JOURNAL OF ECONOMIC THEORY **68**, 1–25 (1996)
ARTICLE NO. 0001

Deviations, Dynamics, and Equilibrium Refinements*

MATTHEW RABIN

Department of Economics, University of California—Berkeley, Berkeley, California 94720

AND

JOEL SOBEL

Department of Economics, University of California—San Diego, La Jolla, California 92093

Received May 11, 1993; revised October 24, 1994

Existing equilibrium refinements rule out Nash equilibria susceptible to deviations. We propose a framework for considering not only equilibria impervious to deviations, but also equilibria likely to recur in the long run because they are repeatedly deviated to. We explore which equilibria are recurrent with respect to the deviations underlying some existing signaling refinements. We show that the set of recurrent equilibria based on Cho and Kreps's (1987) intuitive criterion is equivalent to their solution concept, but that applying our framework to existing cheap-talk refinements make those solution concepts more realistic and guarantee existence where their current formulations do not. *Journal of Economic Literature* Classification Numbers: B49, C72, C73.

1. INTRODUCTION

It is a basic tenet of non-cooperative game theory that if players settle upon predictable play, this play will be a Nash equilibrium. Yet many standard solution concepts propose that some Nash equilibria are themselves susceptible to deviations. Such deviation-based refinements have been developed in the signaling literature (see [2, 7]), most prominently in the cheap-talk literature (see [10, 17]). While these refinements identify equilibria that are susceptible to deviations, they leave an important question unanswered: If players are likely to deviate from some equilibrium, what will they deviate *to*? We feel this question is important because some equilibria that are subject to deviations may still persist in

* We thank Andreas Blume, Antonio Cabrales, Eddie Dekel-Tabak, Joe Farrell, In-Uck Park, Iñigo Zapater, two anonymous referees, an associate editor, and participants of theory workshops at Berkeley, University of Iowa, and the University of Chicago for helpful comments. Each author also thanks the National Science Foundation for financial support.

0022-0531/96 $18.00

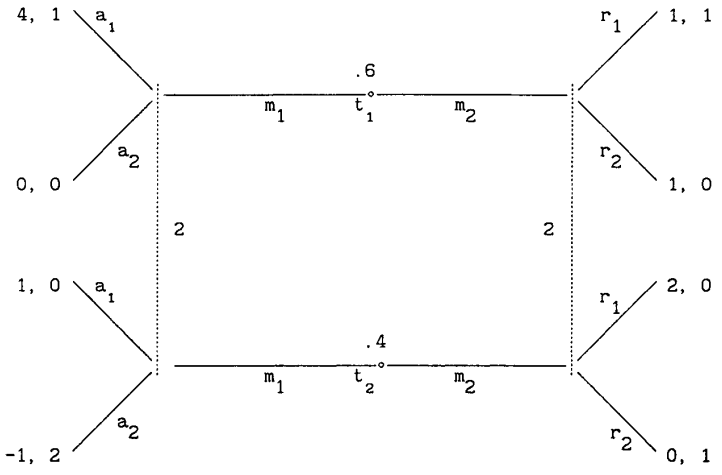

FIGURE 1

the long run because they may be repeatedly deviated to. It seems inappropriate to omit such recurrent equilibria from the solution concepts with which we make predictions in economic models.

In this paper, we develop a framework to address this question and use this framework to critique existing solution concepts. Our framework involves two steps. First, based on intuitions from existing refinements for when and how deviations occur, we develop an algorithm for specifying explicitly the outcomes that might arise following a deviation from an equilibrium. Second, we then develop solution concepts based on which equilibria recur infinitely often given a theory of deviations. We show that our framework gives a more complete dynamic justification for a prominent existing solution concept, but also that it alters and (we feel) improves upon the solution concepts resulting from other theories of deviations.

Our first step, specifying which equilibria might follow deviations, addresses a concern commonly expressed regarding signaling refinements. Signaling refinements have been criticized because proposed deviations from equilibria often rely on an apparent inconsistency: Some players defect from an equilibrium while others continue to believe in that equilibrium. Our approach to deviations takes this general criticism, often called the *Stiglitz critique*, into account.[1] The standard Sender–Receiver signaling game in Fig. 1 illustrates this critique. (The first of each pair of payoffs is that of the informed Sender; the second is that of the uninformed Receiver.)

[1] As we discuss in the next section, we do not, however, concur with the more specific arguments of the example in [7] in which the critique is most often formulated.

There is a sequential equilibrium of this game in which both types of the Sender choose m_2 and the Receiver chooses a_2 given m_1 and r_1 given m_2. This outcome fails the Cho and Kreps's intuitive criterion. They argue that, given the equilibrium response to the message m_2, only type t_1 could conceivably wish to deviate by sending the message m_1, because type t_2 would get worse than her equilibrium payoff by doing so no matter how she thought the Receiver would respond. The Receiver should therefore interpret the deviation m_1 as being sent by t_1 and should respond with action a_1. If the Receiver acts in this way, then t_1 would prefer to send m_1 rather than the equilibrium message, m_2. Cho and Kreps conclude that this equilibrium is not stable.

We feel this analysis stops too soon. If the Receiver realizes that t_1 might deviate by sending message m_1, then he might respond to m_2 with r_2, which is the optimal response given that only t_2 is sending m_2. Given such a response to m_2, now t_2 may be better off sending m_1. If the Receiver believes that t_2 may therefore play m_1, he may react to m_1 by playing a_2, his original equilibrium response to m_1. This in turn may lead both types of Sender to adhere to her equilibrium strategy. Thus, if the players with common knowledge anticipate a deviation from this equilibrium, they may reasonably choose not to deviate after all. While we accept Cho and Kreps's argument that this equilibrium is unstable in some sense, we feel that the Stiglitz critique does indeed have force—rational players might respond to the hypothesized deviation from this equilibrium by continuing to play the equilibrium.[2]

In Section 2, we formalize the arguments above by defining an expansion process: Starting with a proposed deviation, we iteratively add best responses by each player until no more strategies need be added. This expansion process formally parallels the reasoning outlined above and guarantees that the set of possible reactions to a posited deviation is consistent with common knowledge of rationality. We label the resulting set of strategies the *deviation correspondence*. For a given equilibrium and theory of deviations, a deviation correspondence specifies a set of strategies that players may use following a deviation from the equilibrium.

We complete our framework in Section 3 by describing some broad properties of dynamics that allow us to characterize a set of equilibria that may persist in the long run. The environment we are considering is that of "non-strategic" repeated play—two players repeatedly play a game, observing the outcome each period; we implicitly assume that the players discount the future heavily, so that in every round players each try to maximize

[2] But we also show that the Stiglitz critique does not always recover unstable equilibria. For earlier arguments and examples of when the Stiglitz critique does and does not have force, see [16].

their one-shot payoff.[3] We assume that there is a tendency for play to settle down on equilibrium behavior, but that deviations may upset certain equilibria. In the periods following an equilibrium, players may play strategies in the deviation correspondence, and play re-equilibrates to an equilibrium in the deviation correspondence. We define as *recurrent* those equilibria that are likely to be played repeatedly in the long run in such an environment.[4]

Figure 1 illustrates our approach. Let X be the pooling equilibrium that Cho and Kreps argue is susceptible to a deviation. The deviation correspondence from this equilibrium X will include the other equilibrium in this game, in which t_1 sends the message m_1, t_2 sends the message m_1 with probability 3/4 and m_2 with probability 1/4, and the Receiver responds to m_1 with an equal mixture of a_1 and a_2 and to m_2 with r_2. We call this partially separating equilibrium Y. Then our deviation correspondence for X contains both X and Y. By contrast, because Cho and Kreps argue that Y is stable, its deviation correspondence is itself.

Which of these two equilibria might occur in the long run? While we argued above that X won't necessarily be deviated away from in any given period, we assume that there is a small probability that there will be a deviation to Y, and thus posit that eventually this deviation will occur. Once such a deviation occurs, because Y is stable, it will be played forever. Despite the Stiglitz critique, therefore, the equilibrium Y seems the appropriate long-run prediction in Fig. 1. Indeed, we show in Section 3 that the deviation correspondence based on Cho and Kreps's [7] intuitive criterion always contains an equilibrium that is stable.[5] We therefore assume that play will *eventually* deviate to a stable equilibrium, at which point no unstable equilibrium will again be played. Despite basing our solution concept on recurrent equilibria, our proposed solution concept turns out to be equivalent to Cho and Kreps's solution concept, which includes only equilibria that are not susceptible to deviations.[6]

[3] Our model can also be interpreted as a situation in which each person plays the game only once, where the players each period are a new generation of people who observe previous play of the game.

[4] Our exclusive focus on equilibrium rather than non-equilibrium outcomes makes sense in environments in which deviations are relatively rare and equilibration is relatively quick. We feel that some such story underlies the equilibrium-refinement literature. [14] and [29] develop explicitly dynamic models which guarantee that play will generally be in equilibrium, but that equilibrium behavior will periodically be interrupted by mutations.

[5] Throughout the paper, we refer to equilibria as "stable" if they are not subject to deviations and as "unstable" if they are subject to deviations.

[6] In [26], we show also, that by slightly modifying our framework, Kohlberg and Mertens' [15] never-a-weak-best-response criterion has the same property–equilibria are recurrent with respect to the never-a-weak-best-response criterion if and only if they are stable with respect to the never-a-weak-best-response criterion.

The equivalence between recurrent and stable equilibria is not a general feature of our framework, however, and need not hold for other theories of deviations. Suppose, for instance, that a game has three equilibria, A, B, and C. Suppose that (according to some theory) A is not subject to deviations, and B and C tend to deviate to each other, but *not* to A. The standard refinement literature would have us limit attention to equilibrium A, because it is the only stable equilibrium. This seems inappropriate—if either equilibrium B or C occur, then equilibrium A will never occur again. Our notion of plausible long-run equilibria would therefore include B and C; while neither is stable, each is likely to appear over and over again.

In Section 4, we provide an extended example in cheap-talk games that illustrates such a situation and show that our framework provides an alternative approach to making predictions in such games. Based on the arguments of [17], we develop a solution concept called *recurrent mop*, and provide some examples where recurrent mop improves upon existing cheap-talk solution concepts. We also show that recurrent mop guarantees communication in a class of games where other common cheap-talk refinements do not.

While our framework indicates that existing signaling refinements may eliminate too many equilibria, Section 4 illustrates that our approach has attractions for those who like their solution concepts strong. We feel that many existing solution concepts with non-existence problems are "too strong" not because they have the wrong intuition for which equilibria are prone to deviations, but rather because these solution concepts inappropriately rule out all deviation-prone equilibria. Applying our framework to cheap-talk solution concepts and other concepts such as Grossman and Perry [11] therefore provides a practical approach to making strong predictions. Because we focus on recurrent equilibria rather than stable equilibria, we show in Section 3 that we always guarantee existence, even in situations where current solution concepts do not.

In Section 5, we discuss two potential further applications of our approach. First, we discuss an alternative framework to that taken by Bernheim *et al.* [4] for thinking about "coalition-proof" equilibria. We then conclude the paper by discussing some ideas on how to extend our approach to allow for non-equilibrium behavior.

2. Deviation Correspondences and Expansions

In this section, we develop the first step of our framework, constructing deviation correspondences for two-player games. Players 1 and 2 have finite pure-strategy sets S_1 and S_2, with the set of strategy profiles $S \equiv S_1 \times S_2$, and payoff functions (u_1, u_2). For any finite set Z, we denote

by $\Delta(Z)$ the set of probability distributions on Z, so that a mixed strategy for player i is an element in $\Delta(S_i)$. We extend (u_1, u_2) to mixed strategies in the obvious way. Given a set of strategies $Z_j \subseteq \Delta(S_j)$, we denote by $BR_i(Z_j)$ the set of best responses in S_i to Z_j. That is, $BR_i(Z_j) \equiv \{s_i^* \in S_i |$ there exists $z_j \in Z_j$ such that $u_i(s_i^*, z_j) \geqslant u_i(s_i, z_j)$ for all $s_i \in S_i\}$. Similarly, we define the notion of strong best responses as $SBR_i(Z_j) \equiv \{s_i^* \in S_i |$ there exists a full-support probability distribution p_j over the strategies in Z_j such that $u_i(s_i^*, p_j) \geqslant u_i(s_i, p_j)$ for all $s_i \in S_i\}$.

While our framework could be applied more generally, we limit our attention in this paper to simple signaling games. Player 1 is an informed Sender, who has private information drawn from a finite set of types T according to a common-knowledge distribution $\pi(\cdot)$. Her set of pure strategies consists of rules that assign to each $t \in T$ a message m, which is a member of a finite set M. Player 2 is an uninformed Receiver, who observes which message m is chosen, and then chooses an action a, which belongs to a finite set A. Players have utility functions $u_1(t, m, a)$ and $u_2(t, m, a)$. In Fig. 1, the Sender's type is either t_1 or t_2, and she can choose message m_1 or m_2, after which the Receiver chooses his action. For ease of reference, we label the actions by the Receiver differently depending on which message they follow, so the Receiver chooses between a_1 and a_2 if he observes message m_1, and between r_1 and r_2 if he observes m_2.

In order to describe a deviation correspondence we begin with an equilibrium $\gamma = (\gamma_1, \gamma_2)$ and a set of possible deviations from this equilibrium, $Q(\gamma = (Q_1, Q_2)$, where Q_i is a (possibly empty) set of mixed strategies for player i. We then construct the deviation correspondence, which is a set $D(\gamma) \subseteq \Delta(S_1) \times \Delta(S_2)$ containing those strategies that rational players might consider following a deviation from γ, by iteratively adding strategies that respond optimally to the deviations. We take the sets $Q(\gamma)$ as input into our framework—these sets are meant to directly incorporate intuitions of existing refinements. In this section, we shall illustrate our approach using Cho and Kreps's [7] *intuitive criterion*; in Section 4, we apply our approach to the theory of deviations proposed by Matthews *et al.* [17].

Starting from the set of possible deviations Q we form an increasing sequence of sets $\{\Sigma_1(n), \Sigma_2(n))\}$. The union of these sets will be the deviation correspondence. First we define $\Sigma_1(0)$ and $\Sigma_2(0)$:

$$\Sigma_i(0) \equiv \begin{cases} \gamma_i & \text{if} \quad Q_i = \varnothing \\ Q_i & \text{if} \quad Q_i \neq \varnothing. \end{cases}$$

If there are no allowable deviations, $\Sigma_i(0)$ is simply player i's equilibrium strategy; if there are allowable deviations, then $\Sigma_i(0)$ consists of these

deviations. This formulation allows that the equilibrium strategy γ_i need not be part of $\Sigma_i(0)$; the logic of the refinements that we study suggests to us that when an equilibrium is subject to deviations, then the equilibrium itself ought be eliminated from the sets $(\Sigma_1(0), \Sigma_2(0))$.

To illustrate our approach, consider again the sequential equilibrium in Fig. 1 in which both types of the Sender choose m_2 and the Receiver chooses a_2 given m_1 and r_1 given m_2. We denote this equilibrium by $((m_2m_2), (a_2r_1))$. Cho and Kreps [7] argue that t_1 would deviate from this equilibrium by sending message m_1 rather than m_2, her message specified by the equilibrium. Thus it is concluded that the equilibrium $((m_2m_2), (a_2r_1))$ is not stable.

Formally, we construct the sets $(\Sigma_1(0), \Sigma_2(0))$ based on the intuitive criterion as follows. Let $BR_2(K, m)$ denote the set of best responses by the Receiver to the message m if his beliefs are concentrated on the set of types $K \subseteq T$. For the equilibrium γ, let $u^*(t)$ be the payoff for type t of the Sender, and let M^* be the set of signals sent with probability zero. We say that a strategy is in Q_1 if it involves all types besides some type t^* sending their equilibrium signals, and involves t^* sending some message $m^* \in M^*$ such that $u(t^*, m^*, a) > u^*(t^*)$ for all $a \in BR_2(J, m^*)$, where $J \equiv \{t|$ there exists $a \in BR_2(T, m^*)$ such that $u(t, m^*, a) \geqslant u^*(t)\}$. Type t^* gains (relative to her equilibrium payoff) by using m^* if the Receiver responds to this message with an element of $BR_2(J, m^*)$. It is plausible to restrict the Receiver's best responses to m^* to be in $BR_2(J, m^*)$ since every type of Sender outside of J does better playing her equilibrium strategy than sending m^*, provided the Receiver responds to m^* by responding optimally to some conjecture over the Sender's types. The intuitive criterion rules out, an equilibrium precisely when Q_1 is non-empty. For the equilibrium $((m_2m_2), (a_2r_1))$, this definition says that $Q_1 = \{(m_1m_2)\}$, and thus that $\Sigma_1(0) = \{(m_1m_2)\}$.

We shall also define "deviations" by the Receiver, which correspond to the anticipated behavior by the Receiver that Cho and Kreps invoke in arguing that the Sender should deviate. In particular, we assume that a strategy is in Q_2 if it has the Receiver playing a best response to the message m^* for some $s \in Q_1$, and playing the original equilibrium strategy in response to all strategies $m \neq m^*$. Hence if Q_1 is empty, then $(\Sigma_1(0), \Sigma_2(0))$ is equal to the original equilibrium. Otherwise, Q_2 does not include the Receiver's equilibrium strategy, and does include all strategies involving off-the-equilibrium-path best responses that motivate the Sender to deviate. One way in which we assess the validity of the Stiglitz critique is to see if the Receiver's equilibrium strategy is an element of $\Sigma_2(n)$ (defined below) for some n. To summarize, if we apply our framework to the pooling equilibrium of Fig. 1, we see that $\Sigma_1(0) = Q_1 = \{(m_1m_2)\}$, and $\Sigma_2(0) = Q_2 = \{(a_1r_1)\}$.

We consider such sets $(\Sigma_1(0), \Sigma_2(0))$ to be only a preliminary hypothesis and note that these strategies might not be consistent with common knowledge of rationality. If rational players hypothesize a set of deviations $Q(\gamma)$ might occur from the equilibrium γ, then each player ought rationally to respond to such deviations. For instance, if the Receiver truly believes that the Sender is going to play strategy $(m_1 m_2)$ rather than $(m_2 m_2)$, then he should play the strategy $(a_1 r_2)$ rather than $(a_1 r_1)$. Consequently, we wish to add this as a possible strategy for the Receiver.

We wish to define a deviation correspondence based on $(\Sigma_1(0), \Sigma_2(0))$ as the subset of strategies in $(\Delta(S_1), \Delta(S_2))$ that players believe possible following the equilibrium γ. Define $\Sigma_i(n)$ iteratively by setting $\Sigma_i(n) \equiv \Sigma_i(n-1) \cup SBR_i(\Delta(\Sigma_{-i}(n-1)))$ if $\Sigma_{-i}(n-1) \neq \gamma_{-i}$ and by $\Sigma_i(n) = \gamma_i$ if $\Sigma_{-i}(n-1) = \gamma_{-i}$. That is, if we have added strategies to player $-i$, we then add all strong best responses by player i to all beliefs over the new strategies; if we have not added strategies to player $-i$, then we continue to assume that player i will play his equilibrium strategy.[7] Because each S_i is finite and $\Sigma_i(n) \subseteq \Sigma_i(n+1)$, we know there exists an n^* such that for all i and all k, $\Sigma_i(n^*) = \Sigma_i(n^* + k)$. Then let $\Sigma_i^* = \Sigma_i(n^*)$.

In the expansion process based on the intuitive criterion, $\Sigma_1 = \Sigma_1(0)$, but $\Sigma_2(1)$ adds in the strategy $(a_1 r_2)$, because perceiving a deviation will mean that the Receiver will change his response on the equilibrium path. Depending on the Sender's beliefs about the likelihood of the Receiver's two strategies in $\Sigma_2(1)$, therefore, the Sender might now prefer to play the strategy $(m_1 m_1)$. Thus, $\Sigma_1(2)$ contains $(m_1 m_1)$. In turn, this means that $\Sigma_2(3)$ contains $(a_2 r_1)$, which is optimal if the Sender chooses $(m_1 m_1)$, and that $\Sigma_1(4)$ contains $(m_2 m_2)$, so the original equilibrium $((m_2 m_2), (a_2 r_1))$ is contained in (Σ_1^*, Σ_2^*).

We refer to this process more generally as an *expansion*:

DEFINITION 1. For a given equilibrium $\gamma \equiv (\gamma_1 \gamma_2)$ and deviation sets $Q(\gamma) \equiv (Q_1, Q_2)$, let the *expansion* of $(\gamma, Q(\gamma))$, denoted $Exp(\gamma, Q(\gamma))$, be the sets of strategies (Σ_1^*, Σ_2^*) constructed as outlined above.

Throughout the paper, we shall equate the *deviation correspondence* $D(\gamma)$—which summarizes whether and how deviations from the equilibrium γ will occur—with the expansion of the $(\gamma, Q(\gamma))$. By iteratively adding in best responses, the expansion process guarantees that the sets of strategies the players might play are consistent with common knowledge of

[7] The results in this section do not change if we replace SBR by BR in the definition of $\Sigma_i(n)$. In Section 4 we describe a deviation correspondence where it is necessary to add in only strong best responses. For consistency we have chosen to use strong best response in this section as well.

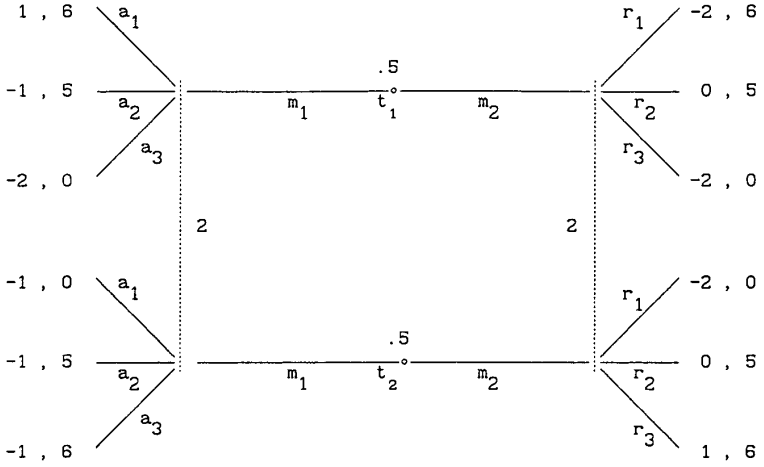

FIGURE 2

rationality.[8] An important feature of the expansion process is that it will always contain at least one Nash equilibrium. This fact follows once we note that the hypothetical game defined by permitting players to use only mixtures of strategies in $D(\gamma)$ (and in which they receive the payoffs of the real game) has a Nash equilibrium. That Nash equilibrium must also be a Nash equilibrium of the real game; otherwise the expansion process would not have stopped.[9]

Because the intuitive-criterion deviation correspondence constructed from the pooling equilibrium X in Fig. 1 contains X itself, the Stiglitz critique applies. Figure 2 demonstrates, however, that the Stiglitz critique does not always redeem unintuitive equilibria.

The pooling equilibrium $((m_2 m_2), (a_2 r_2))$ fails the intuitive criterion in this example because t_1 (and only t_1) could gain by deviating. Because this is the only deviation designated by the intuitive criterion, $\Sigma_1(0) = Q_1 = \{(m_1 m_2)\}$ and $\Sigma_2(0) = Q_2 = \{(a_1 r_2)\}$. In the iteration process, $(a_1 r_3)$ will be included in $\Sigma_2(1)$, but nothing more is added into either player's

[8] Because deviation correspondences are sets of strategies consistent with common knowledge of rationality without imposing the equilibrium assumption, our approach here is similar to that taken in [21, 24], where an approach to combining behavioral assumptions with rationalizability is developed. Game-theoretic applications using similar ideas include [6], [22, 23, 25], [28], and [30].

[9] The sets $(\Sigma_1(0), \Sigma_2(0))$ need not contain an equilibrium (they don't in our example for Fig. 1). It is necessary to expand the set of strategies in the deviation correspondence beyond $(\Sigma_1(0), \Sigma_2(0))$ in order to develop a theory which guarantees that it is possible for players to play an equilibrium following a deviation. This expansion process also guarantees that our theory of deviations is consistent with common knowledge of rationality.

best-response set. The deviation correspondence will therefore be (Σ_1^*, Σ_2^*), where $\Sigma_1^* = \{(m_1 m_2)\}$ and $\Sigma_2^* = \{(a_1 r_2), (a_1 r_3)\}$. These sets contain only the separating equilibrium to this game, $((m_1 m_2), (a_1 r_3))$, and not the original equilibrium. The unintuitive equilibrium does not therefore become plausible even if we assume common knowledge of the posited deviation, so the outcome here is not stable in any sense: Even when we take the Stiglitz critique fully into account, the pooling equilibrium does not survive a deviation.

We have not described the only possible way to formalize the Stiglitz critique. In fact, if one applies our approach to the Beer–Quiche example in which Cho and Kreps [7] discuss the Stiglitz critique, one finds that the bad equilibrium is not included in its deviation correspondence. Figure 3, which is closely related toil the Beer–Quiche example, clarifies the contrast between our approach and the original formulation of the Stiglitz critique.

Here, as in Fig. 1, there is a pooling sequential equilibrium, $(m_2 m_2, a_2 r_1)$; as in Fig. 1, this equilibrium is subject to a deviation in which t_1 plays m_1, and the Receiver responds to m_1 with a_1. It is easy to confirm that the deviation correspondence based on this deviation does not add back the equilibrium strategy for either player, because once the players realize that t_1 might deviate and play m_1, the Receiver would prefer to respond to m_1 with the strategy a_1 no matter what else he believed; the expansion process does not add back the original equilibrium.

Our formulation captures what the players might rationally choose in response if they come to believe the proposed deviation, and only the proposed deviation, is likely. The specific argument by Joseph Stiglitz

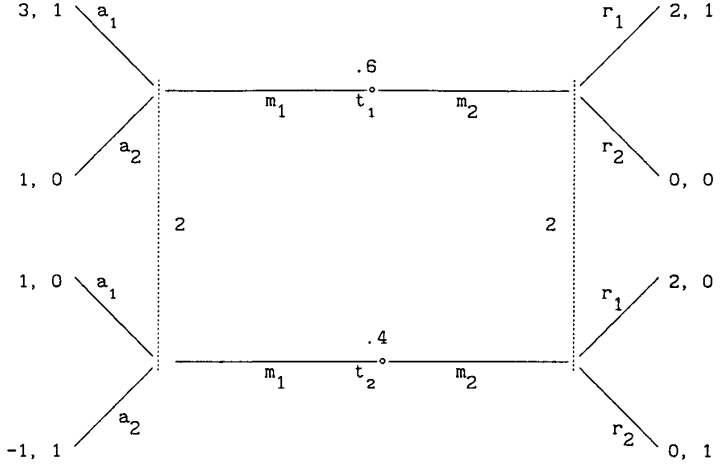

FIGURE 3

discussed by Cho and Kreps was somewhat different: He posited that the Receiver should not necessarily come to believe that a deviation by only type t_1 is likely, because if type t_2 came to believe that the Receiver would react differently on the equilibrium path because of a predicted deviation by t_1, then she too would be tempted by the deviation; if the Receiver for some reason came to believe such a deviation by t_2 was more likely than one by t_1, then he might reasonably respond to the deviation by a_2 rather than a_1. Thus, the original equilibrium might be played.

Stiglitz's original critique seems to be to propose that there is a larger set of deviations, which, if deemed just as likely as the one Cho and Kreps proposed, might lead the players back to their equilibrium strategies. Because we first imagine that the players come to believe in the proposed deviation and only then use our expansion process to make sure to include all possible rational responses to a posited deviation, our approach leads us more rarely to add back equilibrium strategies than if we used the specific arguments originally proposed by Stiglitz. We use the term Stiglitz critique for our approach because it leads us to conclude, as do the original arguments, that sometimes rational players might respond to a contemplated deviation by playing the original equilibrium.[10]

One could in any event modify our general approach to incorporate the alternative theory of deviations by including additional strategies in the deviation sets (Q_1, Q_2) and applying our expansion process to these modified deviation sets.

Whether one applies Stiglitz's Stiglitz critique or Rabin and Sobel's Stiglitz critique, our framework could be used to define a "Stiglitz-proof" version of the intuitive criterion or any other signaling refinement. Translating into our framework, the modified approach could be to eliminate an equilibrium γ only if $\gamma \notin D(\gamma)$. Of course, whenever $D(\gamma) \neq \{\gamma\}$, one might still say that an equilibrium γ is unstable in that players might deviate. In a sense, some of the debate over the merits of signaling refinements and the Stiglitz critique may be about which of these two notions of "stability" is appropriate—should we, as is currently done, build our solution concepts around equilibria from which surely there will be no deviation (i.e., where

[10] We note that even Stiglitz's original formulation does not argue against throwing out the unituitive equilibrium in Fig. 2. In that example, changes in the Receiver's behavior on the equilibrium path make the equilibrium path even more attractive for the non-deviating type, so that she will not contaminate the proposed deviation by the other type.

It is straightforward to show that in signaling games with two types of Sender, two signals, and two responses, if only pure strategies are used on the equilibrium path of every equilibrium, as in the Beer–Quiche example, then no equilibrium that fails the intuitive criterion will be an element of its deviation correspondence. This fact does not apply to Fig. 1, because the equilibrium Y involves some mixing. It is also straightforward to demonstrate that the result does not extend beyond two-type, two-signal, two-action games.

$D(\gamma) = \{\gamma\}$), or should be build our solution concepts based on equilibria from which there *might not* be deviations (i.e., where $\gamma \in D(\gamma)$)? In the next section, we argue that *neither* view takes a sufficiently dynamic view of deviations.

3. Dynamics (Sort of)

We now consider the dynamic implications for equilibrium selection of our framework. Of course, since deviations might not lead immediately to alternative equilibria, we would expect to observe non-equilibrium play frequently. We will discuss a possible non-equilibrium version of our framework in Section 5; here we focus exclusively on equilibrium outcomes. Partly we justify this by supposing that the rate of re-equilibration following a deviation is fast relative to the frequency of deviations. But we are also simply carrying over the equilibrium assumption from the solution concepts we are critiquing and modifying.

We define the set of Nash-equilibrium outcomes contained in a deviation correspondence $D(\cdot)$ as $ND(\cdot)$. We assume that each set $ND(\cdot)$ is finite.[11] While often there can be an infinite number of equilibria supporting any one equilibrium outcome, we shall sometimes abuse terminology and use "equilibrium" to mean an equilibrium outcome.

We suppose that after a deviation from an equilibrium γ play eventually equilibrates on some outcome $\gamma' \in ND(\gamma)$. A trivial class of equilibria that might persist are those that are stable in the sense that they are not susceptible at all to deviations. It is of course stable equilibria that have been emphasized in the refinement literature. Unless one wants to make arguments that such equilibria will not occur to begin with, a solution concept should clearly contain all equilibria that are not susceptible to deviations. Formally, an equilibrium is stable when $Q_i = \varnothing$ for each i. Because deviation correspondences are constructed using expansions, this is equivalent to the following definition:

Definition 2. An equilibrium γ is *stable* if $D(\gamma) = \{\gamma\}$.

Our framework does not not always rule out other equilibria. The simplest type of equilibrium ignored by the refinement literature that can persist in the long run is a *quasi-stable equilibrium*:

Definition 3. An equilibrium is *quasi-stable* if $ND(\gamma) = \{\gamma\}$.

[11] This assumption will be valid for generic signaling games (see [7, page 190, Fact 2]).

If a deviation from a quasi-stable equilibrium occurs, play must return to the original equilibrium when it next equilibrates.[12] Especially if we are unwilling to abandon exclusive focus on equilibrium outcomes, ruling out quasi-stable equilibria is inappropriate.

More generally, what can we say about the set of equilibria that might occur in the long run? To motivate our answer, we return again to the game in Fig. 1. While the partially separating equilibrium, Y, is stable with respect to the intuitive criterion, the pooling equilibrium X is not. Formally, we argued earlier that $ND(Y) = \{Y\}$ and $ND(X) = \{X, Y\}$. The equilibrium Y clearly might occur repeatedly. Because $X \in ND(X)$, arguably the equilibrium X might occur repeatedly as well. But because it is also true that $Y \in ND(X)$, a deviation from X "might" occur. And, because Y is stable, if it occurs, the equilibrium X will never again occur.

Based on the assumption that at any play of the game there might be a deviation from X to Y, and that once there is, play will never return to X, we propose to rule out X as a long-run equilibrium. More generally, let $B_1(\gamma) = ND(\gamma)$, $B_n(\gamma) = \{\gamma'' | \gamma'' \in ND(\gamma') \text{ for some } \gamma' \in B_{n-1}(\gamma)\}$ and $B^*(\gamma) = \bigcup_n B_n(\gamma)$. $B^*(\gamma)$ consists of those equilibria that can arise following a sequence of deviations from γ. If $\gamma' \in ND(\gamma)$ but $\gamma \notin B^*(\gamma')$, then once γ' occurs the equilibrium γ will never occur again. Ruling out the equilibrium γ will be justified if we make the following assumptions about the dynamics of repeated play. Suppose that the number of periods it takes for an equilibrium to occur following a deviation is finite; further suppose that there exists some $p > 0$ such that whenever γ occurs, any equilibrium in $ND(\gamma)$ occurs in the next equilibrated period with probability at least p.[13] We call a dynamic process that satisfies these assumptions nontrivial. When the dynamic is nontrivial, following the equilibrium γ each equilibrium in $B^*(\gamma)$, and only these equilibria, will arise with positive probability. It follows that as the number of periods approaches infinity $\gamma' \in ND(\gamma)$ and $\gamma \notin B^*(\gamma')$ imply that the probability of seeing the equilibrium γ converges to zero.

Under these conditions, the set of equilibria that appear with positive probability in the long run are described in the following definition.[14]

DEFINITION 4. A set of Nash equilibria $\mathcal{G}$ is an *absorbing set* if $ND(\gamma) \subseteq \mathcal{G}$ for all $\gamma \in \mathcal{G}$. The set $\mathcal{G}$ is *recurrent* if it is an absorbing set and contains no proper, non-empty subsets that are absorbing sets. An equilibrium γ is recurrent if it is contained in some recurrent set.

[12] The clearest examples of quasi-stable equilibria are when there is a unique equilibrium in a game, and a theory of deviations says that it is not stable. For such an example, see Farrell [10].

[13] It is here that we use our assumption that the number of equilibrium outcomes is finite. We can relax this assumption. (See [13].)

[14] [8], [12], and [13] investigate related ideas.

Since the set of all equilibria of the game is itself an absorbing set, and since any intersection of absorbing sets is also an absorbing set, we know that recurrent sets exist.

It is clear from Definition 4 that if play settles on an equilibrium in a recurrent set, then no equilibrium outside of the recurrent set will be observed in the future. Moreover, for each γ, $B^*(\gamma)$ is an absorbing set. It follows that $B^*(\gamma)$ contains a recurrent set for each γ. Once this happens, play will never return to γ unless γ itself is an element of a recurrent set. Hence, the probability that a non-recurrent equilibrium will be played infinitely often is zero. For this reason, we conclude that the definition of recurrence fully captures the set of equilibria that will persist in the long run.[15] To summarize:

PROPOSITION 1. *For all games, there exist recurrent equilibria. If the dynamic is nontrivial, an equilibrium occurs infinitely often with positive probability if and only if it is recurrent.*

Proposition 1 guarantees existence of recurrent equilibria with respect to even those theories of deviations that do not guarantee the existence of stable equilibria. We study such a theory of deviations in the next section. We conclude this section by applying our framework to the intuitive criterion. Proposition 2 demonstrates that the recurrent equilibria with respect to the intuitive deviation correspondence coincide with those that are stable according to the intuitive deviation correspondence (i.e., the set of equilibria that pass the intuitive criterion). In order to prove Proposition 2, we use the following simple result:

LEMMA 1. *If there exists a quasi-stable equilibrium $\gamma' \in ND(\gamma)$ such that $\gamma' \neq \gamma$, then γ is not recurrent.*

Lemma 1 follows directly from the definitions of quasi-stability and recurrence. Since any quasi-stable equilibrium is trivially recurrent, Lemma

[15] As we discuss in the text, our definition of recurrence sometimes rules out equilibria that "could" occur infinitely often, but which are very unlikely to appear in the very long run if there are lower bounds on the probability of deviations to all equilibria in a deviation correspondence. If either we wish to consider the not-so-long long run, or if we do not want to assume such a lower bound on deviation probabilities, a weaker definition of recurrent equilibria could be defined iteratively as follows. Let $E(O)$ be the set of all Nash equilibria in a game. For $n>0$, let $E(n) \equiv \bigcup_\gamma B_n(\gamma)$. $E(1)$ is the set of equilibria that are in the deviation correspondence of some other equilibrium. If an equilibrium is not in $E(1)$, it clearly cannot persist in the long run, because it is itself not robust to deviations, and would never be deviated to by another equilibrium. Likewise, any equilibrium not contained in $E(n)$ for some n will not persist in the long run. Any equilibrium which is contained in $E(n)$ for all n, on the other hand, might occur infinitely often. Under this definition, the equilibrium X in Fig. 1 would be deemed recurrent because it is contained in its own deviation correspondence.

1 means that if the deviation correspondence from every equilibrium contains a quasi-stable equilibrium, then an equilibrium is recurrent if and only if it is quasi-stable. Indeed, this case applies to the intuitive deviation correspondence:

PROPOSITION 2. *The set of equilibria that are recurrent with respect to the intuitive deviation correspondence are precisely the set of stable equilibria, which in turn is precisely the set of equilibria that passes the intuitive criterion.*

Proof. Applying Lemma 1, we need only show that, for all equilibria γ, there exists γ' surviving the intuitive criterion such that $\gamma' \in ND(\gamma)$. This result holds trivially if γ survives the intuitive criterion. Suppose that it does not hold for some γ failing the intuitive criterion. Let (Σ_1^*, Σ_2^*) be the expansion of γ. Now consider the hypothetical game in which the payoffs are the same as the game being examined, but in which only the messages and actions in (M, Q^*) are available to the players, where Q^* is the set of actions played with positive probability (on or off the equilibrium path) by some strategy in Σ_2^*. We know by the Cho and Kreps existence theorem that, with respect to this hypothetical game, there exists $\sigma \in (\Sigma_1^*, \Sigma_2^*)$ that passes the intuitive criterion.

To complete our proof, we claim that if an equilibrium survives the intuitive criterion in the hypothetical game, then it also survives the intuitive criterion in the actual game. Suppose not. Then there exists an unsent message, m^*, of the sort invoked in the intuitive criterion. To complete the proof we will show that the best responses to m^* must be contained in Q^*, which contradicts the assumption that σ survives the intuitive criterion in the hypothetical game. When σ fails the intuitive criterion, the definition of deviation sets requires that a strategy in which one type of Sender sends m^* must be an element of Q_1. Moreover, all best responses to this strategy must be in Q_2. Hence, since all strategies contained in Q must also be in (Σ_1^*, Σ_2^*), any deviation by the Sender that causes an equilibrium to fail the intuitive criterion, as well as a best response to that deviation, must be available in the hypothetical game.

The proof of Proposition 2 is direct. We show that for any equilibrium, γ, $D(\gamma)$ contains an equilibrium that passes the intuitive criterion. The result then follows from Lemma 1.

Proposition 2 defends the intuitive criterion against the Stiglitz critique in a dynamic framework. It implies that any equilibrium that fails the intuitive criterion will not be played infinitely often with positive probability.

4. Cheap Talk and Recurrent Mops

A major point of our framework is that there may be unstable equilibria that are nonetheless recurrent. Yet the deviation correspondence based on the intuitive criterion did not reflect this point; Proposition 2 indicated that only stable equilibria were recurrent. In this section, we present in the context of cheap-talk games a deviation correspondence where some unstable equilibria are recurrent.

Cheap-talk games are signaling games with the property that players' payoffs do not depend directly on the message sent by the Sender, so that the payoffs are representable by the utility functions $u_1(t, a)$ and $u_2(t, a)$. We shall construct deviation correspondences based on deviations that are *weakly credible* in the sense of Matthews *et al.* [17]. They partition the Sender's types into disjoint groups, where one group sticks with the equilibrium and each other group sends a different message. They require that each type t gets a higher payoff from the Receiver's optimal response to t's group than what t could get if imitated another group's message. Formally:

Definition 5. For all $t \in T$, let $u_1^*(t)$ be the payoffs type t gets in equilibrium γ. Call a subset J of T a *self-signaling family* relative to γ if there exists a partition of J into J_i, $i = 1, ..., j$, and actions a_i^* such that:

$$\text{(i)}; \quad a_i^* \in \arg\max_{a \in S_2} \sum_{t \in J_i} u_2(t, a)\, \pi(t) \text{ for } i = 1, ..., j;$$

$$\text{(ii)}; \quad u_1(t, a_k^*) > \max\{u_1^*(t), u_1(t, a_i^*)\} \qquad \text{for } t \in J_k \text{ and } i \neq k,$$

$$\text{(iii)}; \quad u_1(t, a_i^*) < u_1^*(t) \qquad \text{for all } i \text{ if } t \notin J.$$

While inspired by Matthews *et al.* [17], Definition 5 incorporates a more liberal notion of allowable deviations than do Matthews *et al.*[16] While we find Matthews *et al.*'s restrictions compelling if interpreted as a theory of which deviations necessarily occur in a given play of the game, our perspective leads us to consider the broader class of deviations that may occur in the long run.

We must modify the definition of deviation correspondences to take into account the special nature of cheap-talk games and the deviation that Matthews *et al.* permit. We follow Farrell [10] in assuming that in every equilibrium, and for every subset of types $X \subseteq T$, there exists

[16] Specifically, when there are multiple statements that might be believed, or if the Receiver has more than one optimal response to given beliefs, Matthews *et al.* not assume that he will choose the action that the relevant types of Sender prefer. We allow a deviation if some beliefs, and optimal response by the Receiver benefits the relevant types. See [17] and [22] for a discussion of this issue. We make a further restriction for simplicity that does not substantively change our results: While Matthews *et al.* present a definition valid for mixed-strategy deviations, we only allow for pure-strategy deviations.

	A	B	C	D	E	F	G
t_1	9, 10	6, 7	7, 9	4, 0	4, 0	4, 0	3, 6
t_2	6, 7	5, 10	7, 9	4, 0	4, 0	4, 0	3, 6
t_3	4, 0	4, 0	4, 0	9, 10	6, 7	7, 9	3, 6
t_4	4, 0	4, 0	4, 0	6, 7	5, 10	7, 9	3, 6

FIGURE 4

a "neologism"—an unused message $m(X)$ that means "I am some type $t \in X$." We assume that there are a finite number of messages used in equilibrium and a finite set of potential neologisms, $\{m(L)\}_{L \subseteq T}$. For a given equilibrium γ, Definition 6 can be used to construct sets (Q_1, Q_2). For every self-signaling family of types J, partition $\{J_1, ..., J_N\}$ and best responses a_i^* to J_i meeting the criteria of Definition 6, let σ_2 be the strategy that responds to the neologism $m(J_i)$ by action a_i^* and to every other message according-to the equilibrium strategy. Then let Q_2 be the set of all such σ_2 and let Q_1 be the set of optimal responses to beliefs over the set Q_2.

Once the set Q has been specified, we create the deviation correspondence by iteratively adding strong best responses, where the Sender's strategies are restricted to sending those messages that are used with positive probability in either the original equilibrium or in Q_1.[17]

We call the solution concept created by applying our framework to this deviation correspondence *recurrent mop*. The game in Fig. 4 illustrates some implications of recurrent mop, and how it differs from existing solution concepts.

Assuming that the four types of Sender are equally likely, there is a pooling equilibrium where all types of the Sender send the same messages and the Receiver always takes action G. The pooling outcome is not plausible according to the arguments of Matthews *et al.*: types t_1 and t_2 could improve their payoffs by jointly deviating with the message "I am either t_1 or t_2," leading to action C, and types t_3 and t_4 could jointly deviate by sending the message "I am either t_3 or t_4," leading to action F. For that reason, the pooling equilibrium is not stable with respect to the mop

[17] This restriction is necessitated by the structure of cheap-talk games, and is made consistent with rationality by assuming that the Receiver interprets any message as if it were one of the messages sent in the original equilibrium or in the set Q_1.

deviation correspondence. Moreover, the deviation correspondence will contain the only other sequential equilibrium outcome of the game.

This other equilibrium is a partially pooling outcome in which types t_1 and t_2 pool and induce the action C, while types t_3 and t_4 pool and induce action F. This equilibrium is also not stable with respect to the mop deviation correspondence. It is susceptible, for instance, to type t_1 deviating by self-signaling herself, inducing the action A rather than C. Likewise, type t_3 is tempted to self-signal herself to induce the action D.

Matthews *et al.*'s original solution concept is empty in this game. But the partially pooling equilibrium is a recurrent mop, because it is quasi-stable with respect to the mop deviation correspondence. That is, a deviation from the partially pooling equilibrium will never lead players back to the pooling equilibrium. Deviations from the partially pooling equilibrium involve the different types trying to separate themselves further, but nothing in the logic of responding to these deviations suggests that types $\{t_1, t_2\}$ would wish to pool with types $\{t_3, t_4\}$.[18]

We feel that recurrent mop makes the right prediction in this game—the subsets of types $\{t_1, t_2\}$ and $\{t_3, t_4\}$ will separate from each other. By using recurrence as the standard for prediction-making, we formulate a solution concept that captures that fact. By contrast, the non-existence of Matthews *et al.*'s *announcement-proof equilibrium* leaves the analyst agnostic about the possible outcomes in this game. We also note that the solution concept *neologism-proof equilibrium* developed by Farrell [10] predicts precisely the opposite conclusion—that only the pooling

[18] Formally, consider the partially pooling equilibrium in which types t_1 and t_2 send the message m_{12} and types t_3 and t_4 send the message m_{34}. The mop deviation set Q_2 will include responding to the message $m(t_1)$ with action A, and message $m(t_3)$ with action D, and responding to m_{12} and m_{34} with actions C and F. The set Q_1 will therefore contain only the strategy where t_1 sends $m(t_1)$, t_2 sends m_{12}, t_3 sends $m(t_3)$, and t_4 sends m_{34}. (Note here why it is important that we not include "weak" best responses during iteration; doing so would allow, for instance, the Receiver to respond to message $m(t_1)$ with action D because he thought the message would be sent with probability zero. Similarly, the Receiver could respond to previously unsent messages with D. That is why we do not allow the Sender to use messages that were not used with positive probability in either the original equilibrium or Q_1.) In the iterative process, the Receiver would now add in strategies where he responds to m_{12} with B instead of C, and to m_{34} with E instead of F; this response is optimal because he knows that t_1 and t_3 are not sending their equilibrium messages. Given this, we would add to the Sender's strategies the possibility that t_2 would in fact send the message $m(t_1)$ and t_4 would send $m(t_3)$. This, in turn, means that the Receiver might respond to $m(t_1)$ and $m(t_3)$ with actions C and F instead of B and E. Thus, the deviation correspondence will consist of strategies where types t_1 and t_2 send some combination of messages $m(t_3)$ and m_{34}. In particular, the deviation correspondence contains the original partially pooling equilibrium (using either the original messages, or the neologisms). We do not add the fully pooling equilibrium, and hence (since there are no other equilibrium outcomes) the partially pooling equilibrium is quasi-stable.

	A	B	C	D	E	F
t_1	0, -10	5, 4	4, 5	2, 6	1, 0	1, 0
t_2	4, 5	0, -10	5, 4	1, 0	2, 6	1, 0
t_3	5, 4	4, 5	0, -10	1, 0	1, 0	2, 7

FIGURE 5

equilibrium will occur. In our opinion, this is because Farrell is overly conservative in determining what constitutes credible deviations from an equilibrium, but is overly liberal in rejecting the partially pooling equilibrium because it is not stable.[19]

So far, all our examples of recurrent equilibria have been quasi-stable. In our final example, we illustrate that non-quasi-stable equilibria can be recurrent. Consider the game in Fig. 5.[20]

There are three equilibrium outcomes that involve partial pooling (for example, type t_1 induces the action D, while the other two types induce A), a separating equilibrium, and a pooling equilibrium in which the Receiver takes action F. It is straightforward to check that all equilibria permit a mop deviation. Starting from the outcome in which the Receiver only plays actions A and D with positive probability, there is a deviation that induces C. The deviation theory must then include the action F (as the Receiver allows the possibility that typed t_1 and t_2 will induce C and type t_3 will send another message). Consequently, the deviation correspondence must contain the outcome in which types t_1 and t_2 separate from the other type. In this way, one can verify that any minimal set for the game in Fig. 5 must

[19] While our framework can be used to guarantee existence, Fig. 4 illustrates why we don't feel it should be applied only when there is non-existence using other solution concepts. There are instances where existing refinements clearly are over-selective permitting equilibria even when existence is not a problem. In Fig. 4, for instance, a recurrent version of Farrell's [10] solution concept would conclude that both equilibria are possible, whereas Farrell allows only the fully pooling equilibrium. While Farrell identifies a sense in which the partially pooling equilibrium is prone to deviations, the deviation that he proposes involves the players separating *even more*—nothing in the logic of his notion of stability suggests that the players would deviate back to the pooling equilibrium. Even applying his own theory of deviations in this game, we feel Farrell's solution concept, by including only the pooling equilibrium, is too selective.

[20] This game is modeled after an example in [19], which is itself a variation of a game introduced by [18].

contain all three of the semi-pooling equilibria. While none of the equilibria is quasi-stable in this game, all four non-pooling equilibria are recurrent. Furthermore, one can show that the deviation correspondences from any of these equilibria do not contain the pooling equilibrium and that the deviation correspondence from the pooling equilibrium includes the other equilibria but not the equilibrium itself.[21]

We have no broad characterization theorem for recurrent mop, but we conclude with a result showing that, in a restrictive class of games, recurrent mop can guarantee a minimal degree of communication that other solution concepts do not. In particular, we can show that all recurrent mops involve meaningful communication in games of *partial common interests*, which are games where the Sender has a compelling interest to share with the Receiver some, but not necessarily all, of her private information.[22]

To define such games formally, we need to develop some initial definitions. For any non-empty subset of types L, let $A^*(L) \equiv \{\arg\max \sum_{t \in L} u_2(t, a)\, \mu(t) \mid \mu$ is a probability distribution supported on $L\}$ and $A^*(L, \pi) \equiv \{\arg\max \sum_{t \in L} u_2(t, a)\, \pi(t)\}$. The set $A^*(L)$ contains all of the responses that an optimizing Receiver would consider assuming that the Sender's type is an element of L, while $A^*(L, \pi)$ requires the Receiver to derive the relative probabilities of the types in L from the prior π. Also let $\underline{u}(t; L) \equiv \min\{u_1(t, a): a \in A^*(L)\}$ denote the lowest payoff that type t would obtain if the Receiver believed that her type was in L, and let $u_1^p(t)$ denote the maximum payoff that type t can obtain in a completely pooling equilibrium. (For cheap-talk games with generic payoffs, there will only be one pooling equilibrium payoff.)

DEFINITION 6. A cheap-talk game has partial common interests if there exists a partition $J_1, ..., J_j$ of T such that:

[21] Formally, consider the pooling equilibrium, where all types send the message m_{123}, and the Receiver responds with F. The deviations from this equilibrium are for t_1 and t_2 to send m_{12}, t_1 and t_3 to send m_{13}, and t_2 and t_3 to send m_{23}, and for the Receiver to respond to these with actions C, B, and A, respectively. Given the Receiver's responses, type t_1 strictly prefers message m_{13}, t_2 strictly prefers m_{12}, and t_3 strictly prefers m_{23}, so that the expansion includes strategies where the Receiver responds to m_{12} with C or E, to m_{13} with B or D, and to m_{23} with A or F. Given such stratgies for the Receiver, t_1 might send either m_{13} or m_{23}. Given these strategies, the Receiver's stratgies would be to respond to m_{12} with C, E, or D, to m_{13} with B, D, or F, and to m_{23} with A, F, or E. At this point no further strategies would be added back. The set of strategies does not involve the pooling equilibrium, because for all beliefs by the Sender, both t_1 and t_2 will strictly prefer sending m_{12} to sending m_{123}.

[22] An example of a game of partial common interests is Example 5 of [22], which Rabin gives as an example where, intuitively, one would want to guarantee meaningful communication, but where the solution concept he develops does not so. Propertion 3 shows that recurrent mop does guarantee communication in that and related examples.

(i) $\underline{u}_1(t_i; J_i) > \max\{u_1(t_i, a_k) : a_k \in A^*(J_k)\}$ for all $t_i \in J_i \neq J_k$;

(ii) for each i, there exists $a_i \in A^*(J_i, \pi)$ such that $u_1(t_i, a_i) > u_1^P(t_i)$ for all $t_i \in J_i$; and

(iii) if $L \cap J_k \neq \varnothing$ for at least two k, then for each $a \in A^*(L)$ there exists an i and $t_i \in L \cap J_i$ such that $u_1(t_i; J_i) > u_1(t_i, a)$.

Definition 6 is meant to capture the intuition that it is in the interest of both players for types in sets J_i to reveal at least that they are in sets J_i Condition (i) is a strong condition that guarantees that types in J_i prefer to identify themselves as members of J_i rather than as members of any other element of the partition. Condition (ii) states that each type would prefer to identify herself as a member of the partition that contains her type rather than be pooled.[23] Condition (iii) requires that, relative to what is available by being treated as a member of the set it belongs to, at least one type loses when members of different elements of the partition pool. We use this condition to show that once the population arrives at a strategy that reveals the partition J, it will never move to a less informative strategy profile. This condition follows from (i) whenever $A = \bigcup_{i=1}^{j} BR_2(J_i)$. This fact in turn allows us to establish the following result:

PROPOSITION 3. *No pooling equilibrium is a recurrent mop in a game with partial common interests.*

Proof. We prove the proposition in two steps. First we show that in a game with partial common interests the deviation correspondence beginning from a pooling equilibrium must contain a partially revealing equilibrium in which types in different sets J_i send different messages. Second we show that the deviation correspondence that starts from an equilibrium in which types in different J_i send different messages contains only equilibria with the same property. Formally, let $\{J_i\}$, $i = 1, ..., j$ be the partition in Definition 8, let γ^P be a pooling equilibrium and let $B = \{(\gamma_2, \gamma_1) | (\gamma_2, \gamma_1)$ is a pure-strategy Nash equilibrium and $\gamma_1(m, t_i) \neq \gamma_1(m, t_k)$ whenever $t_i \in J_i$ and $t_k \in J_k$ for $i \neq k\}$. Step 1 demonstrates that $ND(\gamma^P) \cap B \neq \varnothing$, while Step 2 demonstrates that $ND(B) \subseteq B$. It follows from the definitions that there is no recurrent mop containing γ^P. It follows from Definition 8 that B is non-empty (it must contain a partially pooling outcome in which types in the same J_i send the same message, but separate from types in other J_i); to support the equilibrium, select actions $a_i \in A^*(J_i, \pi)$ for $i = 1, ..., j$. Specify that the Receiver responds to each message with one of these actions and that for each i there exists a message to which the Receiver takes the action a_i. It is apparent that the deviation

[23] Conditions (i) and (ii) therefore together guarantee that J is a self-signaling family of sets relative to any pooling equilibrium.

correspondence starting from γ^P contains this outcome: Q_2 and Q_1 add the appropriate equilibrium strategies.

It remains to show that $ND(B) \subseteq B$. Start with any equilibrium in B. It follows from (iii) of Definition 6 that any self-signaling family that exists relative to an equilibrium in B must be a subpartition of $\{J_i\}$. It follows from (i) and the full-support assumption on conjectures that the expansion process only admits actions contained in $A^*(L)$ for $L \subseteq J_i$ for some i. Therefore, if type $t_i \in J_i$ sends message m_i with positive probability under $\gamma \in B$, then every strategy for the Receiver that is in $D(\gamma)$ must respond to m_i with an element of $A^*(J_i)$. It follows from (iii) that no element of $ND(\gamma)$ can pool types from different elements of the partition. This completes the proof.

Proposition 3 states that in games with partial common interests, players will not babble uninformatively forever. Blume *et al.* [5] obtain a related result in their study of evolutionary stability in cheap-talk games, but to our knowledge, the result holds for no other cheap-talk solution concept.[24] Along with the examples of this section, Proposition 3 demonstrates that we can construct solution concepts within our framework that are both more powerful and more realistic than existing concepts.

5. Discussion

We conclude by discussing two further areas of game-theoretic research to which the general principles outlined in this paper could be applied.

Bernheim *et al.* [4] consider the frequent supposition in game theory that, in communication-rich environments, only Pareto-efficient equilibria will be played. They follow [1] in pointing out a problem with this hypothesis in multi-person games: A Pareto-efficient Nash equilibrium may be susceptible to a coalition of players renegotiating their behavior so as to yield them all higher payoffs, given that other players continue to play their equilibrium strategies. While this will leave some other players worse off, such renegotiation may be likely if there is opportunity for private communication.

To deal with this issue, Aumann [1] proposed the solution concept of *strong Nash equilibrium*, which rules out all Nash equilibria that are susceptible to any such deviating coalition. Bernheim *et al.* argue that strong Nash equilibrium rules out too many equilibria, because it applies no test as to whether the outcome negotiated by the deviating coalition is itself susceptible to renegotiation. They define a solution concept,

[24] Sanchirico [27] suggests that a related result might be possible in his dynamic learning model.

coalition-proof Nash equilibrium, in which equilibria are ruled out only if there exists some beneficial coalitional renegotiation that is itself free from further coalitional renegotiation.

Our model does not directly apply to this issue, partly because we have not formalized it for multi-player games, but mostly because it is hard to conceptualize common knowledge, etc., without an explicit model of how players communicate. Yet our framework suggests that both the strong Nash equilibrium and the coalition-proof Nash equilibrium may be misleading in focusing too much on whether an equilibrium is "stable," rather on whether it is "recurrent."

While Bernheim *et al.* may be correct in suggesting that many of the renegotiations allowed by strong Nash equilibrium need not lead to stable behavior, we question their inference that such renegotiations are unlikely. In our framework, we would instead allow all such renegotiations and consider their long-run implications. Doing so, we could obtain an equilibrium concept that incorporates intuitive notions of when equilibria are subject to renegotiation but, unlike Aumann and Bernheim *et al.* guarantees existence.[25]

While we have focused on equilibrium outcomes throughout this paper, the equilibrium hypothesis itself has come under attack in recent years (see, e.g., [3] and [20]). We wish to conclude by discussing how our framework could be useful both in exploring the foundations of equilibrium analysis and in developing useful non-equilibrium theories.

Suppose that we defined deviation correspondences with respect to non-equilibrium outcomes as well as equilibrium outcomes, and said that a set of outcomes $\mathcal{O}$ is recurrent if it is minimal with respect to the property $D(\gamma) \subseteq \mathcal{O}$ for all $\gamma \in \mathcal{O}$. Recurrent outcomes would be those outcomes that are contained in some recurrent set. While this definition maintains the hypothesis that play eventually converges to a recurrent set, it no longer leads us automatically to focus only on equilibrium outcomes.

Of course, because stable equilibria are themselves recurrent sets, they would still be natural candidates as outcomes upon which play will settle. In fact, if it turned out that every deviation correspondence (even those generated from non-equilibrium outcomes) constructed from the intuitive criterion contains a stable equilibrium, then our framework would suggest that only stable equilibria will occur in the long run, even if we do not *a priori* focus only on equilibria. Thus, Cho and Kreps's theory of deviations would provide a (partial) justification not only for selecting their equilibria among all equilibria, but also among all outcomes.

As we showed in Section 4, however, there are games where no equilibrium is stable according to the mop deviation correspondence. Thus, the

[25] Chwe [8] suggests a related approach.

implied solution concept "recurrent mop rationalizability" would include non-equilibrium as well as equilibrium outcomes. It would nonetheless have some predictive power. In Example 3, for instance, although it would not uniquely predict the partially pooling equilibrium, recurrent mop rationalizability would include only outcomes in which there is partial separation of the differing types.

The above conjectures all suggest that a variant of our framework can be used not only to investigate which equilibria will occur in the long run, as we have done in this paper, but also to investigate whether play will converge to equilibrium and, if play does not necessarily converge, which non-equilibrium outcomes might recur infinitely often.

REFERENCES

1. R. AUMANN, "Acceptable Points in General Cooperative *N*-Person Games," Contributions to the Theory of Games IV, Princeton Univ. Press, Princeton, NJ, 1959.
2. J. BANKS AND J. SOBEL, Equilibrium selection in signaling games, *Econometrica* **55** (1987), 647–662.
3. B. D. BERNHEIM, Rationalizable strategic behavior, *Econometrica* **52** (1984), 1007–1028.
4. B. D. BERNHEIM, B. PELEG, AND M. WHINSTON, Coalition-proof Nash equilibria I: Concepts, *J. Econ. Theory* **42** (1987), 1–12.
5. A. BLUME. Y.-G. KIM, AND J. SOBEL, Evolutionary stability in games of communication, *Games Econ. Behav.* **5** (1993), 547–575.
6. I.-K. CHO, "Stationarity, Rationalizability and Bargaining," Department of Economics Working Paper 92-127, University of Chicago, May 1992.
7. I.-K.-CHO AND D. M. KREPS, Signaling games and stable equilibria, *Quart. J. Econ.* **102** (1987), 179–221.
8. M. CHWE, Farsighted coalitional stability, *J. Econ. Theory* **63** (1994), 299–325.
9. V. CRAWFORD AND J. SOBEL, Strategic information transmission, *Econometrica* **50** (1982), 1431–1451.
10. J. FARRELL, Meaning and credibility in cheap talk games, *Games Econ. Behav.* **5** (1993), 514–531.
11. S. GROSSMAN AND M. PERRY, Perfect sequential equilibrium, *J. Econ. Theory* **39** (1986), 97–119.
12. E. KALAI, A. PAZNER, AND D. SCHMEIDLER, Collective choice correspondences as admissible outcomes of social bargaining processes, *Econometrica* **44** (1976), 233–240.
13. E. KALAI AND D. SCHMEIDLER, An admissible set occurring in various bargaining situations, *J. Econ. Theory* **14** (1977), 402–411.
14. M. KANDORI, G. MAILATH, AND R. ROB, Learning, mutation, and long run equilibria in games, *Econometrica* **61** (1993), 29–56.
15. E. KOHLBERG AND J. F. MERTENS, On the strategic stability of equilibria, *Econometrica* **54** (1986), 1003–1038.
16. G. MAILATH, "A Reformulation of a Criticism of the Intuitive Criterion and Forward Induction," manuscript, University of Pennsylvania, 1988.
17. S. MATTHEWS, M. OKUNO-FUJIWARA, AND A. POSTLEWAITE, Refining cheap-talk equilibria, *J. Econ. Theory* **55** (1991), 247–273.
18. H. MOULIN AND J.-P. VIAL, Strategically zero-sum games: The class whose completely mixed equilibria cannot be improved upon, *Int. J. Game Theory* **7** (1978), 201–221.

19. R. Myerson, Credible negotiation statements and coherent plans, *J. Econ. Theory* **48** (1989), 264–291.
20. D. Pearce, Rationalizable strategic behavior and the problem of perfection, *Econometrica* **52** (1984), 1029–1050.
21. M. Rabin, "Predictions and Solution Concepts In Non-Cooperative Games," Ph.D. thesis, MIT, June 1989.
22. M. Rabin, Communication between rational agents, *J. Econ. Theory* **51** (1990), 144–170.
23. M. Rabin, "Focal Points in Pre-Game Communication," University of California—Berkeley Working Paper 91-179, September 1991.
24. M. Rabin, Incorporating behavioral assumptions into game theory, *in* "Problems of Coordination in Economic Activity" (James Friedman, Ed.), pp. 69–87, Kluwer Academic, Boston 1994.
25. M. Rabin, A model of pre-game communication, *J. Econ Theory* **63** (1994), 370–391.
26. M. Rabin and J. Sobel, "Deviations, Dynamics, and Equilibrium Refinements," University of California—Berkeley Working Paper 93-211, May 1993.
27. C. Sanchirico, "Strategic Intent and the Salience of Past Play: A Probabilistic Model of Learning in Games," manuscript, Yale University, 1993.
28. J. Watson, A "reputation" refinement without equilibrium, *Econometrica* **61** (1993), 199–205.
29. P. Young, The evolution of conventions, *Econometrica* **61** (1993), 57–84.
30. I. Zapater, "Credible Proposals in Communication Games," mimeo, Brown University, 1993.