's strategy given the play path. (The first-best is S(4).) By our concept, player 2's strategy is justifiable because he can counter-argue: "If I had deviated from $S(8)$ to $S(4)$, 'always C' would have been the unique MPCB of player 1's strategy given the path induced by my deviation, and I would have been criticized for having done worse than $S(8)$ itself against 'always C'".

A profile of justifiable strategies (given their induced path and the plausibility criterion R) constitutes *"equilibrium in justifiable strategies"* (EJS). EJS can be viewed as an extension of the concept of self-confirming equilibrium due to Fudenberg and Levine (1993). When R is degenerate (such that all strategies are equally plausible), the two concepts are equivalent.

The implications of EJS depend on the structure of R. In this paper, I employ a *simplicity-based* plausibility criterion: sRs' if and only if s' is not simpler than s'. Simplicity of strategies is defined in terms of their finite automata representation: sRs' if s has no more states and no more transitions than s'. Needless to say, this definition is merely an example. However, I hope to persuade the reader that it captures considerations that are natural and pertinent to extensive games. Alternative plausibility criteria are discussed in Section 5.5.

I apply EJS with the simplicity-based plausibility criterion to familiar games: the centipede and chain store games, as well as a two-person variant on Samuelson's inter-temporal exchange model (Samuelson, 1958). The departure from maximizing behaviour inherent in our notion of justifiability leads to non-standard play patterns. The purpose of the applications is *not* to resolve the paradoxes associated with these games, but to illustrate EJS and the various considerations it captures in these games. Applying EJS to bargaining environments remains a particularly interesting challenge for future research because bargaining is often a delegated activity, in which justifiability considerations may have non-trivial effects on delay.

2. THE EQUILIBRIUM CONCEPT

I study two-person extensive-form games with complete information.[2] Let S_i be player i's strategy space. Let $h(s_1, s_2)$ be the play path induced by the strategy pair (s_1, s_2). Let $b_j \in S_j$ denote a belief of player j's pure strategy. A pair consisting of player i's strategy and a belief of player j's strategy is denoted (s_i, b_j). Player i's payoff function is denoted u_i.

The model describes a post-game justification procedure for each player. The procedure can be viewed as a debate between each player and his "critic", who can be interpreted as the player's principal. However, the critic is not a player and the debate is not modelled as a non-cooperative game. The critic reviews the consequences of the player's strategy, as well as the plausibility of the belief that he uses to justify it.

2. All the definitions are extendible to incomplete-information games.

Players must defend their choice of strategy with beliefs that are *consistent* with the realized path—*i.e.* $h(s_i, b_j) = h(s_1, s_2)$ for every $i \in \{1, 2\}$. *Plausibility* of player i's belief is determined by some exogenous criterion, which induces a weak order R on S_j. For every path h, we can define the set of MPCB of player j's strategy, $B_j{}^*(s_1, s_2) = \{b \in S_j; b$ is consistent with $h(s_1, s_2)$ and bRb' for every other belief b' that is consistent with $h(s_1, s_2)\}$. Player i must justify s_i with some $b \in B_j{}^*(s_1, s_2)$.

When will we say that a player's strategy is justifiable with respect to the observed play path and the plausibility criterion R? It seems obvious that if s_i is a best reply to some $b \in B_j^*(s_1, s_2)$, it is justifiable because it optimizes against an admissible belief. However, since $B_j{}^*(\cdot)$ can vary with the play path, it may be the case that any best-reply s to the belief b implies $b \notin B_j^*(s, b)$. In fact, this will be the *prevalent* case under the plausibility criterion that will serve us in the applications section. Therefore, the question arises: can a strategy be sub-optimal and justifiable at the same time?

When player i uses $b \in B_j{}^*(s_1, s_2)$ to justify s_i and s_i is not a best-reply to b, player i's critic can argue *ex post* that player i should have played some s_i', for which $u_i(s_i', b) > u_i(s_i, b)$. Player i will nevertheless win the debate (and attain justifiability) if he can counter-argue:

"If I had followed your recommendation and played s_i', the realized path would have been different and so would the set of MPCB's of player j's strategy. For any MPCB b' given that path, you could have argued that s_i does better than s_i' against b'."

Let us now state the formal definition of justifiable strategies:

Definition 2.1. s_i is **justifiable**, given $h(s_1, s_2)$ and a plausibility criterion R, if there exists $b \in B_j{}^*(s_1, s_2)$, such that for every s_i' satisfying $u_i(s_i', b) > u_i(s_i, b)$ and for every $b' \in B_j{}^*(s_i', b)$, it is the case that $u_i(s_i, b') > u_i(s_i', b')$.

Definition 2.2. The strategy profile (s_1, s_2) is an **equilibrium in justifiable strategies** (EJS) under a plausibility criterion R, if s_1 and s_2 are justifiable, given $h(s_1, s_2)$ and R.

The structure of Definition 2.1 embodies an important rhetorical principle: *a counter-argument that turns the initial argument against itself is effective*. The player's counter-argument shows that accepting the critic's recommendation would have subjected the player to the same kind of criticism, only in the opposite direction. Essentially, what he says is: "*You would have continued to criticize me on similar grounds, even if I had done what you are saying I should have done; therefore, your criticism has no real bite.*"

Of course, the idea of formulating solution concepts in such an "argumentative" manner has many precedents in *cooperative* game theory. The definitions of the bargaining set and the kernel have an objection/counter-objection structure that is analogous to EJS. In Rubinstein *et al.* (1992) and Osborne and Rubinstein (1994, Chapters 14–15), the Nash bargaining solution, the Shapley value and the Nucleolus are defined in a similar way.

By Definition 2.1, a profitable unilateral deviation does not automatically knock out a putative equilibrium. In contrast, the following are two ways of showing that a strategy pair is not an EJS, which will be used in the applications.

(1) A profitable deviation that does not alter the set of MPCB's of the opponent's strategy.
(2) A profitable deviation, which is also a best reply to some MPCB given the path induced by the deviation.

EJS can be viewed as an extension of self-confirming equilibrium. A strategy pair (s_1, s_2) is a pure-strategy self-confirming equilibrium (SCE) if for every $i, j \in \{1, 2\}, i \neq j$, there exists

a belief $b \in S_j$ such that s_i is a best-reply to b and b is consistent with $h(s_1, s_2)$.[3] In both EJS and SCE, a strategy is justified by a consistent belief. In SCE, players optimize with respect to their beliefs. In EJS, they do not necessarily optimize but their beliefs are subjected to further constraints. When all beliefs are equally plausible, the two concepts are equivalent:

Remark. Suppose that sRs' for every $s, s' \in S_i, i \in \{1, 2\}$. Then, a strategy profile (s_1, s_2) is an EJS if and only if it is a pure-strategy self-confirming equilibrium. (The proof is simple and therefore omitted.)

Definition 2.1 does not guarantee existence or uniqueness of a justifiable strategy for player i, given s_j. Moreover, since Definition 2.2 involves pure strategies, existence of EJS is not guaranteed in general. Of course, existence and characterization of EJS are sensitive to the structure of R.

3. A SIMPLICITY-BASED PLAUSIBILITY CRITERION

This section presents a particular plausibility criterion, according to which *simpler* beliefs are considered more plausible. Thus, $B_j^*(s_1, s_2)$ is the set of simplest consistent beliefs of player j's strategy, given $h(s_1, s_2)$.

The idea of simplicity as a theory-selection criterion has a very long tradition in the philosophy of science, as well as a more recent one in the algorithmic-complexity literature (Li and Vitanyi, 1997). To motivate the introduction of simplicity as a plausibility criterion, consider the repeated Prisoner's Dilemma and suppose that along the play path, both players always cooperated. "Always cooperate" and "Tit-for-Tat" are consistent beliefs of either player's behaviour. However, "always cooperate" is simpler, in the sense that it dispenses with the threat to punish defection, which "Tit-for-Tat" ascribes to the player.

In order to formalize simplicity judgments, I use the language of finite automata, following the definition given by Osborne and Rubinstein (1994). Let us focus on finite games, in which player i has a constant action set A_i throughout his decision nodes. A finite automaton that represents (player j's belief of) a pure strategy for player i is a quadruple $(Q_i, q_i^\circ, f_i, \tau_i)$, where Q_i is a finite set of states, $q_i^\circ \in Q_i$ is the automaton's initial state, $f_i\colon Q_i \to A_i$ is an output function that assigns an action to every state,[4] and $\tau_i : Q_i \times A_j \to Q_i$ is a transition function that assigns a state to every pair consisting of player i's own machine state and player j's last observed action.

A pure strategy in a finite game has infinitely many finite automata representations. Therefore, b_j will always stand for a *particular* automata representation of a belief of player j's pure strategy. It should be stressed that the automata formalism is used to represent the players' beliefs, *not* their own strategies.

Let $b = (Q, q^\circ, f, \tau)$ be a belief of player j's strategy. Denote the total number of states and transitions in the automaton by $c(b)$ and $t(b)$, respectively. Formally, $c(b) = |Q|$ and $t(b) = \Sigma_{q \in Q} t(q)$, where $t(q)$ is the number of transitions from the state $q \in Q$, defined by $t(q) = |\cup_{a \in A_i} \tau(q, a)|$. Note that $1 \leq t(q) \leq |A_i|$.

Definition 3.1. We will say that b' is *simpler than* b if $c(b') \leq c(b)$ and $t(b') \leq t(b)$, with at least one strict inequality. We will say that $b' R b$ if b is not simpler than b'.

3. Fudenberg and Levine (1993) allow mixed consistent beliefs as well as strategies. I restrict attention to pure strategies and beliefs.

4. I refer to q as an a-state if $f(q) = a$.

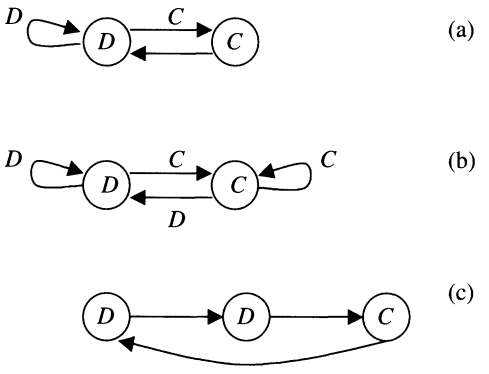


FIGURE 2

Thus, $b'Rb$ if b has exactly as many states and transitions as b', or if b has more states or more transitions than b'. Clearly, R is a weak order. This definition reflects the idea that strategies with a smaller number of contingencies are simpler. Banks and Sundaram (1990) employ a similar, though stronger definition. Of course, this is just an example of a simplicity criterion, which can become unnatural and mechanistic when pushed too far.[5]

Let us use the repeated Prisoner's Dilemma to illustrate this definition. Consider Figure 2. Strategy (a) is simpler than strategy (b) because it has the same number of states (two) and fewer transitions (three versus four). Also, strategy (a) is simpler than strategy (c) because it has the same number of transitions (three) and fewer states (two versus three). In contrast, there is no strict simplicity ranking between (b) and (c), because (b) has more transitions and (c) has more states.

The comparison between (a) and (b) captures the idea that removing threats simplifies a strategy. If ascribing a threat to the opponent is unnecessary for a consistent description of his behaviour, then it cannot be part of an admissible belief. The comparison between (a) and (c) captures the following idea. If one theory of the opponent's behaviour uses a single "state of mind" to describe his behaviour at two different periods, whereas another theory uses two different "states of mind" to describe his behaviour in those periods, then the former is simpler. In both cases, the simpler strategy is the one containing a smaller number of contingencies.

Given the simplicity-based plausibility criterion, usually $s_j \notin B_j^*(s_i, s_j)$ when s_i is a best-reply to s_j. Optimizing against s_j usually means that some of the contingencies in s_j are deliberately avoided by player i. Therefore, these contingencies cannot be part of a simplest consistent belief of player j's strategy. For example, in the centipede game represented by Figure 1, if player 1's strategy is $S(5)$ and player 2 optimizes by playing $S(4)$, then "always C" (in a single-state automata representation) is the *unique* MPCB of player 1's strategy. This property makes justifiability differ in an interesting way from maximization.

Simplicity can clash with other plausibility criteria. For instance, in the finitely repeated Prisoner's Dilemma, "always cooperate" is a dominated strategy, hence implausible by rationality-based criteria. Thus, when the critic attacks the player's performance against an MPCB given the plausibility criterion R given by Definition 3.1, the player can counter-argue that R is an inappropriate criterion to begin with. In this paper, however, we are only concerned

5. An earlier version of the paper contained an alternative definition of simplicity, following the approach taken by Binmore *et al.* (1998), according to which b' is simpler than b if b' can be derived from b by successive operations of merging transitions, merging states and deleting states. This is a weaker and in a sense, more intuitive definition. However, it would be more cumbersome to present and it leads to the same results in the applications as Definition 3.1.

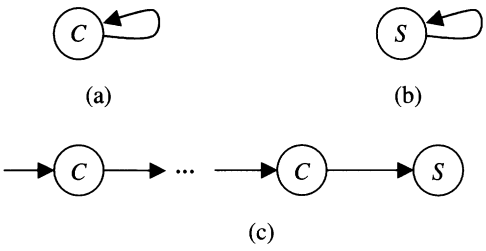


FIGURE 3

with justifying one's strategy *without* challenging the plausibility criterion underlying the critic's argument. The fact that the player accepts the premise of the critic's argument makes his counter-argument all the more effective.

4. APPLICATIONS

In this section, EJS (with the simplicity-based plausibility criterion) is applied to three familiar extensive-form games. It is shown that EJS leads to departures from maximizing behaviour. At the individual level, the departures are small in the sense that players never do worse than second-best replying. However, the play paths that emerge are substantially different from Nash equilibrium. Once again, I wish to emphasize that the applications are not meant to resolve the paradoxes associated with the games, but to illustrate EJS. However, I do believe that they capture considerations and patterns of behaviour that are pertinent to these games.

Example 4.1 *(The centipede game).* I study the game represented by Figure 1.[6] Let K be the number of nodes in the game. Denote the strategies "always continue" and "always stop" for player i by $c_i{}^*$ and $s_i{}^*$ (the subscripts are occasionally omitted). Denote the strategy "stop at the kth node" ($k > 2$) by $S(k)$. "Always stop" and "always continue" are represented by single-state automata, whereas $S(k)$ is represented by an automaton with $k/2$ states, rounded off to the higher integer (see Figure 3). We will say that s_j stops before s_i if player j terminates the game in (s_i, s_j).

The crucial observation for our analysis is that $c_i{}^*$ is simpler than $S(k)$ (for any $k > 1$), since it has a smaller number of states and transitions. Thus, if player j stops before player i, $c_i{}^*$ is an MPCB of player i's strategy, even if his true strategy is different.

The Nash outcome (immediate termination) is consistent with EJS. Consider the Nash equilibrium $(s_1{}^*, s_2{}^*)$. Since $s_i{}^*$ is an MPCB of player i's strategy and $s_j{}^*$ is a best reply to $s_i{}^*$, $s_j{}^*$ is justifiable. I now prove that there exists another EJS:

Proposition 1. *$(c_1{}^*, S(K))$ is an EJS in the centipede game. Moreover, $h(c_1{}^*, S(K))$ is the unique EJS outcome other than the Nash outcome.*

Proof. First, let us show that a path which is not terminated at the first or last nodes cannot be supported by an EJS. Suppose that s_1 stops before s_2. Because player 2 chooses nothing but C in (s_1, s_2), $c_2{}^*$ is an MPCB of his strategy, given $h(s_1, s_2)$. By the payoff structure of the centipede game (see Figure 1), $c_1{}^*$ does better than s_1 against $c_2{}^*$. Moreover, $c_2{}^*$ is an MPCB of player 2's strategy, given $h(c_1{}^*, c_2{}^*)$. We have constructed a profitable deviation (from s_1 to

6. The analysis holds for any "centipede-game" payoff structure.

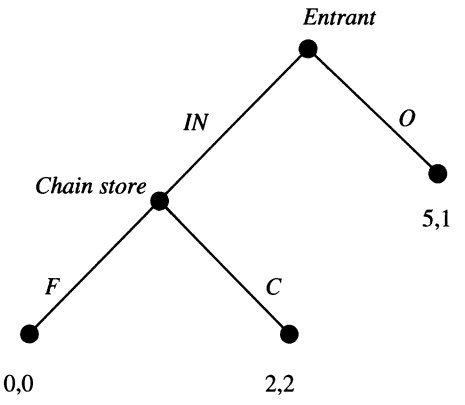


FIGURE 4

c_1^*) that induces a path, in which c_2^* remains an MPCB of player 2's strategy. Thus, s_1 is not justifiable, given $h(s_1, s_2)$. A virtually identical argument works in case s_2 stops before s_1.

It remains to show that $(c_1^*, S(K))$ is an EJS. Consider player 2's reasoning first. Since he obtains the highest payoff that is possible for him in the game, his strategy is trivially justifiable. As to player 1's reasoning, $S(K)$ is an MPCB of player 2's strategy, given $h(c_1^*, S(K))$. Any profitable deviation to $s_1' \neq c_1^*$ stops before $S(K)$ (and after the 1st node). But then, given $h(s_1', S(K))$, c_2^* is the unique MPCB of player 2's strategy. Furthermore, $u_1(c_1^*, c_2^*) > u_1(s_1', c_2^*)$. Thus, c_1^* is justifiable, given $h(c_1^*, S(K))$. ||

In this equilibrium, Player 1 is willing to bear the loss from not stopping before player 2 because otherwise, it would be hard to explain why he stopped "all of a sudden".

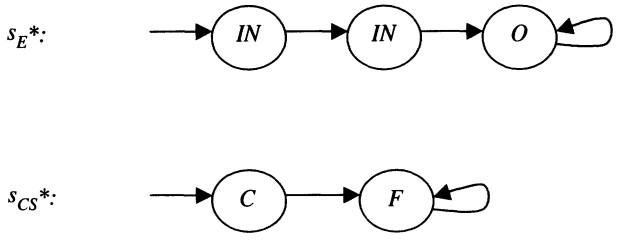
Example 4.2 *(The Chain Store Game).* I study a two-person version of the game. A chain store and an entrant repeatedly play the game represented by Figure 4 for $K > 2$ periods. The entrant's available actions are "out" (O) and "in" (IN). The chain store's available actions are "fight" (F) and "accommodate" (C). In every Nash equilibrium, the outcome of any stage game is either O or (IN, C). The well-known paradox associated with this game concerns the sub-game perfect outcome, which is (IN, C) at every period. However, the possibility of fighting along the play path, which is the focus of our attention in this example, is already ruled out by Nash equilibrium.

Consider the strategy profile given by Figure 5 and its induced path, denoted h^*.

Proposition 2. (s_{CS}^*, s_E^*) *is an EJS.*

Proof. Consider the entrant's reasoning first. The chain store's strategy, s_{CS}^*, is an MPCB, given h^*. The only profitable deviation for the entrant against s_{CS}^* is to play O at period 2. This deviation induces the following path:

Period	1	2	3	4	···	K
Entrant	IN	O	O	O	···	O
Chain store	C	—	—	—	···	—



Period	1	2	3	4	···	*K*
Entrant	*IN*	*IN*	*O*	*O*	···	*O*
Chain store	*C*	*F*	-	-	···	-

FIGURE 5

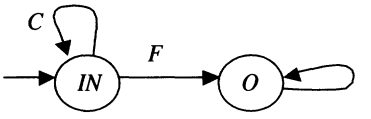


FIGURE 6

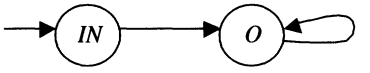


FIGURE 7

Given this path, "always C" is the unique MPCB of the chain store's strategy. The entrant's original strategy s_E* does better than the deviant strategy against "always C". Therefore, s_E* is justifiable, given h*.

Let us turn to the chain store's reasoning. Figure 6 represents an MPCB of the entrants' strategy given h*, denoted by b_E*. The chain store's only profitable deviation against b_E* is to play F at the first period, which induces the following path:

Period	1	2	3	4	···	K
Entrant	IN	O	O	O	···	O
Chain store	F	—	—	—	···	—

Given this path, the unique MPCB of the entrant's strategy, denoted b_E', is given by Figure 7. Note that b_E' is simpler than b_E*. The chain store's original strategy s_{CS}* does better than the deviant strategy against b_E'. Therefore, s_{CS}* is justifiable, given h*. ||

This EJS captures the following pattern of behaviour. The chain store pursues the goal of driving the entrant out of the market. It first tries to achieve this goal "peacefully" and switches to aggressive behaviour as soon as the "peaceful tactics" fail. The opposite behaviour pattern,

	A	B
A	$X_1(k)$- ε, $X_2(k)$-ε	-ε, $X_2(k)$
B	$X_1(k)$, -ε,	0 , 0

FIGURE 8

namely aggressive play followed by accommodating behaviour by the chain store, is inconsistent with EJS. Consider the following play path, denoted h':

Period	1	2	3	4	···	K
Entrant	IN	IN	IN	IN	···	IN
Chain store	F	C	C	C	···	C

The unique MPCB of the entrant's strategy, given this path, is "always IN". It would continue to be an MPCB if the chain store profitably deviated to "always C". Therefore, the chain store's original strategy is not justifiable, given h'.

Example 4.3 *(Inter-temporal Exchange in a Finite-horizon Model).* The following example is a two-person variant on a well-known OLG model dating back to Samuelson (1958). Nash equilibrium in this class of models prescribes no inter-temporal exchange when the time horizon is finite.

Consider a K-period economy with two players, 1 and 2, and two non-durable goods, 1 and 2. Player i can produce good i but obtains utility only from consuming good j. At each period, player i decides whether to produce one unit of good i (A) or not (B). Player i's utility from consuming one unit of good j (i) is 1 (0). The cost of producing either good is ε per unit, $0 < \varepsilon < 1$. The exchange technology is imperfect—direct simultaneous exchange is infeasible. Instead, goods are sold for "coupons". If player i produces at period k, he receives one coupon, with which he can purchase one unit of good j that is produced at any $k' > k$.

Formally, for every period $k \in \{1, \ldots, K\}$, let $p_i(k) = 1$ if player i plays A at k and $p_i(k) = 0$ if he plays B at k ($p_i(0) = 0$ for all $i \in \{1, 2\}$). Let $X_i(k) = 1$ if $\Sigma_{h<k}[p_i(h) - p_j(h-1)] > 0$ and $X_i(k) = 0$ otherwise. Thus, $X_i(k)$ signifies whether player i has any coupons left at period k from previous periods of production and trade. Each player's total utility is the sum of his periodic payoffs. The players' payoffs at period k are given by Figure 8.

In the unique Nash equilibrium of this game, both players play always B. This is also an EJS because "always B" is an MPCB and each player optimizes against this belief. Now, consider the following play path, denoted h^*, where $1 < k^* < K$:

Period	1	...	k^*-1	k^*	k^*+1	...	K
Player 1	A	...	A	A	B	...	B
Player 2	A	...	A	A	B	...	B

Proposition 3. *h^* is supported by an EJS.*

Proof. Figure 9 represents an MPCB of either player's strategy given h^*, denoted by b^*. Suppose that (b^*, b^*) is also the true strategy profile.

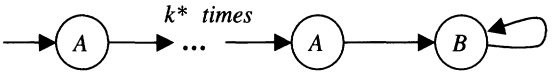


FIGURE 9

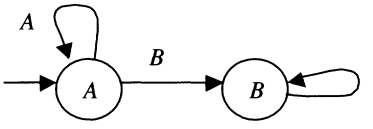


FIGURE 10

By the game's payoff structure, the only profitable deviation for either player is to play B, instead of A, at period k^*. The path that results from such a deviation by player 1 is:

Period	1	...	k^*-1	k^*	k^*+1	...	K
Player 1	A	...	A	B	B	...	B
Player 2	A	...	A	A	B	...	B

The unique MPCB of player 2's strategy, given this path is represented by Figure 10. Against this belief, player 1's deviant strategy does worse than his original strategy, which is therefore justifiable, given h^*.

This EJS captures the following reasoning. The optimal strategy for either player is to stop producing right before his opponent intends to stop. However, if the player optimizes, the most plausible *ex post* theory will be that his decision to stop producing *caused* the opponent to stop as well. This is an intuitive inference because a close temporal link is intuitively interpreted as a causal link. Since neither player wants to be held responsible for his opponent's decision to stop producing, both players stop simultaneously in equilibrium.

5. DISCUSSION

An agent's accountability often concerns the *reasons* behind his decisions, in addition to their consequences. In this paper, I have modelled this form of accountability as a post-game debate, which reviews the optimality of the player's strategy, as well as the plausibility of the belief that supports it. In the remainder of this section, I would like to discuss matters of interpretation, which are raised by this model.

5.1. *Can justifiability be rationalized by utility maximization?*

So far, we have compared between justifiability and utility maximization, where both are defined with respect to a given payoff function. One may ask whether justifiability with respect to some payoff function is equivalent to maximization of *another* utility function over the terminal nodes. In other words, perhaps the difference between justifiability and maximization can be reduced

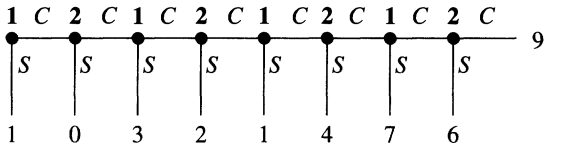


FIGURE 11

to a different utility function. The following example demonstrates that this is not the case in general.

Consider the game, represented by Figure 11, which is identical to the centipede game of Figure 1, except for a single modification of player 1's payoff at the fifth node. Player 2's payoffs are omitted.

Two strategy pairs, (c_1*, $S(4)$) and ($S(5)$, $S(4)$) (using the notation of example 4.1), induce the same play path, which is terminated by player 2 at the fourth node. As we saw in Section 4, the strategy c_1* for player 1 is justifiable, given this path. In contrast, the strategy $S(5)$ is not justifiable. To see why, observe that deviating to $S(3)$ is profitable for player 1. The unique MPCB of player 2's strategy, given the path induced by this deviation, is c_2*. Since $S(5)$ does *worse* than $S(3)$ against c_2*, $S(5)$ is not justifiable, given $h(S(5), S(4))$.

It follows that two strategies, which lead to the same play path, may differ in their justifiability. The reason is that justifiability evaluates strategies not only by their performance in the actual play path, but also by their hypothetical performance in counter-factual play paths. Thus, *justifiability cannot be rationalized by maximization of some utility function over the game's terminal nodes.*

If the definition of a consequence is extended so as to consist of a terminal history h and the player's own strategy s, justifiability can clearly be rationalized by a two-valued function U, which takes the value of 1 if s is justifiable given h and 0 otherwise. This, however, is a trivial representation, which amounts to a restatement of Definition 2.1.

5.2. *Can EJS be embedded in a strict non-cooperative game?*

A related question is whether the justification process can be modelled as a strict non-cooperative game, in which the critic is also a player, such that EJS would appear as a "reduced form" model of some standard equilibrium concept in a "larger game". If we think of the critic as an agent, whose sole motive is to win the post-game debate, then modelling the debate as a zero-sum sub-game between the player and his critic is straightforward. However, it adds no further insight and merely complicates the exposition.[7]

On the other hand, if we think of the critic as a principal, who basically wishes "his" player to perform well, then it is not clear how to pursue a strict non-cooperative approach. Standard equilibrium analysis presumes that all players share the same beliefs of the strategy profile. But if player 1, say, and his critic hold the same belief of player 2's strategy, what are they arguing

7. Note that the set of Nash equilibria in this extended game can also contain outcomes that are not EJS according to Definition 2.2. For example, suppose that neither player has a justifiable response to the opponent's strategy. Then, the players' strategies constitute a NE of the extended game.

about? The question of how to embed EJS in standard equilibrium analysis of a larger game, in which the critic is a player, is thus left for future research.

5.3. *Justifying strategies versus justifying actions*

In this model, players justify their *strategies*. Alternatively, players could be modelled as justifying the *actions* they took along the play path. If each player announces his intended strategy to his critic at the outset, then Definition 2.1 seems appropriate. Otherwise, we need not assume that player i's critic observes s_i better than s_j and the alternative formulation should be preferred. Note that in such a model, the argument of subsection 5.1 against the rationalization of justifiability fails.

The main reason for preferring the present formulation of EJS is methodological: it is useful to have a version of EJS that departs as little as possible from self-confirming equilibrium. Assuming that the critic takes as given *everything* except for the opponent's strategy is a minimal relaxation of standard equilibrium analysis.

A related point is that players in this model justify their behaviour at the end of the game, once and for all, using an *ex post* belief. The beliefs they may have held during the game or the way they responded to such beliefs do not enter into the justification procedure. It would be interesting to extend the model in this direction.

5.4. *Plausibility as a burden-of-proof criterion*

One way of thinking about R is as a prior-probability ranking: bRb' if b has a higher prior probability than b'. There is a slight problem with this interpretation, namely that it implies that EJS displaying sub-optimal behaviour are relatively improbable events, since the role of sub-optimal play in EJS is to rule out a plausible belief, in favour of a less plausible one.

I prefer to interpret R as a criterion for assigning the burden of proof in the post-game debate. To use an analogy, the presumption of innocence in criminal trials means that the onus is imposed on the prosecution. It does not always reflect a prior belief—even a reputed mobster is presumed innocent at the beginning of his trial. The burden-of-proof assignment may have something to do with prior beliefs, but it has other rationales as well.

Similarly in our model, if player 1 wishes to justify s_1 with $b \in S_2$, he must accept the burden of proof that more plausible beliefs are incorrect. Recall the Prisoner's Dilemma example in Section 3. Under the simplicity-based plausibility criterion, the presumption is that player 2's strategy contains no threats to punish defection, unless proven otherwise.

The rationale for this burden-of-proof assignment is not necessarily that player 2's threat is improbable *a priori*. One goal that is served by the simplicity-based criterion is that players do not get away too easily with almost any kind of behaviour. If the onus were not imposed on player 1, he could justify any individually rational outcome by ascribing "grim" threats to the opponent. Basically, he would be able to argue: "If I had not done exactly what I did, it would have been a catastrophe". This seems like an excessively facile justification and a stronger burden-of-proof criterion is intuitively required.

5.5. *Alternative plausibility criteria*

EJS can be fruitfully applied with other plausibility criteria than simplicity. The chain store game, for example, belongs to an important class of dynamics games, in which strategies are sometimes distinguishable by the type of *reciprocity* that they exhibit. If a player reacts to soft

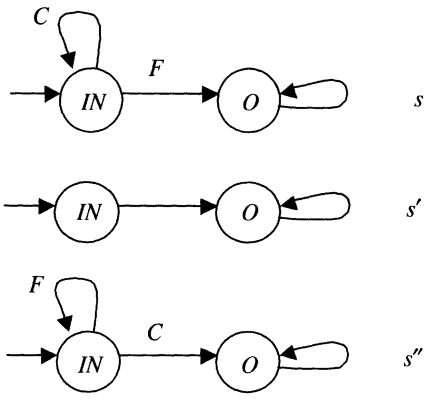


FIGURE 12

(aggressive) play with soft (aggressive) play, his strategy displays "positive reciprocity" (tit-for-tat in the repeated Prisoner's Dilemma displays the ultimate positive reciprocity). Conversely, if a player reacts to soft play with aggressive play and vice versa, his strategy displays "negative reciprocity". Finally, when the player's behaviour is independent of his opponent's actions, his strategy displays no reciprocity.

Players (and critics) often approach strategic interactions with a tendency to attribute to their opponent a certain type of reciprocity. For example, they may find it more plausible to attribute negative reciprocity to the opponent than to attribute to him positive or no reciprocity. This subsection illustrates the effect of this plausibility criterion in the chain store game.

For simplicity, suppose that when justifying its strategy, the chain store can only use beliefs with a *two-state* automata representation: one IN-state and one O-state. Beliefs that display greater *negative reciprocity* are considered more plausible *a priori*. Thus, the beliefs represented by Figure 12 are ranked as follows: $sRs'Rs''$.

Thus, it is more plausible to believe that the entrant reacts softly to aggressive play and vice versa (relative to the belief that it reacts softly to soft play and aggressively to aggressive play, or that its entry decisions do not depend on the chain store's actions). A similar assumption can be made for the entrant's beliefs of the chain store's strategy.

Under these assumptions, the following play path can be supported by EJS:

Period	1	2	3	...	K
Entrant	IN	IN	IN	...	IN
Chain store	F	C	C	...	C

Consider the chain store's reasoning. Suppose that its strategy is to play F against the first entry and C against every subsequent entry. The chain store justifies this strategy with the belief s' represented by Figure 12, which is an MPCB, given the play path. The only profitable deviation is to play C at the first period. This deviation induces the following path:

Period	1	2	...	K
Entrant	IN	IN	...	IN
Chain store	C	C	...	C

The unique MPCB of the entrant's strategy, given this path, is represented by s in Figure 1. The chain store's original strategy does better than the deviant strategy against s. Therefore, it

is justifiable given the original play path. The entrant's behaviour can be supported in similar fashion.

In general, EJS with a plausibility criterion based on "negative reciprocity" gives rise to play patterns with fighting followed by accommodation by the chain store (in contrast to the patterns implied by the simplicity-based criterion). I conjecture that in a game such as the repeated Prisoner's Dilemma, the two plausibility criteria would lead to similar results.

5.6. *Related literature*

This paper is related to several strands in the literature. Rubinstein (1986), Abreu and Rubinstein (1988), Kalai and Stanford (1988), Banks and Sundaram (1990) and others introduced complexity considerations into players' reasoning. The difference is that in the present paper, players are concerned with the complexity of their *beliefs*, rather than their own strategy. Simplicity is a rhetorical consideration, which is not directly payoff-relevant. For another model that applies complexity considerations to beliefs, see Eliaz (2002).

As an analysis of games with procedurally rational players, this paper is related to Rosenthal (1989), McKelvey and Palfrey (1995) and Osborne and Rubinstein (1998). Finally, relaxing the assumption that players know the strategy profile in favour of the assumption that they have beliefs, which satisfy consistency in conjunction with other criteria, has precedents in Battigalli and Guatoli (1988) and Rubinstein and Wolinsky (1994), among other works.

Acknowledgements. This paper is based on a section of my Ph.D. dissertation (Tel-Aviv University). I am deeply grateful to Ariel Rubinstein for his supervision. I thank Colin Camerer, Kfir Eliaz, Gilat Levy, two anonymous referees and numerous seminar participants, for useful comments. I owe special thanks to Eddie Dekel for detailed comments on an earlier draft. Financial support from the Israel Science Foundation and the Arthur Goodhart fund is gratefully acknowledged.

REFERENCES

ABREU, D. and RUBINSTEIN, A. (1988), "The Structure of Nash Equilibrium in Repeated Games with Finite Automata", *Econometrica*, **56**, 1259–1282.

BANKS, J. and SUNDARAM, R. (1990), "Repeated Games, Finite Automata, and Complexity", *Games and Economic Behavior*, **2**, 97–117.

BATTIGALLI, P. and GUATOLI, D. (1988), "Conjectural Equilibria and Rationalizability in a Macroeconomic Game with Incomplete Information" (University Bocconi).

BINMORE, K., PICCIONE, M. and SAMUELSON, L. (1998), "Evolutionary Stability in Alternating-offers Bargaining Games", *Journal of Economic Theory*, **80**, 257–292.

ELIAZ, K. (2002), "Nash Equilibrium when Players Account for the Complexity of their Forecasts", *Games and Economic Behavior,* forthcoming.

FUDENBERG, D. and LEVINE, D. (1993), "Self-Confirming Equilibrium", *Econometrica*, **61**, 523–545.

KALAI, E. and STANFORD, W. (1988), "Finite Rationality and Interpersonal Complexity in Repeated Games", *Econometrica*, **56**, 397–410.

LI, M. and VITANYI, P. (1997), *Introduction to Kolmogorov Complexity and Applications,* 2nd edition (Springer).

MCKELVEY, R. D. and PALFREY, T. R. (1995), "Quantal Response Equilibria for Normal Form Games", *Games and Economic Behavior*, **10**, 6–38.

OSBORNE, M. J. and RUBINSTEIN, A. (1994), *A Course in Game Theory* (Cambridge, MA: MIT Press).

OSBORNE, M. J. and RUBINSTEIN, A. (1998), "Games with Procedurally Rational Players", *American Economic Review*, **88**, 834–847.

ROSENTHAL, R. (1989), "A Bounded-rationality Approach to the Study of Non-cooperative Games", *International Journal of Game Theory*, **18**, 273–92.

RUBINSTEIN, A. (1986), "Finite Automata Play a Repeated Prisoner's Dilemma", *Journal of Economic Theory*, **39**, 83–96.

RUBINSTEIN, A., SAFRA, Z. and THOMSON, W. (1992), "On the Interpretation of the Nash Bargaining Solution and its Extension to Non-Expected Utility Preferences", *Econometrica*, **60**, 1171–1186.

RUBINSTEIN, A. and WOLINSKY, A. (1994), "Rationalizable Conjectural Equilibrium: Between Nash and Rationalizability", *Games and Economic Behavior*, **6**, 299–311.

SAMUELSON, P. A. (1958), "An Exact Consumption-Loan Model of Interest with or without the Social Contrivance of Money", *Journal of Political Economy*, **66**, 467–482.

SHAFIR, E., SIMONSON, I. and TVERSKY, A. (1993), "Reason-based Choice", *Cognition*, **49**, 11–36.

TETLOCK, P. E. and BOETTGER, R. (1989), "Accountability: A Social Magnifier of the Dilution Effect", *Journal of Personality and Social Psychology*, **57**, 388–298.
TETLOCK, P. E. and BOETTGER, R. (1994), "Accountability Amplifies the Status Quo Effect when Change Creates Victims", *Journal of Behavioral Decision Making*, **7**, 1–23.
TVERSKY, A. and SHAFIR, E. (1992), "Choice under Conflict: The Dynamics of the Deferred Decision", *Psychological Science*, **3**, 358–361.
WALTON, D. N. (1989), *Informal Logic* (Cambridge: Cambridge University Press).

[Footnotes]

[3] **Self-Confirming Equilibrium**

Drew Fudenberg; David K. Levine

Econometrica, Vol. 61, No. 3. (May, 1993), pp. 523-545.

Stable URL:

http://links.jstor.org/sici?sici=0012-9682%28199305%2961%3A3%3C523%3ASE%3E2.0.CO%3B2-I

References

The Structure of Nash Equilibrium in Repeated Games with Finite Automata

Dilip Abreu; Ariel Rubinstein

Econometrica, Vol. 56, No. 6. (Nov., 1988), pp. 1259-1281.

Stable URL:

http://links.jstor.org/sici?sici=0012-9682%28198811%2956%3A6%3C1259%3ATSONEI%3E2.0.CO%3B2-X

Self-Confirming Equilibrium

Drew Fudenberg; David K. Levine

Econometrica, Vol. 61, No. 3. (May, 1993), pp. 523-545.

Stable URL:

http://links.jstor.org/sici?sici=0012-9682%28199305%2961%3A3%3C523%3ASE%3E2.0.CO%3B2-I

LINKED CITATIONS

Finite Rationality and Interpersonal Complexity in Repeated Games
Ehud Kalai; William Stanford
Econometrica, Vol. 56, No. 2. (Mar., 1988), pp. 397-410.
Stable URL:
http://links.jstor.org/sici?sici=0012-9682%28198803%2956%3A2%3C397%3AFRAICI%3E2.0.CO%3B2-F

Games with Procedurally Rational Players
Martin J. Osborne; Ariel Rubinstein
The American Economic Review, Vol. 88, No. 4. (Sep., 1998), pp. 834-847.
Stable URL:
http://links.jstor.org/sici?sici=0002-8282%28199809%2988%3A4%3C834%3AGWPRP%3E2.0.CO%3B2-6

On the Interpretation of the Nash Bargaining Solution and Its Extension to Non-Expected Utility Preferences
Ariel Rubinstein; Zvi Safra; William Thomson
Econometrica, Vol. 60, No. 5. (Sep., 1992), pp. 1171-1186.
Stable URL:
http://links.jstor.org/sici?sici=0012-9682%28199209%2960%3A5%3C1171%3AOTIOTN%3E2.0.CO%3B2-T

An Exact Consumption-Loan Model of Interest with or without the Social Contrivance of Money
Paul A. Samuelson
The Journal of Political Economy, Vol. 66, No. 6. (Dec., 1958), pp. 467-482.
Stable URL:
http://links.jstor.org/sici?sici=0022-3808%28195812%2966%3A6%3C467%3AAECMOI%3E2.0.CO%3B2-Z