

**Abstract:** This entry describes ways that the definition of an equilibrium among players' strategies in a game can be sharpened by invoking additional criteria derived from decision theory. Refinements of John Nash's 1950 definition aim primarily to distinguish equilibria in which implicit commitments are credible due to incentives. One group of refinements requires sequential rationality as the game progresses. Another ensures credibility by considering perturbed games in which every contingency occurs with positive probability, which has the further advantage of excluding weakly dominated strategies.

## Refinements of Nash equilibrium

Game theory studies decisions by several persons in situations with significant interactions. Compared to other theories of multi-person decisions, it has two distinguishing features. One is explicit consideration of each person's available strategies and the outcomes resulting from combinations of their choices; that is, a complete and detailed specification of the 'game.' In noncooperative contexts, the other is a focus on optimal choices by each person separately. John Nash (1950, 1951) proposed that a combination of mutually optimal strategies can be characterized mathematically as an *equilibrium*. According to Nash's definition, a combination is an equilibrium if each person's choice is an optimal response to others' choices. His definition assumes that a choice is optimal if it maximizes the person's expected utility of outcomes, conditional on knowing or correctly anticipating the choices of others. In some applications, knowledge of others' choices might stem from prior agreement or communication, or accurate prediction of others' choices might derive from 'common knowledge' of strategies and outcomes and of optimizing behavior. Because many games have multiple equilibria, the predictions obtained are incomplete. However, equilibrium is a weak criterion in some respects, and therefore one can refine the criterion to obtain sharper predictions (Harsanyi and Selten, 1988; Hillas and Kohlberg, 2002; Kohlberg, 1990; Kreps, 1990).

Here we describe the main refinements of Nash equilibrium used in the social sciences. Refinements were developed incrementally, often relying on *ad hoc* criteria, which makes it difficult for a non-specialist to appreciate what has been accomplished. Many refinements have been proposed but we describe only the most prominent ones. First we describe briefly those refinements that select equilibria with simple features, and then focus mainly on those that invoke basic principles adapted from single-person decision theory.

### ***Equilibria with simple features***

Nash's construction allows each person to choose randomly among his strategies but randomization is not always plausible. The equilibria in 'pure' strategies are those that do not use randomization. Similarly, strict equilibria are those for which each person has a unique optimal strategy in response to others' strategies. In games with some symmetries

among the players, the symmetric equilibria are those that reflect these symmetries. In applications to dynamic interactions the most useful equilibria are those that, at each stage, depend only on that portion of prior history that is relevant for outcomes in the future. In particular, when the dynamics of the game are stationary one selects equilibria that are stationary, or that are Markovian in that they depend only on state variables that summarize the history relevant for the future. Applications to computer science select equilibria, or more often approximate equilibria, using strategies that can be implemented by simple algorithms. Particularly useful are equilibria that rely only on limited recall of past events and actions and thus economize on memory or computation.

### ***Refinements that require strategies to be admissible***

One strategy is strictly dominated by another if it yields strictly inferior outcomes for that person regardless of others' choices. Because an equilibrium never uses a strictly dominated strategy, the same equilibria persist when strictly dominated strategies are deleted, but after deletion it can be that some remaining strategies become strictly dominated. A criterion that exploits this feature deletes strictly dominated strategies until none remain, and then selects those equilibria that remain in the reduced game. If a single equilibrium survives then the game is called *dominance solvable*. An equilibrium can, however, use a strategy that is weakly dominated in that it would be strictly dominated were it not for ties—in decision theory such a strategy is said to be inadmissible. A prominent criterion selects equilibria that use only admissible strategies, and sometimes this is strengthened by iterative deletion of strictly dominated strategies after deleting the inadmissible strategies. A stronger refinement uses *iterative deletion of* (both strictly and weakly) *dominated strategies* until none remain; however, this procedure is ambiguous because the end result can depend on the order in which weakly dominated strategies are deleted.

A particular order is used for dynamic games that decompose into a succession of subgames as time progresses. In this case, those strategies that are weakly dominated because they are strictly dominated in final subgames are deleted first, then those in penultimate subgames, etc. In games with 'perfect information' as defined below this procedure implements the criterion called **backward induction** and the equilibria that survive are among those that are *subgame-perfect* (Selten, 1965). In general a subgame-perfect equilibrium is one that induces an equilibrium in each subgame. The informal criterion of **forward induction** has several formulations. Kohlberg and Mertens (1986) require that a refined set of equilibria contains a subset that survives deletion of strategies that are not optimal responses at any equilibrium in the set. Van Damme (1989, 1991) requires that if player 1 rejects a choice A in favor of B or C then another player who knows only that B or C was chosen should consider C unlikely if it is chosen only in equilibria that yield player 1 outcomes worse than choosing A, whereas B is chosen in an equilibrium whose outcome is better. A typical application mimics backward induction but in reverse—if a person previously rejected a choice with an outcome that would have been superior compared to the outcomes from all but one equilibrium of the ensuing subgame, then presumably the person is anticipating that favorable equilibrium and intends to use his strategy in that equilibrium of the subgame.

## ***Dynamic games***

Before proceeding further we describe briefly some relevant features of dynamic games; that is, games in which a player acts repeatedly, and can draw inferences about others' strategies, preferences, or private information as the game progresses. A dynamic game is said to have 'perfect information' if each person knows initially all the data of the game, and the prior history of his and others' actions whenever he takes an action, and they do not act simultaneously. In such a game each action initiates a subgame; hence backward induction yields a unique subgame-perfect equilibrium if there are no ties. But in many dynamic games there are no subgames. This is so whenever some person acts without knowing all data of the game relevant for the future. The source of this deficiency is typically that some participant has private information; e.g., his own preferences or information about outcomes, or because his actions are observed imperfectly by some others. Among parlor games, chess is a game with perfect information (if players remember whether each king has been castled). Bridge and poker are games with imperfect information because the cards in one player's hand are not known to others when they bet. In practical settings, auctions and negotiations resemble poker because each party acts (bids, offers, etc.) without knowing others' valuations of the transaction. Analyses of practical economic games usually assume (as we do here) 'perfect recall' in the sense that each player always remembers what he knew and did previously. If bridge is treated as a two-player game between teams then it has imperfect recall because each team alternately remembers and then forgets the cards in one member's hand as the bidding goes round the table, but bridge has perfect recall if it is treated as a four-player game. In card games like bridge and poker each player can derive the probability distribution of others' cards from the assumption that the deck of cards was thoroughly shuffled. Models of economic games impose analogous assumptions; e.g., a model of an auction assumes that each bidder initially assesses a probability distribution of others' valuations of the item for sale, and then updates this assessment as he observes their bids. More realism is obtained from more complicated scenarios; e.g., it could be that player A is uncertain about player B's assessment of player A's valuation. In principle the model could allow a hierarchy of beliefs—A's probability assessment of B's assessment of A's assessment of ... . Adopting a proposal by John Harsanyi (1968) developed by Mertens and Zamir (1985), such situations are modeled by assuming that each player is one of several types. The initial joint distribution of types is commonly known among the players, but each player knows his own type, which includes a specification of his available strategies, his preferences over outcomes, and most importantly, his assessment of the conditional probabilities of others' types given his own type. In poker, for instance, a player's type includes the hand of cards he is dealt, and his hand affects his beliefs about others' hands.

Refinements of Nash equilibrium are especially useful in dynamic games. Nash equilibria do not distinguish between the case in which each player commits initially and irrevocably to his strategy throughout the game, and the case in which a player continually re-optimizes as the game progresses. The distinction is lost because the definition of Nash equilibrium presumes that players will surely adhere to their strategies chosen initially. Most refinements of Nash equilibrium are intended to resurrect this important distinction. Ideally one would like each Nash equilibrium to bear a label telling

whether it assumes implicit commitment or relies on incredible threats or promises. Such features are usually evident in the equilibria of trivially simple games, but in more complicated games they must be identified augmenting the definition of Nash equilibrium with additional criteria.

In the sequel we describe two classes of refinements in detail, but first we summarize their main features, identify the main selection criteria they use, and mention the names of some specific refinements. Both classes are generalizations of backward induction and subgame-perfection, and they obtain similar results, but their motivation and implementation differ.

**(1) The criterion of sequential rationality.** The presumption that commitment is irrevocable is flawed if commitment to a strategy is not viewed as credible by other participants in the game. Commitment can be advantageous, of course, but if commitment is possible (e.g., via enforceable contractual arrangements) then it should properly be treated as a distinct strategy. Absent commitment, some Nash equilibria are suspect because they rely implicitly on promises or threats that are not credible. For example, one Nash equilibrium might enable an incumbent firm to deter another firm from entering its market by threatening a price war. If such a threat succeeds in deterring entry then it is costless to the incumbent because it is never challenged; indeed, it can be that this equilibrium is sustained only by the presumption that the incumbent will never need to carry out the threat. But this threat is not credible if the incumbent would recognize after entry occurs that accommodation is more profitable than a price war. In such contexts, the purpose of a refinement is to select an alternative Nash equilibrium that anticipates correctly that entry will be followed by accommodation.

Refinements in the first class exclude strategies that are not credible by requiring explicitly that a strategy is optimal in each contingency, even if it comes as a surprise. (We use the term ‘contingency’ rather than the technical term ‘information set’ used in game theory—it refers to any situation in which the player chooses an action.) These generally require that a player’s strategy is optimal initially (as in the case of commitment), and also that in each subsequent contingency in which the player might act his strategy remains optimal for the remainder of the game, even if the equilibrium predicts that the contingency should not occur. This criterion is called **sequential rationality**. As described later, three such refinements are *perfect-Bayes*, *sequential*, and *lexicographic* equilibria, each of which can be strengthened further by imposing additional criteria such as **invariance**, the **intuitive criterion** and **divinity**.

**(2) The criterion of perfection or stability.** The presumption that commitment is irrevocable is also flawed if there is some chance of deviations. If a player might ‘tremble’ or err in carrying out his intended strategy—or his valuation of outcomes might be slightly different than others anticipated—then other players can be surprised to find themselves in unexpected situations. Refinements that exploit this feature are implemented in two stages. In the first stage one identifies the Nash equilibria of a perturbation of the original game, usually obtained by restricting each player to randomized strategies that assign positive probabilities to all his original pure strategies.

In the second stage one identifies those equilibria of the original game that are limits of equilibria of the perturbed game as this restriction is relaxed to allow inferior strategies to have zero probabilities.

Refinements in the second class also exclude strategies that are not credible, but refinements in this class implement sequential rationality indirectly. The general criterion that is invoked is called **perfection** or **stability** depending on the context. In each case a refinement is obtained from analyses of perturbed games. This second class of refinements is typically more restrictive than the first class due to the stronger effects of perturbations. As described later, two such refinements are *perfect* and *proper* equilibria. These are equilibria that are perturbed slightly by some perturbation of the players' strategies. A more stringent refinement selects a subset of equilibria that is *truly-perfect* or *stable* in the sense that it is perturbed only slightly by every perturbation of players' strategies. This refinement selects a subset of equilibria rather than a single equilibrium because there need not exist a single equilibrium that is **essential** in that it is perturbed slightly by every perturbation of strategies. A stringent refinement selects a subset that is *hyperstable* in that it is stable against perturbations of both players' strategies and their valuations of outcomes, or against perturbations of their optimal responses; and further, it is **invariant** in that it is unaffected by addition or deletion of redundant strategies.

The crucial role of perturbations in the second class of refinements make them more difficult for non-specialists to understand and appreciate, but they have a prominent role in game theory because of their desirable properties. For example, in a two-player game a perfect equilibrium is equivalent to an equilibrium that uses only admissible strategies. In general, refinements in the second class have the advantage that they satisfy several selection criteria simultaneously.

After this overview, we now turn to detailed descriptions of the various refinements.

### ***Refinements that require sequential rationality***

In dynamic games with perfect information, the implementation of backward induction is unambiguous because in each contingency the player taking an action there knows exactly the subgame that follows. In chess, for example, the current positions of the pieces determine how the game can evolve subsequently. Moreover, if he anticipates his opponent's strategy then he can predict how the opponent will respond to each possible continuation of his own strategy. Using this prediction he can choose an optimal strategy for the remainder of the game by applying the **principle of optimality**—his optimal strategy in the current subgame consists of his initial action that, when followed by his optimal strategies in subsequent subgames, yields his best outcome. Thus in principle (although not in practice, since chess is too complicated), his optimal strategy can be found by working backward from final positions through all possible positions in the game.

In contrast, in a game with imperfect information a player's current information may be insufficient to identify the prior history that led to this situation, and therefore insufficient to identify how others will respond in the future, even if he anticipates their strategies. In

poker, for example, knowledge of his own cards and anticipation of others' strategies are insufficient to predict how they will respond to his bets. Their strategies specify how they will respond conditional on their cards but, since he does not know their cards, he remains uncertain what bets they will make in response to his bets. In this case, it is his assessment of the probability distribution of their cards that enables construction of his optimal strategy. That is, this probability distribution can be combined with their strategies to provide him with a probabilistic prediction of how they will bet in response to each bet he might make. Using this prediction he can again apply the principle of optimality to construct an optimal strategy by working backward from the various possible conclusions of the game.

Those refinements that select equilibria satisfying sequential rationality use an analogous procedure. The analog of the probability distribution of others' cards is a system of 'beliefs,' one for each contingency in which the player might find himself. Each belief is a conditional probability distribution on the prior history of the game given the contingency at which he has arrived. Thus, to whatever extent he is currently uncertain about others' preferences over final outcomes, or their prior actions, then his current belief provides him with a probability distribution over the various possibilities. As in poker, this probability distribution can be combined with his anticipation of their strategies to provide him with a probabilistic prediction of how they will act in response to each action he might take—and again, using this prediction he can apply the principle of optimality to construct an optimal strategy by working backward from the various possible conclusions of the game.

There is an important proviso, however. These refinements require that whenever one contingency follows another with positive probability then the belief at the later one must be obtained from the belief at the earlier one by Bayes' rule. This ensures consistency with the rules of conditional probability. But importantly, it does not restrict a player's belief at a contingency that was unexpected; i.e., had zero probability according to his previous belief and the other players' strategies.

The weakest refinement selects a *perfect-Bayes* equilibrium (Fudenberg and Tirole, 1991). This requires that each player's strategy is consistent with some system of beliefs such that (1) his strategy is optimal given his beliefs and others' strategies, and (2) his beliefs satisfy Bayes' rule (wherever it applies) given others' strategies. A stronger refinement selects *sequential* equilibria (Kreps and Wilson, 1982). A sequential equilibrium requires that each player's system of beliefs is consistent with the structure of the game. Consistency is defined formally as the requirement that each player's system of beliefs is the limit of the conditional probabilities induced by players' strategies in some perturbed game, as described previously. A further refinement selects *quasi-perfect* equilibria (van Damme, 1984), which requires admissibility of a player's strategy in continuation from each contingency, excluding any chance that he himself might deviate from his intended strategy. And even stronger are *proper* equilibria (Myerson, 1978) described later.

This sequence of progressively stronger refinements is typical. Because proper implies quasi-perfect implies sequential implies perfect-Bayes, one might think that it is sufficient to always use properness as the refinement. However, the prevailing practice in the social sciences is to invoke the weakest refinement that suffices for the game being studied. This reflects a conservative attitude about using unnecessarily restrictive refinements. If, say, there is a unique sequential equilibrium that uses only admissible strategies, then one refrains from imposing stronger criteria.

These refinements can be supplemented with additional criteria that restrict a player's beliefs in unexpected contingencies. The most widely used criteria apply to contexts in which one player A could interpret the action of another player B as revealing private information; that is, B's action might signal something about B's type. These criteria restrict A's belief (after A observes B deviating from the equilibrium) to one that assigns positive probability only to B's types that might possibly gain from the deviation, provided it were interpreted by A as a credible signal about B's type. The purpose of these criteria is to exclude beliefs that are blind to B's attempts to signal what his type is when it would be to B's advantage for A to recognize the signal. In effect, these criteria reject equilibria that commit a player to unrealistic beliefs. Another interpretation is that these criteria reject equilibria in which B is "threatened by A's beliefs" because A stubbornly retains these beliefs in spite of plausible evidence to the contrary.

The simplest version requires that A's belief assigns zero probability to those types of B that cannot possibly gain by deviating, regardless of how A responds. The **intuitive** criterion (Cho and Kreps, 1987) requires that there cannot be some type of B that surely gains from deviating in every continuation for which A responds with a strategy that is optimal based on a belief that assigns zero probability to those types of B that cannot gain from the deviation. That is, an equilibrium fails the intuitive criterion if A's belief fails to recognize that B's deviation is a credible signal about his type. These authors also define an alternative version, called the **equilibrium domination** criterion, that requires for each continuation in which A responds with a strategy that is optimal based on a belief that assigns zero probability to those types of B that cannot gain from the deviation, there cannot be some type of B that gains from deviating. More restrictive is the criterion **D1** (Banks and Sobel, 1987), also called 'divinity' when it is applied iteratively, which requires that if the set of A's responses for which one type of B gains from deviating is larger than the set for which a second type gains then A's beliefs must assign zero probability to the second type. The criterion **D2** is similar except that some (rather than one) type of B gains. All these criteria are weaker than the **never weak best reply** criterion that requires an equilibrium to survive deletion of a player's strategy that is not an optimal reply to any equilibrium with the same outcome.

A *lexicographic* equilibrium (Blume, Brandenburger and Dekel, 1991) uses a different construction. Each player is supposed to rely on a sequence of 'theories' about others' strategies. He starts the game by assuming that his first theory of others' strategies is true, and uses his optimal strategy according to that theory. He continues doing so until he finds himself in a situation that cannot be explained by his first theory. In this case, he abandons the first theory and assumes instead that the second theory is true—or if it too

cannot explain what has happened then he proceeds to the next theory in the sequence. This provides a refinement of Nash equilibrium because each player anticipates that deviation from his optimal strategy for any theory will provoke others to abandon their current theories and strategies and thus respond with their optimal strategies for their next theories consistent with his deviant action. Lexicographic equilibria can be used to represent nearly any refinement. The hierarchy of a player's theories serves basically the same role as his system of beliefs, but the focus is on predictions of other players' strategies in the future rather than probabilities of what they know or have done in the past. The lexicographic specification has the same effect as considering small perturbations of strategies; e.g., the sequence of strategies approximating a perfect or proper equilibrium can be used to construct the hierarchy of theories.

### ***Refinements derived from perturbed games***

The other major class of refinements relies on perturbations to select among the Nash equilibria. The motive for this approach stems from a basic principle of decision theory—the **equivalence** of alternative methods of deriving optimal strategies. This principle posits that constructing a player's optimal strategy in a dynamic game by invoking auxiliary systems of beliefs and the iterative application of the principle of optimality (as in perfect-Bayes and sequential equilibria) is a useful computational procedure, but the same result should be obtainable from an initial choice of a strategy, i.e., an optimal plan of action for the entire game. Indeed, the definition of Nash equilibrium embodies this principle. Proponents therefore argue that whatever improvements come from dynamic analysis can and should be replicated by static analysis of initial choices among strategies, supplemented by additional criteria. (We use the terms 'static' and 'dynamic' analysis rather than the technical terms 'normal-form' and 'extensive-form' analysis used in game theory.) The validity of this argument is evident in the case of subgame-perfect equilibria of games with perfect information, which can be derived either from the principle of optimality using backward induction, or by iterative elimination of weakly dominated strategies in a prescribed order. The argument is reinforced by major deficiencies of dynamic analysis; e.g., we mentioned previously that a sequential equilibrium can use inadmissible strategies. Another deficiency is failure to satisfy the criterion of **invariance**; viz., the set of sequential equilibria can depend on which of many equivalent descriptions of the dynamics of the game is used (in particular, on the addition or deletion of redundant strategies)

In this view one should address directly the basic motive for refinement, which is to exclude equilibria that assume implicitly that each player commits initially to his strategy—since Nash equilibria do not distinguish between cases with and without commitment. Thus one considers explicitly that during the game any player might deviate from his equilibrium strategy for some exogenous reason that was not represented in the initial description of the game. Recognition of the possibility of deviations, however improbable they might be, then ensures that a player's strategy includes a specification of his optimal response to others' deviations from the equilibrium. The objective is therefore to characterize those equilibria that are affected only slightly by small probabilities of deviant behaviors or variations in preferences. This program is implemented by considering perturbations of the game. These can be perturbations of strategies or

payoffs, but actually the net effect of a perturbation of others' strategies is to perturb a player's payoffs.

In the following we focus on the perturbations of the static (i.e., the normal form) of the game but similar perturbations can also be applied to the dynamic version (i.e., the extensive-form) by applying them to each contingency separately. This is done by invoking the principle that a dynamic game can also be analyzed in a static framework by treating the player acting in each contingency as a new player (interpreted as the player's agent who acts solely in that contingency) in the 'agent-normal-form' of the game, where the new player's payoffs agree with those of the original player.

The construction of a *perfect* equilibrium (Selten, 1975) illustrates the basic method, which uses two steps.

1. For each small positive number  $\varepsilon$  one finds an  $\varepsilon$ -*perfect* equilibrium, defined by the requirement that each player's strategy has the following property: every one of his pure strategies is used with positive probability, but any pure strategy that is an inferior response to the others' strategies has probability no more than  $\varepsilon$ . Thus an  $\varepsilon$ -perfect equilibrium supposes that every strategy, and therefore every action during the game, might occur, even if it is suboptimal.
2. One then obtains a perfect equilibrium as the limit of a convergent subsequence of  $\varepsilon$ -perfect equilibria.

One method of constructing an  $\varepsilon$ -perfect equilibrium starts by specifying for each player  $i$  a small probability  $\delta_i < \varepsilon$  and a randomized strategy  $\sigma_i$  that uses every pure strategy with positive probability—that is, the strategy combination  $\sigma$  is 'completely mixed.' One then finds an ordinary Nash equilibrium of the perturbed game in which each player's payoffs are as follows: his payoff from each combination of all players' pure strategies is replaced by his expected payoff when each player  $i$ 's pure strategy is implemented only with probability  $1 - \delta_i$  and with probability  $\delta_i$  that player uses his randomized strategy  $\sigma_i$  instead. In this context one says that the game is perturbed by less than  $\varepsilon$  toward  $\sigma$  – we use this phrase again later when we describe stable sets of equilibria. An equilibrium of this perturbed game induces an  $\varepsilon$ -perfect equilibrium of the original game.

An alternative definition of perfect equilibrium requires that each player's strategy is an optimal response to a convergent sequence of others' strategies for which all their pure strategies have positive probability—this reveals explicitly that optimality against small probabilities of deviations is achieved, and that a perfect equilibrium uses only admissible strategies.

In fact, a perfect equilibrium of the agent-normal-form induces a sequential equilibrium of the dynamic version of the game. Moreover, if the payoffs of the dynamic game are generic (i.e., not related to each other by polynomial equations) then every sequential equilibrium is also perfect.

A stronger refinement selects *proper* equilibria (Myerson, 1978). This refinement supposes that the more inferior the expected payoff from a strategy is, the less likely it is

to be used. The construction differs only in step 1: if one pure strategy  $S$  is inferior to another  $T$  in response to the others' strategies then  $S$  has probability no more than  $\varepsilon$  times the probability of  $T$ . A proper equilibrium induces a sequential equilibrium in every one of the equivalent descriptions of the dynamic game.

A perfect or proper equilibrium depends on the particular perturbation used to construct an  $\varepsilon$ -perfect or  $\varepsilon$ -proper equilibrium. Sometimes a game has an equilibrium that is *essential* or *truly-perfect* in that any  $\sigma$  can be used when perturbing the game by less than  $\varepsilon$  toward  $\sigma$ , as above. This is usual for a static game with generic payoffs because in this case its equilibria are isolated and vary continuously with perturbations. However, such equilibria rarely exist in the important case that the static game represents a dynamic game, since in this case some strategies have the same equilibrium payoffs. This occurs because there is usually considerable freedom about how a player acts in contingencies off the predicted path of the equilibrium; in effect, the same outcome results whether the player 'punishes' others only barely enough to deter deviations, or more than enough. Indeed, for a dynamic game with generic payoffs, all the equilibria in a connected set yield the same equilibrium outcome because they differ only off the predicted path of equilibrium play. One must therefore consider sets of equilibria when invoking stringent refinements like truly-perfect. One applies a somewhat different test to sets of equilibria. When considering a set of equilibria one requires that every sufficiently small perturbation (within a specified class) of the game has an equilibrium near some equilibrium in the set. Some refinements insist on a minimal closed set of equilibria with this property, but here we ignore minimality.

The chief refinement of this kind uses strategy perturbations to generate perturbed games. Kohlberg and Mertens (1986) say that a set of equilibria is *stable* if for each neighborhood of the set there exists a positive probability  $\varepsilon$  such that, for every completely mixed strategy combination  $\sigma$ , each perturbation of the game by less than  $\varepsilon$  toward  $\sigma$  has an equilibrium within the neighborhood. Stability can be interpreted as truly-perfect applied to sets of equilibria and using the class of payoff perturbations generated by strategy perturbations. Besides the fact that a stable set always exists, it satisfies several criteria: it uses only **admissible** strategies, it contains a stable set of the reduced game after deleting a strategy that is weakly dominated or an inferior response to all equilibria in the set (these assure **iterative elimination of weakly dominated strategies** and a version of **forward induction**), and it is **invariant** to addition or deletion of redundant strategies. However, examples are known in which a stable set of a static game does not include a sequential equilibrium of the dynamic game it represents. This failure to satisfy the backward induction criterion can be remedied in various ways that we describe next.

One approach considers the larger class of all payoff perturbations. In this case, invariance to redundant strategies is not assured so it is imposed explicitly. For this, say that two games are equivalent if deletion of all redundant strategies results in the same reduced game. Similarly, randomized strategies in these two games are equivalent if they yield the same randomization over pure strategies of the reduced game. Informally, a set of equilibria is hyperstable if every payoff perturbation of an equivalent game has an

equilibrium equivalent to one near the set. Two formal versions are the following. Kohlberg and Mertens (1986) say that a set  $S$  of equilibria is *hyperstable* if, for each neighborhood  $N$  of those strategies in an equivalent game that are equivalent to ones in  $S$ , there is a sufficiently small neighborhood  $P$  of payoff perturbations for the equivalent game such that every game in  $P$  has an equilibrium in  $N$ . A somewhat stronger version is the following. A set  $S$  of equilibria of a game  $G$  is *uniformly hyperstable* if, for each neighborhood  $N$  of  $S$ , there is a  $\delta > 0$  such that every game in the  $\delta$ -neighborhood of any game equivalent to  $G$  has an equilibrium equivalent to one in  $N$ . This version emphasizes that uniform hyperstability is closely akin to a kind of continuity with respect to payoff perturbations of equivalent games. Unfortunately, both of these definitions are complex, but the second actually allows a succinct statement in the case that the set  $S$  is a ‘component’ of equilibria; viz., a maximal connected set of the Nash equilibria. In this case the component is uniformly hyperstable if and only if its topological index is nonzero, and thus **essential** in the sense used in algebraic topology to characterize a set of fixed points of a function that is slightly affected by every perturbation of the function. This provides a simply computed test of whether a component is uniformly hyperstable.

Hyperstable sets tend to be larger than stable sets of equilibria because they must be robust against a larger class of perturbations, but for this same reason the criterion is actually stronger. Within a hyperstable component there is always a stable set satisfying the criteria listed previously. There is also a proper equilibrium that induces a sequential equilibrium in every dynamic game with the same static representation—thus, the criterion of **backward induction** is also satisfied. Selecting a stable subset or a proper equilibrium inside a hyperstable component may be necessary because there can be other equilibria within a hyperstable component that use inadmissible strategies. Nevertheless, for a dynamic game with generic payoffs, all the equilibria within a single component yield the same outcome, since they differ only off the path of equilibrium play, so for the purpose of predicting the outcome rather than players’ strategies it is immaterial which equilibrium is considered. However, examples are known in which an inessential hyperstable component contains two stable sets with opposite indices with respect to perturbations of strategies.

The most restrictive refinement is the revised definition of stability proposed by Mertens (1989). Although this definition is highly technical, it can be summarized briefly as follows for the mathematically expert reader. Roughly, a closed set of equilibria is (Mertens-) *stable* if the projection map (from its neighborhood in the graph of the Nash equilibria into the space of games with perturbed strategies) is essential. Such a set satisfies all the criteria listed previously, and several more. For instance, it satisfies the **small-worlds** criterion (Mertens, 1992), which requires that adding other players whose strategies have no effect on the payoffs for the original players has no effect on the selected strategies of the original players. The persistent mystery in the study of refinements is why such sophisticated constructions seem to be necessary if a single definition is to satisfy all the criteria simultaneously. The clue seems to be that, because Nash equilibria are the solutions of a fixed-point problem, a fully adequate refinement must ensure that fixed points exist for every perturbation of this problem.

The development of increasingly stronger refinements by imposing *ad hoc* criteria incrementally was a preliminary to more systematic development. Eventually, one wants to identify decision-theoretic criteria that suffice as axioms to characterize refinements. The two groups of refinements described above approach this problem differently. Those that consider perturbations seek to verify whether there exist refinements that satisfy many or (in the case of Mertens-stability) most criteria. From its beginning in the work of Selten (1975), Myerson (1978), and Kohlberg and Mertens (1986), this has been a productive exercise, showing that refinements can enforce more stringent criteria than Nash (1950, 1951) requires. However, the results obtained depend ultimately on the class of perturbations considered, since Fudenberg, Kreps, and Levine (1988) show that each Nash equilibrium of a game is the limit of strict equilibria of perturbed games in a very general class. Perturbations are mathematical artifacts used to identify refinements with desirable properties, but they are not intrinsic to a fundamental theory of rational decision making in multi-person situations. Those in the other group directly impose decision-theoretic criteria—admissibility, iterative elimination of dominated or inferior strategies, backward induction, invariance, small worlds, etc. Their ultimate aim is to characterize refinements axiomatically. But so far none has obtained an ideal refinement of the Nash equilibria.

Srihari Govindan and Robert Wilson

## **Bibliography**

Banks, J., and Sobel, J. 1987. Equilibrium Selection in Signaling Games. *Econometrica* 55, 647-661.

Blume, L., Brandenburger, A., and Dekel, E. 1991. Lexicographic Probabilities and Choice Under Uncertainty. *Econometrica* 59, 61-79.

Blume, L., Brandenburger, A., and Dekel, E. 1991. Lexicographic Probabilities and Equilibrium Refinements. *Econometrica* 59, 81-98.

Cho, I., and Kreps, D. 1987. Signaling Games and Stable Equilibria. *Quarterly Journal of Economics* 102, 179-221.

Fudenberg, D., Kreps, D., and Levine, D. 1988. On the Robustness of Equilibrium Refinements. *Journal of Economic Theory* 44, 351-380.

Fudenberg, D., and Tirole, J. 1991. Perfect Bayesian Equilibrium And Sequential Equilibrium. *Journal of Economic Theory* 53, 236-260.

Harsanyi, J. 1967. Games with Incomplete Information Played by 'Bayesian' Players, I-III. *Management Science* 14, 159-182, 320-334, 486-502.

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

- Hillas, J., and Kohlberg, E. 2002. The Foundations of Strategic Equilibrium. Handbook of Game Theory III, edited by R. Aumann and S. Hart. Amsterdam: North-Holland/Elsevier Science Publishers.
- Kohlberg, E. 1990. Refinement of Nash Equilibrium: The Main Ideas. Game Theory and Applications, edited by T. Ichiishi, A. Neyman, and Y. Tauman. San Diego: Academic Press.
- Kohlberg, E., and Mertens, J-F. 1986. On the Strategic Stability of Equilibria. *Econometrica* 54, 1003-1038.
- Kreps, D. 1990. Game Theory and Economic Modeling. New York: Oxford University Press.
- Kreps, D., and Wilson, R. 1982. Sequential Equilibria. *Econometrica* 50, 863-894.
- Mertens, J-F. 1989. Stable Equilibria—A Reformulation, Part I: Definition and Basic Properties. *Mathematics of Operations Research* 14, 575-624.
- Mertens, J-F. 1992. The Small Worlds Axiom for Stable Equilibria. *Games and Economic Behavior* 4, 553-564.
- Mertens, J-F., and Zamir, S. 1985. Formulation of Bayesian Analysis for Games with Incomplete Information. *International Journal of Game Theory* 14, 1-29.
- Myerson, R. 1978. Refinement of the Nash Equilibrium Concept. *International Journal of Game Theory* 7, 73-80.
- Nash, J. 1950. Equilibrium Points in n-Person Games. *Proceedings of the National Academy of Sciences USA* 36, 48-49.
- Nash, J. 1951. Non-Cooperative Games. *Annals of Mathematics* 54, 286-295.
- Selten, R. 1965. Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragertragheit. *Zeitschrift für die gesamte Staatswissenschaft* 121, 301-324, 667-689.
- Selten, R. 1975. Reëxamination of the Perfectness Concept for Equilibrium Points in Extensive Games. *International Journal of Game Theory* 4, 25-55.
- van Damme, E. 1984. A Relation Between Perfect Equilibria in Extensive Form Games and Proper Equilibria in Normal Form Games. *International Journal of Game Theory* 13, 1-13.
- van Damme, E. 1989. Stable Equilibria and Forward Induction. *Journal of Economic Theory* 48, 476-496.

van Damme, E. 1991. *Stability and Perfection of Nash Equilibria*. Berlin: Springer-Verlag.