Coalitions, Leadership, and Social Norms: The Power of Suggestion in Games

PETER M. DEMARZO*

Kellogg Graduate School of Management, Northwestern University, Evanston, Illinois 60208

Received May 2, 1989

This paper examines the set of outcomes sustainable by a leader with the power to make suggestions in games. By acting as focal points, these suggestions are important even if players can communicate and form coalitions. For finite-horizon games, I show that sustainable outcomes are supported by "scapegoat" strategies, which hold a single player accountable for the actions of a group. For infinite-horizon, two-player repeated games, I show that by using an appropriate sequence of punishments and rewards, a leader can induce sufficiently patient players to play any feasible, individually rational outcome. Finally, leadership power is shown to increase if coalitions must consider the credibility of deviations in a manner similar to Coalition-Proof Nash Equilibrium. Journal of Economic Literature Classification Number: 026.

1. INTRODUCTION

This paper analyzes the power of suggestion in strategic environments. By "power of suggestion" I am referring to the ability of an individual or an institution to create a focal point for players' expectations of the behavior of others. I determine whether a leader with such power can influence the outcome of a strategic game in which communication and coalition formation is possible. Specifically, I characterize the set of outcomes that could be sustained by such a leader, and describe the structure of the suggestions that support them.

* This paper originated from discussions with Michihiro Kandori, Paul Milgrom, and Karen Van Nuys. In addition, I received helpful comments on an earlier draft from Geir Asheim, Dan Klein, Barry Weingast, and two anonymous referees.
The importance of focal points has long been recognized in the study of strategic games. When multiple equilibria exist (which is typical in the case of nonzero sum games), there is an issue of coordination or equilibrium selection that must be resolved. Many authors have proposed that the ability to create a focal point in such games might serve as a potent selection mechanism. Indeed, Thomas Schelling (1980) argues that:

A mediator, whether imposed on the game by its original rules or adopted by the players to facilitate an efficient outcome . . . can influence the [players’] expectations on his own initiative, in a manner that both parties cannot help mutually recognizing. When there is no apparent focal point for agreement, he can create one by his power to make a dramatic suggestion. . . . His directions have only the power of suggestions, but coordination requires the common acceptance of some source of suggestion.

*The Strategy of Conflict*, p. 144.

In this paper, I examine focal points for players’ expectations about the behavior of others that arise from exogenous suggestions made by some external leader. There are many possible sources of such leadership, depending upon the environment being modeled. For example, in the context of dispute resolution, a commonly recognized external mediator may act as a leader; whereas in a political body, party leadership and ideology may provide the focal point. Leadership may also emanate from more abstract sources, including historical precedent (e.g., driving on the right-hand side of the road), a common moral or behavioral code (e.g., a "fairness" principle), or a shared social norm (e.g., discrimination against certain groups). In short, a leader is any exogenous source of suggested behavior that is commonly recognized by the players and so acts as the most powerful focal point for their expectations.

By taking the leader, whether an institution or an individual, to be exogenous to the game, I can model leadership simply as a set of suggested actions for the players after any contingency. That is, leadership corresponds to an exogenously specified strategy profile for the players in the game. I call a strategy a Sustainable Social Norm if, once suggested to the players and established as the focal point for their expectations, the players will indeed choose to conform to the strategy.

Of course, a strategy can only be sustainable if the players do not have an incentive to behave differently. Consider first a purely noncooperative environment where only individual incentives are important. Then if a leader were to suggest a particular Nash Equilibrium of a simultaneous move game, one would expect the players to follow the leader. Similarly, in an extensive form noncooperative game, any Subgame Perfect Nash Equilibrium should be sustainable as a suggestion by a commonly recognized leader. This is so since, in the absence of any other means of coordination, the players will naturally be led to coordinate on the suggestion, given that no individual has an incentive to deviate. Thus, for nonco-
operative environments, the set of sustainable strategies simply corresponds to standard equilibrium notions for the game.

Most forms of social interaction, however, are more aptly described by environments in which players can communicate. In this case, it is not clear that focal points should be important or effective—players have other means of coordinating their behavior. The goal of this paper is to demonstrate that even in such environments, a leader with the power of suggestion may have an important influence on the game. Under the assumption that the players cannot choose to eliminate the leader, so that the leader can continue to make suggestions that are commonly known, I define the set of Sustainable Social Norms for environments in which coalition formation is possible. I show that for multistage games, this set no longer corresponds to any existing notion of equilibrium. Moreover, for finite-horizon games, the set of sustainable outcomes can be supported by strategies with a particular structure, called Scapegoat Strategies. These strategies hold a single player accountable for the actions of a group, and punish that player if the group deviates. For infinite-horizon games, I show that a leader, by using an appropriate sequence of punishments and rewards, can induce sufficiently patient players to play any feasible, individually rational outcome. Finally, leadership power is shown to increase if coalitions must consider the credibility of their deviations.

2. Definitions and Notation

2.1. Multistage Games

Consider a multistage game $G$ with some number $T$ (finite or infinite) of stages. During each stage, players choose their actions simultaneously, and these choices are revealed at the end of the stage. Hence, at the beginning of each stage, each player knows the entire prior history of the game. Note that standard repeated games are a special instance of this class, in which each stage game is identical.

Let $N = \{1, \ldots, n\}$ be the set of players. In the first stage of the game, each player $i$ must choose an action $a_{i1}$ from the set $A_{i1}$. The profile of all players' actions at time 1 is denoted $a_1 = (a_{11}, \ldots, a_{n1}) \in A_1 = \prod_{i \in N} A_{i1}$. This profile is then revealed and the second stage of play commences. Since the set of feasible actions in the second stage may depend upon the outcome of the first stage, player $i$ chooses a second stage action from the set $A_i(a_{i1})$, and collectively the players' choices determine the second stage profile $a_2 \in A(a_1)$.

Proceeding in this fashion, it is clear that at stage $t$, the current history of the game is the sequence $a^{t-1} = (a_1, \ldots, a_{t-1})$ of action profiles
chosen in prior stages. In general, the set of feasible actions at stage \( t \) for player \( i \) is given by \( A_i(\mathcal{a}^{t-1}) \), adopting the convention that \( a^0 = (\ ) \) and \( A_i(\ ) = A_{1i} \). Hence, the set of feasible histories for the play of the game until time \( t \) can be defined by \( H^t = \{(h^{t-1}, h_t) \mid h^{t-1} \in H^{t-1}, h_t \in A(h^{t-1})\} \), where \( H^0 = \{(\ )\} \).

Given a sequence of actions \( a^T \in H^T \) for the entire game \( G \), the payoff for player \( i \) is specified by a utility function \( U_i: H^T \mapsto \mathbb{R} \). Also define \( U(a^T) \in \mathbb{R}^N \) to be the vector of payoffs for all players. Note that for extensive form games, the function \( U \) is simply the payoff associated with each terminal node. Furthermore, since "dummy stages" which do not affect the payoffs can easily be added, the restriction that the game terminate after \( T \) stages independent of history is without loss of generality. Finally note that for standard repeated games, with stage payoffs given by \( g_i(a) \) for action \( a \) independent of history, and with discount factor \( \delta \), we have \( U_i(a^T) = \sum_{t=1}^{T} \delta^{t-1} g_i(a_t) \).

For technical simplification, it is assumed throughout this paper that for any time \( t \) and any feasible history \( h^t \), the action space \( A_i(h^t) \) is finite. This can, however, be generalized to encompass mixed strategies which are ex post observable; see DeMarzo (1988).

2.2. Strategies and Subgames

A strategy for player \( i \) is a specification of an action to take in each stage, given the current history of the game. Hence, it will consist of an action choice \( s_{i1} \in A_{1i} \) in the first period, followed by \( s_i(h^t) \in A_i(h^t) \) in a subsequent period in which the history \( h^t \) is observed. Following the earlier convention, also define \( s_i(\ ) = s_{i1} \).

A strategy profile \( s \) for the game \( G \) is simply a vector of each player's strategy, so that \( s(h^t) = (s_{1i}(h^t), \ldots, s_{ni}(h^t)) \in A(h^t) \). Denote the set of possible strategies for player \( i \) by \( S_i \), and the set of strategy profiles by \( S = \prod_{i \in N} S_i \). The game \( G \) is thus fully described by the triple \( (N, S, U) \).

Each strategy profile \( s \in S \) naturally induces an action sequence \( q(s) \in H^T \) representing the path of play in \( G \) if players acted according to \( s \). This path is defined as

\[ q(s) = (q_1(s), \ldots, q_T(s)), \]

where \( q_1(s) = s_{i1} \), and for \( t = 2, \ldots, T \),

\[ q_t(s) = s(q_1(s), \ldots, q_{t-1}(s)) = s(q^{t-1}(s)). \]

Note that we may write the payoffs received by the players if strategy profile \( s \) is followed as \( U(q(s)) \).
Any history $h' \in H'$ naturally defines a subgame $G(h')$ of $G$ given by $(N, S(h'), U(h'))$, where

$$U(h')(a^{T-t}) = U(h', a^{T-t}),$$

and

$$S(h') = \bigcup_{s \in S} s/h',$$

in which $s/h'$ denotes the restriction of strategy $s$ to the subgame; i.e., $(s/h')(a^{k}) = s(h', a^{k})$.

Finally, the following standard definitions are given for reference:

**Definition 1.** A strategy profile $s$ is a *Nash Equilibrium (NE)* for the game $G$ if there does not exist a player $i \in N$, with an alternative strategy $s_i \in S_i$ such that

$$U_i(q(s_{-i}, s_i)) > U_i(q(s)),$$

where $s_{-i} = (s_j)_{j \neq i}$.

Since the subgames of a stage game are simply the continuations after any history, Selten’s (1975) notion of Subgame Perfection is easily stated:

**Definition 2.** A strategy profile $s$ is a *Subgame Perfect Equilibrium (SPE)* for the game $G$ if for any history $h' \in H'$, $t < T$, the strategy profile $s/h'$ is a Nash equilibrium for the game $G(h')$.

### 3. Sustainable Social Norms

In this section, the concept of a Sustainable Social Norm is defined to identify the class of exogenously provided suggestions to the players that would actually be followed when the game is played. Recall from the discussion in the Introduction that whether these suggestions come from a prominent social norm, an outside mediator or leader, or historical custom, they can be identified as mappings from histories into recommended action profiles; that is, they are simply strategy profiles for the game. Hence, the definitions provided below are identical in form to those of standard solution concepts for the game, though the reader should keep in mind the particular interpretation of these strategies intended here.
3.1. Finite-Horizon Games

In the case of a single-stage game $G$, the players play a single, simultaneous move, normal form game. Hence a strategy profile is simply an action choice $s \in A_1$. Consider the following set of strategies, defined by Aumann (1959):

**Definition 3.** A strategy profile $s$ is a Strong Nash Equilibrium (SNE) for the game $G$ if there does not exist an alternative strategy profile $s' \neq s$ such that

$$U_i(s') > U_i(s), \quad \text{for all } i \text{ such that } s'_i \neq s_i. \quad (1)$$

Suppose a leader were to propose an SNE $s$ to the players at the start of the game. Would such a strategy be followed? Clearly, if player $i$ expects the others to follow the leader, player $i$ can do no better than to play $s_i$. Moreover, even if a group of players $C \subset N$ were to communicate and hence coordinate their action choice, the decision $s_c$ is optimal if they expect the remaining players to follow the leader and play $s_{-c}$. Thus, it seems that for single-stage games, any SNE $s$ would be sustained if suggested by a commonly recognized leader.

There may also exist strategies which are not SNE and yet still seem sustainable. For example, in the Prisoner’s Dilemma game, the strategy in which both players defect is not an SNE, but might be sustainable. Though both players have an interest in agreeing to cooperate, such an agreement is not self-enforcing, and so might break down. This issue of whether a coalitional deviation is self-enforcing or “credible” has been exploited by Bernheim et al. (1987), who define a weaker equilibrium notion for one-stage games called Coalition-Proof Nash Equilibrium (CPNE). This concept expands the set of SNE by ruling out deviations that are not themselves credible.

Therefore, the set of sustainable suggestions for $G$ should include the set of SNE, but may be as large as the set of CPNE if one believes that self-enforceability is an important constraint on coalitions. Lacking a clear consensus on this point, I shall take a conservative approach:

**Definition 4.** A strategy profile $s$ is a Sustainable Social Norm (SSN) for the one-stage game $G$ if and only if it is a Strong Nash Equilibrium for $G$.

This definition, by ignoring the issue of credibility for deviating coalitions, gives the least power to the leader, and hence underestimates the power of suggestion in games. In Section 6, however, I reconsider this point and offer an alternative definition of sustainability motivated by the notion of CPNE.
Given this definition of SSN for single-stage games, I extend it to finite-stage games. Suppose a strategy \( s \) is sustainable in a multistage game \( G \). This should mean that if such a strategy were suggested to the players, it would be followed by the players, even though they might communicate and contemplate various deviations. Since \( s \) should be followed on every subgame, this implies that \( s/h' \) should be a sustainable suggestion on \( G(h') \) for any history \( h' \). If this were not satisfied, then on some subgame the players would not necessarily follow \( s \), so that \( s \) could not be considered sustainable.

Alternatively, suppose a strategy \( s \) is sustainable on every proper subgame of the game \( G \). Under what conditions is \( s \) also sustainable on \( G \) itself? Again, \( s \) is considered sustainable if, once suggested, it is followed by the players. Since \( s \) is sustainable on subgames, this implies that \( s \) will be followed on every subgame, even if players communicate and contemplate deviations on them. Hence, the only time when \( s \) might not be followed is in the first period of play itself—the players might deviate in this first stage, but after that would be lead to follow \( s \) on the resulting subgame. Therefore, we can determine the sustainability of such a strategy \( s \) on the full game \( G \) by ensuring that \( s_1 \) is sustainable in the first stage of \( G \), under the assumption that the remainder of the game is played according to \( s \). To this end, I define the following:

**Definition 5.** The one-stage game induced by \( s \) on \( G \), denoted \( G_1^s \), is simply the simultaneous move game with players \( N \), action space \( A_1 \), and payoffs given by \( U(a, q(s/a)) \).

This permits the following recursive definition for sustainable suggestions in finite-stage games, based on the previous analysis of one-stage games:

**Definition 6.** A strategy profile \( s \) is a Sustainable Social Norm for the finite-stage game \( G \) if

1. for any history \( h' \) with \( t > 0 \), the strategy \( s/h' \) is an SSN for the subgame \( G(h') \), and
2. the strategy \( s_1 \) is an SNE for the one-stage game \( G_1^s \).

### 3.2. An Example

To see the implications of the definition of SSN, I consider a simple example of a bargaining game between a union and a firm. I suppose there are two possible agreements that can be reached, one involving high wages and the other low. If the firm offers high wages and the union demands low wages, an intermediate outcome is reached. Finally, if the union demands high wages and the firm offers low wages, a strike occurs and both parties lose. This is represented by the following normal form:
The Strong Nash Equilibria for this game are the outcomes (H,H) and (L,L). Suppose a mediator were brought in and suggested the outcome (H,H) to the players. Given such a focal point, this outcome would most likely be played. In fact, if the mediator is truly the most potent focal point for the players, this should be true even if the players discuss coordinating on some other outcome such as (L,L). This is so since the union strictly prefers the outcome (H,H) to (L,L) and so will support the mediator, knowing that the firm will then also find it in its best interest to support the mediator. Thus, the set of SNE are sustainable, and hence are SSN for the game.

Suppose now that this game is repeated twice. Consider an arbitrary sequence of the outcomes (H,H) or (L,L). By reasoning similar to that above, such a sequence proposed by a mediator would be followed by the players. Furthermore, it is easy to check that these strategies are indeed SSN. In fact, these are the only pure strategy SSN outcomes for this game.

Consider now another two-stage game, in which the firm is suffering "hard times" in the first period, so that medium or high wages will necessitate massive layoffs. That is, suppose the first stage of the game has payoffs:

\[
\begin{array}{c|cc}
& H & L \\
\hline
H & 4,8 & 6,6 \\
L & 0,0 & 8,4 \\
\end{array}
\]

Of course, if the mediator were to suggest the outcome (L,L) followed by either (H,H) or (L,L), such a suggestion would clearly be followed and these strategies are indeed SSN. Consider, however, the outcome (H,H) in the first period. This outcome is Pareto dominated by (L,L), and so would appear unsustainable in this game if the players could communicate. Yet suppose some mediator (either an actual individual or a social convention applicable to this environment) were to suggest in the second period that players adopt the current status quo; that is, suppose that the agreement reached in the first stage is a potent focal point in the second-
stage bargaining game. In this case the outcome \((H,H)\) could be sustained in the first stage, if the union expects that settling for low wages today will result in low wages in the future. In this case, the union should resist the firm’s attempt to negotiate lower wages in the first period, since high wages for both periods is better for the union than low wages for both periods. Moreover, even if the firm agreed to pay high wages in the second stage in exchange for low wages today, the union would recognize that the firm would renege on this commitment in the second stage and support the low wage proposal of the mediator. Hence, this “status quo norm,” which starts by suggesting high wages and in the future suggests that the current wages be maintained, is sustainable in this environment. It is easy to check that this strategy satisfies the definition of SSN, since \((H,H)\) is an SNE in the one-period game induced by this norm.

3.3. An Alternative Characterization

The definition for SSN provided earlier makes clear the recursive structure of such strategies. It does not make explicit, however, the precise class of deviations which can be used to block a proposed SSN. In this section, I develop an alternative, nonrecursive definition in which a strategy profile is identified as an SSN if there does not exist such a “blocking” deviation.

For one-stage games, SSN are defined as strategy profiles for which no coalition finds it profitable to deviate and block the proposed strategy. Consider next a two-stage game. In this case, a deviation \(\hat{s}\) for the two-stage game should block the given strategy profile \(s\) only if the members of the deviating coalition have a mutual incentive to follow it in the first stage. In addition, however, if the continuation of \(\hat{s}\) requires players to deviate from \(s\) in the second stage, the coalition that deviates in the second stage should also have an incentive to do so. That is, \(\hat{s}\) should block \(s\) in the last stage, for otherwise the leader would be followed. Extending this argument inductively suggests that for multistage games, a deviation \(\hat{s}\) “blocks” \(s\) if and only if it blocks \(s\) on every subgame in which they disagree. I call this notion sequential blocking, and it has the following equivalent nonrecursive characterization:

**Definition 7.** The strategy profile \(\hat{s}\) sequentially blocks \(s\) if \(\hat{s} \neq s\) and for any history \(h'\),

\[
U_i(h', q(\hat{s}/h')) > U_i(h', q(s/h')),
\]

for all \(i\) such that \(\hat{s}_i(h') \neq s_i(h')\).

Essentially, Eq. (2) implies that at each stage, those players who deviate under \(\hat{s}\) from the original strategy \(s\) have an incentive to do so. That is,
each strictly prefers that the game continue to be played according to \( \hat{s} \), rather than be continued according to \( s \) (given the current history). The definition expresses the idea that a deviation which is to be maintained by a coalition must not only provide its members with an incentive to deviate ex ante, but must also provide them with correct interim incentives to maintain the deviation in the continuation.

This definition of sequential blocking can be used to provide an alternative characterization of a Sustainable Social Norm:

**Theorem 1.** A strategy profile \( s \) is a Sustainable Social Norm for \( G \) if and only if there does not exist a strategy \( \hat{s} \) that sequentially blocks \( s \).

**Proof.** For one-stage games, the result is obvious, since then Eq. (2) coincides with Eq. (1). Hence, suppose the equivalence holds for games with fewer than \( T \) stages.

For a \( T \)-stage game \( G \), \( s \) is an SSN iff \( s/h' \) is an SSN for any history and \( s_1 \) is an SSN for \( G|_1 \). But by the induction hypothesis this is equivalent to \( s/h' \) not being sequentially blocked in \( G(h') \) and \( s_1 \) not being sequentially blocked in \( G|_1 \). It is straightforward to show that this is equivalent to \( s \) not being sequentially blocked in \( G \).

In addition to providing an alternative interpretation of SSN, this characterization is no longer recursive in structure. It can therefore be immediately extended to yield a definition of SSN for infinite-stage games, which are further discussed in Section 5.

## 4. Finite-Horizon Games

In this section, I develop an explicit characterization of SSN outcomes for finite games, examine the relationship between SSN and existing solution concepts, and identify a class of strategies which can support any SSN outcome.

First I show that for finite-stage games, one can verify if a strategy profile is an SSN by checking player incentives in the one-stage games induced by the strategy after any history. This result implies that Sustainable Social Norms can be characterized in a dynamic programming fashion, as Abreu (1988) and others have done for Subgame Perfect Equilibria.

**Theorem 2.** A strategy profile \( s \) is an SSN for a finite-horizon game \( G \) if and only if for every history \( h' \), \( s(h') \) is an SSN in \( G|(h') \).

**Proof.** By definition, if \( s(h') \) is not an SSN, then \( s/h' \) is not an SSN, so that \( s \) is not an SSN. Next suppose \( s \) is not an SSN. From Theorem 1, some \( \hat{s} \) sequentially blocks \( s \). Let \( h' \) be the longest possible history such
that $\hat{s}/h^t \neq s/h^t$, so that $\hat{s}/h^t$ differs from $s/h^t$ only in stage $t + 1$. Since $\hat{s}/h^t$ sequentially blocks $s/h^t$, $\hat{s}(h^t)$ must block $s(h^t)$ in $G^i(h^t)$.  

4.1. Relation to Other Solution Concepts

Since Sustainable Social Norms were designed to eliminate the possibility of coalitional deviations, it is natural that they should rule out single-agent deviations. Hence one expects that an SSN should be a refinement of Subgame Perfect Equilibrium. In fact, the following result holds for both finite- and infinite-horizon games:

**Theorem 3.** If a strategy profile $s$ is an SSN for $G$ then $s$ is a Subgame Perfect Equilibrium for $G$.

*Proof.* Note that since an SSN is an SSN on every subgame, the theorem is proved by showing that SSN is a refinement on Nash Equilibrium. Suppose $s$ were not a Nash Equilibrium. Then there exists a player $i$ and a strategy $\hat{s}_i$ such that if $\hat{s} = (\hat{s}_i, s_{-i})$, $U_i(q(\hat{s})) > U_i(q(s))$. Define a new strategy profile $\bar{s}$ equal to $\hat{s}$ except that if

$$U_i(h', q(\bar{s}/h')) < U_i(h', q(s/h')),$$

then $\bar{s}/h^t = s/h^t$. This modification only improves the payoff to player $i$, and guarantees that $\bar{s}$ sequentially blocks $s$. Hence $s$ is not an SSN.  

In general, SSN is in fact a strict refinement of Subgame Perfect Equilibrium. This is obvious simply by considering one-stage games, in which SSN coincides with Strong Nash. There is an important case, however, in which the two concepts coincide. These are games with perfect information, which correspond to stage games in which only a single player has a nondegenerate action choice in any period.

**Theorem 4.** Suppose $G$ is a finite game with perfect information. Then a strategy profile $s$ is an SSN for $G$ if and only if $s$ is a Subgame Perfect Equilibrium for $G$.

*Proof.* From the previous result, it is only necessary to show that if $s$ is Subgame Perfect, it is also an SSN. Suppose $s$ were not an SSN. Then from Theorem 2, $s(h^t)$ is not an SNE for $G_i^i(h^t)$ for some history. But $G_i^i(h^t)$ is by hypothesis a one-player game. Thus $s/h^t$ cannot be a Nash Equilibrium which implies $s$ is not Subgame Perfect.  

Recall that an SSN is defined by considering possible coalition deviations. Hence it is useful to examine the relationship between SSN and Strong Nash Equilibria. Aumann’s original definition of SNE did not impose any constraints on the strategy off the equilibrium path. Since the definition of SSN given here embodies such out of equilibrium restric-
tions, it is natural to compare SSN to Strong Perfect Equilibria, defined by Rubinstein (1980), as follows:

**Definition 8.** A strategy profile \( s \) is a **Strong Perfect Equilibrium** for \( G \) if for any history \( h' \), the strategy profile \( s/h' \) is an SNE for \( G(h') \).

It is now shown that SSN is a weaker requirement than Strong Perfect Equilibria, so that the set of SSN contains all Strong Perfect Equilibria.

**Theorem 5.** If a strategy profile \( s \) is a Strong Perfect Equilibrium for a finite-stage game \( G \), then \( s \) is an SSN for \( G \).

**Proof.** From Theorem 2, if \( s \) is not an SSN then for some history \( s(h') \) is not an SNE for \( G_i(h') \). But then \( s/h' \) is not an SNE for \( G(h') \), and \( s \) is not a Strong Perfect Equilibrium. ☐

Again note that in general the concept of SSN is strictly weaker than that of a Strong Perfect Equilibrium. This is easily seen in the case of perfect information games considered above, in which SSN coincide with Subgame Perfection, yet SNE and hence Strong Perfect Equilibria do not.

### 4.2. Optimal Scapegoat Strategies

In this section, I develop a characterization of Sustainable Social Norms that demonstrates that a very particular strategic structure can be used to support any outcome which can be achieved by an SSN. These strategies have the form of a penal code in the sense that if a deviation occurs, certain players are selected for punishment. Furthermore, this structure can be used to derive a simple characterization of the set of possible SSN outcomes.

First define the optimal (most severe) punishment that can be inflicted on a player using a sustainable or enforceable strategy:

**Definition 9.** An optimal punishment strategy for player \( i \) given history \( h' \) is a strategy profile

\[
\rho^i(h') \in \arg\min \{ U_i(h',q(s)) \mid s \text{ is an SSN for } G(h') \}.
\]

It has the associated punishment value,

\[
\omega_i(h') = U_i(h',q(\rho^i(h'))).
\]

Again, this optimal punishment strategy represents the worst outcome player \( i \) can expect after reaching the subgame \( G(h') \). In the case of Subgame Perfect Equilibria, in which one is only concerned with single-player deviations, it is natural to construct strategies which punish the deviating player by playing the worst possible equilibrium path for that
player on the following subgame. Indeed, Abreu (1988) and Benoit and Krishna (1985) have demonstrated that any outcome of a Subgame Perfect Equilibrium strategy is also the outcome of such an "optimal penal code." It is of interest, therefore, to determine if some such penal structure could be used to support outcomes of SSN. Further, if such a result is possible, it is not clear how such a strategy should respond to deviations by coalitions. Should the strategy punish certain members of the coalition? And if so, which members should be held responsible?

I first define a particular class of strategies in which a single member of the deviating coalition is held accountable for the actions of the group. For obvious reasons, I call this the class of Scapegoat Strategies:

**Definition 10.** A strategy profile $s$ for a finite game $G$ is a Scapegoat Strategy if whenever $h_t \neq s(h_t^{-1})$,

$$s/h_t = \rho^i(h_t),$$

for some $i$ such that $h_{it} \neq s_i(h_t^{-1})$; i.e., whenever a deviation occurs, one of the deviating players is punished optimally.

It is of interest to determine both the scope of such strategies—that is, how large a subset of the SSN outcomes they can support—as well as the optimal rule by which to select the "scapegoat." The following lemma provides an important first step:

**Lemma 6.** Suppose a strategy profile $s$ is an SSN for $G$. Consider another strategy profile $\tilde{s}$ that is identical to $s$ except that for $a \neq s_1$,

$$\tilde{s}/a = \rho^i(a), \quad \text{where } i \in \arg\max_{j: a_j \neq s_j} U_j(q(s)) - \omega_j(a). \quad (3)$$

Then $\tilde{s}$ is an SSN for $G$ with $q(\tilde{s}) = q(s)$.

**Proof.** Since $\tilde{s}$ only differs from $s$ off the induced path of play, clearly $q(\tilde{s}) = q(s)$. It is also obvious that $\tilde{s}/h_t$ is an SSN after any history $h_t$ with $t > 0$ since $\tilde{s}/h_1$ is in any case defined to be an SSN. Thus, it is only necessary to show that $\tilde{s}_1 = s_1$ is an SSN in $G_1^t$. Consider an arbitrary deviation $a$. Since $s_1$ is an SSN in $G_1^t$, we have

$$U_i(a, q(s/a)) \leq U_i(q(s))$$

for some player $i$ such that $a_i \neq s_{1i}$. Since $\omega_i(a) \leq U_i(a, q(s/a))$ by definition, this implies $\omega_i(a) \leq U_i(q(s))$, so that

$$\max_{j: a_j \neq s_j} U_j(q(s)) - \omega_j(a) \geq 0.$$
Therefore, $a$ cannot block $s_1$ in $G_I^f$ since for the player $i$ who is punished according to (3),

$$U_i(a,q(\hat{s}/a)) = U_i(a,p'(a)) = \omega_i(a) \leq U_i(q(s)) = U_i(q(\hat{s})).$$

Hence, $\hat{s}$ is an SSN. $\blacksquare$

I now use this result to demonstrate that any outcome of the game that can be achieved by a Sustainable Social Norm can be achieved by a Scapegoat Strategy, with the following simple rule selecting the scapegoat:

**Theorem 7.** Suppose $a^T$ is the outcome of some SSN for the game $G$. Then there exists an SSN $s$ with $q(s) = a^T$ such that $s$ is a Scapegoat Strategy. Moreover, the agent $i$ held accountable for the deviation $\hat{a}$ after history $h^t$ satisfies

$$i \in \arg\max_{i,\hat{a} \in s(h^t)} U_i(h^t,q(s/h^t)) - \omega_i(h^t,\hat{a}).$$

**(Proof.** The strategy $s$ can be constructed from the original SSN supporting $a^T$ by inductively applying Lemma 6. $\blacksquare$

As an example of this result, consider its application to a standard repeated game, in which $U(a^T) = \sum_{t=1}^T \delta^{t-1} g(a_t)$. The stationarity of such games implies that the punishment value $\omega_i(h^t)$ depends only on the number of remaining periods $T - t$, and so can be written $\omega_i^{T-t}$. In this case, if a deviation $\hat{a}$ occurs from the SSN $s$ in the first stage of a $T$-period repeated game $G$, the player to be punished is the member of the deviating coalition that maximizes

$$U_i(q(s)) - g_i(\hat{a}) - \delta^t \omega_i^{T-1}.$$ 

Thus, the player held accountable for the group deviation is one who (1) receives little from the deviation ($g_i(\hat{a})$ is low), (2) is easy to punish ($\omega_i^{T-1}$ is low), and/or (3) has much to lose ($U_i(q(s))$ is high).

5. **Infinite-Horizon Games**

In this section, I develop the theory of Sustainable Social Norms for infinite-stage games (i.e., $T = \infty$). Before beginning, the following notation must be introduced. The set of feasible outcomes $H^\infty$ for the game $G$ is the set of action sequences $a^\infty = (a_1, a_2, \ldots)$ such that

$$a^t \in H^t, \quad \text{for} \ t = 1, 2, \ldots$$
Strategy profiles \( s \) and their paths \( q(s) \) are defined in the obvious manner. Preferences are given by the utility function \( U: H^\infty \rightarrow \mathbb{R}^n \), and it is assumed that the payoff associated with any infinite-action sequence can be arbitrarily approximated by finite sequences of increasing length, so that for any \( a^\infty \in H^\infty \),

\[
U_i(a^\infty) = \lim_{t \to \infty} U_i(a^t, b^\infty(t)),
\]  

for any sequence of paths \( b^\infty(t) \in H^\infty(a^t) \). (This is satisfied, of course, if stage payoffs are bounded and the discount factor is strictly less than 1.)

Recall from Section 3.3 that SSN is defined for infinite-horizon games according to the characterization provided by Theorem 1. That is, \( s \) is an SSN for the infinite-horizon game \( G \) if there does not exist a strategy \( \bar{s} \) that sequentially blocks \( s \).

For finite-horizon games, Theorem 2 demonstrates that this definition can be reduced to verifying that the one-stage incentives faced by the players are satisfied. Clearly, this is still a necessary condition for an SSN—if \( s \) is an SSN for the infinite-horizon game \( G \), then \( s(h^t) \) is an SSN for \( G_i(h^t) \) after any history \( h^t \). Unfortunately, however, this condition is no longer sufficient to guarantee that a strategy is an SSN. To see this, consider the following two-player example:

Player 1 moves first, and can choose to "trust" and continue the game, or to quit immediately. If he quits, he receives 8, player 2 receives 0, and the game ends. If he continues, both players receive an intermediate-stage payoff of 5 each, and the second player gets to move. Player 2 must now choose to quit or to trust and continue; if she quits, she receives 8 and player 1 receives 0, and the game ends. Otherwise, they both get 5 apiece and the game repeats with player 1 moving again.

Suppose that the players discount their payoffs using some discount factor \( \delta < 1 \). Consider the following strategy profile for the game: Each player quits whenever put on the move. It is easy to check that this is a
Subgame Perfect Equilibrium for the game and that, after any history, quitting is an optimal strategy in the one-stage induced game, given that players quit in the future. If $\delta > \frac{1}{3}$, however, this strategy is an SSN for the game: it is sequentially blocked by the strategy in which both players play trust at every stage, receiving a payoff of $5/(1 - \delta) > 8$. Thus the characterization and results developed in Section 4 for SSN in finite-horizon games do not apply to infinite-horizon games.$^1$

In the above example, the given strategy satisfied the one-stage incentives of the players, but was not an SSN because it was sequentially blocked by an alternative in which both players deviate infinitely often. I now demonstrate that if a strategy satisfies both one-stage incentives and admits no deviation involving such an infinitely lived coalition, then the strategy is an SSN. First define:

**Definition 11.** Player $i$ deviates infinitely often from $s$ along the path $\hat{a}^\infty$ if for any $t$, there exists $\tau > t$ such that

$$\hat{a}_i^{\tau} \neq s_i(\hat{a}^{\tau-1}).$$

**Theorem 8.** A strategy profile $s$ is an SSN for an infinite-stage game $G$ if and only if after any history $h'$,

1. $s(h')$ is an SSN for $G(h')$, and
2. there is no strategy $\hat{s}$ that sequentially blocks $s/h'$ in which at least two players deviate infinitely often along the path $q(\hat{s})$.

**Proof.** Necessity of these conditions is immediate. For sufficiency, suppose $s$ is not an SSN. Then there exists a strategy $\hat{s}$ that sequentially blocks $s$. Since $\hat{s} \neq s$, there is a history $h^k$ such that $\hat{s}(h^k) \neq s(h^k)$. If two or more players deviate infinitely often along $q(\hat{s}/h^k)$, then condition (2) is violated, since $\hat{s}/h^k$ sequentially blocks $s/h^k$. Otherwise, let $\hat{a}^\infty = (h^k,q(\hat{s}/h^k))$. If no player deviates infinitely often along $\hat{a}^\infty$, then there must exist some period $\tau \geq k$ in which the last deviation occurs. But then $\hat{a}_i$ must block $s(\hat{a}^{\tau-1})$ in $G(\hat{a}^{\tau-1})$ and condition (1) is violated. Next suppose that only one player $i \in N$ deviates infinitely often. Then there exists a period $\tau \geq k$ such that player $i$ alone deviates and no other players deviate in any subsequent periods. Since $\hat{s}$ sequentially blocks $s$, it must be that

$$U_i(\hat{a}_i^\tau,q(\hat{a}_i)/\hat{a}_i^\tau)) = U_i(\hat{a}^\infty) > U_i(\hat{a}_i^{\tau-1},q(s/\hat{a}_i^{\tau-1})).$$

However, using Eq. (5),

$$\lim_{\tau \to \infty} U_i(\hat{a}_i^\tau,q(s/\hat{a}_i^\tau)) = U_i(\hat{a}^\infty),$$

$^1$ This result also distinguishes SSN from Greenberg's (1989) notion of Coalitional-Perfect Equilibrium Paths for infinite-horizon repeated games, which relies on one-stage incentives.
so that there exists a \( t \geq \tau \) in which
\[
U_i(\bar{a}^t, q(s/\bar{a}^t)) > U_i(\bar{a}^{t-1}, q(s/\bar{a}^{t-1}))
\]
for the first time. But since \( t \) is the first such period, this implies
\[
U_i(\bar{a}^t, q(s/\bar{a}^t)) > U_i(\bar{a}^{t-1}, q(s/\bar{a}^{t-1})),
\]
so that \( \bar{a} \) blocks \( s(\bar{a}^{t-1}) \) in \( GS_1(\bar{a}^{t-1}) \) and again condition (1) is violated. \( \blacksquare \)

This result is useful since it implies that one-stage incentives can be used to verify an SSN if infinitely lived coalitions can be ruled out.

5.1. Two-Person Supergames

In this section, I apply the definition of SSN for infinite-horizon games to the special case of two-person repeated games with discounting. These games are naturally divided into two distinct types. If the stage game is a "Common Interest" Game, so that the players can only achieve their maximal payoff jointly, then the repetition of this outcome is the unique SSN of the supergame. In contrast, I show that for "Competitive" Games in which it is impossible for both players to receive their maximal payoffs simultaneously, the set of SSN is radically different: if the players are sufficiently patient, any feasible, individually rational payoff may be sustained by a leader as an SSN.

5.1.1. Definitions and Notation. The two-player supergame \( GS \) is defined as follows: In each period, players play an identical, simultaneous move normal form game. The set of feasible action profiles in this stage game is given by the set \( A \), and the stage payoffs are given by the function \( g: A \rightarrow \mathbb{R}^n \). Players in \( GS \) evaluate payoffs using discount factor \( \delta \), so that for \( a^x \in H^x = A^x \),
\[
U_i(a^x) = (1 - \delta) \sum_{t=1}^{\infty} \delta^{t-1} g_i(a_t),
\]
where the payoffs have been normalized so that they may be interpreted as the average discounted value of the payoff stream.

Define the feasible payoffs in the game as those payoffs which can be achieved by any action profile:

Note that there is a nongeneric set of games that lies on the boundary between these two types; games in which both players can achieve their maximum payoff simultaneously, but in which one player can also achieve her maximum payoff alone. For these games, the limiting behavior as players become sufficiently patient is either joint payoff maximization alone, or the Folk Theorem result of competitive games.
For convenience, I allow public, observable randomizations in the game, so that the set \( F \) is convex.\(^3\) Also, for any payoff \( r \in F \), define the action \( a(r) \) to be any element of \( A \) for which \( g(a(r)) = r \).

Next, define a minmax action profile for each player \( i \in N \) as follows:

\[
M^i \in \arg \min_{a_i \in A_i} \max_{a_{-i} \in A_{-i}} g_i(a^i, a^{-i}).
\]

This implies that \( g_i(M^i) \) is the worst payoff \( i \) could be forced to accept in the one-period game. Without loss of generality, the payoffs may be normalized so that

\[
g_i(M^i) = 0, \quad i \in N.
\]

Thus, the set of feasible, individually rational payoffs consists of those payoffs in \( F \) for which all players are getting more than their minmax payoff,

\[
R = \{ r \in F \mid r_i > g_i(M^i), i \in N \} = F \cap \mathbb{R}^n_{++},
\]

and note that \( R \) is also convex by the prior assumption. I assume that this set is nonempty.

Denote the maximum possible payoff in the game \( G \) for each player as

\[
\bar{g}_i = \max_{a \in A_i} g_i(a).
\]

Also, define the Pareto frontier of the set \( R \) by

\[
\text{Par}(R) = \{ r \in R \mid \exists r' \in R, r' > r \},
\]

that is, \( \text{Par}(R) \) is the set of rational and feasible payoffs which are not strictly dominated by some other feasible payoff.

Finally, let the set of possible equilibrium payoffs for the game be denoted

\[
V^\delta = \{ U(q(s)) \mid s \text{ is an SSN for } G^\delta \},
\]

and note that our use of average payoffs implies \( V^\delta \subseteq F \).

\(^3\) This is a standard assumption used in proving the Folk Theorem for repeated games. See, for example, Fudenberg and Maskin (1986).
5.1.2. Common Interest Games. As stated above, Common Interest Games are those games in which either both players receive their maximal payoff $g_i$ or neither do. That is,

**Definition 12.** The stage game $g$ is a Common Interest Game if $\text{Par}(R) = \{g\}$.

For such games, there is no way that the players can individually or jointly improve upon the strategy profile in which they achieve $g$ in each period. Thus such a strategy profile is clearly an SSN. Furthermore, any strategy profile that yielded a payoff of less than $g$ in any period could obviously be sequentially blocked by the deviation to the above strategy. Hence,

**Theorem 9.** If the stage game $g$ is a Common Interest Game, then for any $\delta \in [0,1)$, the outcome of any SSN for $G^\delta$ is joint payoff maximization; that is, $V^\delta = \{g\}$.

5.1.3. Competitive Games. Next consider a broad class of games which may be described as Competitive Games. For these games, there is no action profile in the one-stage game that both players would unanimously prefer—joint payoff maximization is not feasible. More precisely,

**Definition 13.** The stage game $g$ is a Competitive Game if $\bar{g} \in F$.

For such games, it is obvious that $\text{Par}(F)$ and hence $\text{Par}(R)$ are nondegenerate. Consider, then, two distinct payoffs on the Pareto frontier of $R$. It is clear that the two players have diametrically opposed interests in choosing which of these payoffs is obtained. Thus, there is no incentive for the players to form a coalition if they were to negotiate over such payoffs. This might imply that strategies with payoffs on or near the Pareto frontier would not be upset by coalitional deviations. The following theorem verifies this intuition:

**Theorem 10.** Suppose $v^1, v^2 \in \text{Par}(R)$, with $v^1_1 > v^1_2$ and $v^2_2 > v^2_1$. Then there exists $\delta_* \in (0,1)$ such that for all $\delta \in (\delta_*,1)$, if $r \in R$ and $r \geq (v^1_1,v^2_1)$, then $r \in V^\delta$.

**Proof.** Since $\bar{g} \notin F$, the convexity of $F$ and the Pareto efficiency of $v^1$ and $v^2$ imply that $(v^1_1,v^2_2) \notin F$. Then, since $F$ is closed, there exists an $m > 0$ such that

$$\{x \mid x \geq (v^1_1 - m,v^2_2 - m)\} \cap F = \emptyset.$$  

See the construction in Fig. 1.

Next, I explicitly construct SSN strategy profiles for this game having the desired payoffs. Choose an integer
FIG. 1. Construction for Theorem 10.

\[ T > \max_{i,j} \bar{g}_i / v_i^j \]  

(7)

to be the length of the "punishment" phase used in the strategy. Then, select \( \delta_* \) to satisfy the relations

\[
\sum_{t=0}^{T} \delta_*^t = \frac{1 - \delta_*^{T+1}}{1 - \delta_*} \geq T,
\]

(8)

\[
(1 - \delta_*^T) \leq \frac{1}{2} \min_{i,j} m_i (\bar{g}_i - g_i(M^i)),
\]

(9)

which are monotonic in \( \delta_* \), and thus are also satisfied for any \( \delta \in (\delta_*, 1) \).

Now construct the following simple strategy \( S \), which starts with path \( P_0 \):

- **P**. Play \( a(r) \) repeatedly. If player \( i \) deviates alone then start \( P_i \). If both players deviate, start \( P_1 \).

- **P**. Play \( M_i \) for \( T \) periods. Then play \( a(v') \) thereafter. If player \( j \) deviates alone or with \( i \) then start \( P_j \). If \( i \) deviates alone, restart \( P_i \).

First, it is obvious that \( U(q(s)) = r \), so that the strategy does indeed produce the desired payoff. Now check to see that it is an equilibrium. First, I verify one-stage incentives. The following result is useful:
Lemma 11. Suppose the current continuation value $v$ has $v_i \geq v_j$. Then punishment path $P_i$ will deter $i$ from deviating.

Proof. The net gain from deviating to $\hat{a}$ for player $i$ if punished is given by

\[
(1 - \delta) g_i(\hat{a}) + \delta U_i(q(P_i)) - v_i \leq (1 - \delta) \tilde{g}_i + \delta^{T+1} v_j - v_i \\
= (1 - \delta) v_j \{ \tilde{g}_i / v_j - (1 - \delta^{T+1})/(1 - \delta) \} \\
\leq (1 - \delta) v_j \{ \tilde{g}_i / v_j - T \} \\
< 0.
\]

Thus, $i$ will not wish to deviate if punishment will result. □

Thus, no player will have an incentive to deviate from the path $P_0$ alone, and player 1 will not wish to deviate together with 2.

Next, consider the path $P_i$. First, consider player $j$. If there are $t \leq T$ remaining punishment periods, $j$'s continuation payoff is given by

\[
(1 - \delta') g_j(M') + \delta' v_j = v_j - (1 - \delta')(v_j - g_j(M')) \\
\geq v_j - m/2 > v_j.
\]

and player $j$ has no incentive to deviate alone or with player $i$ by Lemma 11. Now, consider a deviation by player $i$ alone. During the first $T$ periods, $i$ is being minmaxed, so that deviating merely postpones the end of the punishment period, and $i$ will not deviate. After $T$ periods, $i$ receives $v_j$ in continuation, so Lemma 11 again applies.

Therefore, I have shown that $s(h^1)$ is an SSN for $G_1^1(h^1)$ for any history $h^1$. Then from Theorem 8, the only deviations that may sequentially block the constructed strategy must require both players to deviate infinitely often.

Suppose $s$ sequentially blocks $s/h^1$ and hence has both players deviating infinitely often. Let $\delta^s = q(s)$. From the construction of $s$, there must exist a time $t + t_1$ when player 1 deviates from path $P_2$, and a time $t + t_2$ which is the first time after this that player 2 deviates. For player 1 to have an incentive to deviate from $P_2$, the deviation must yield a higher continuation payoff, so that, using Eq. (10),

\[
U_1(q(s/\hat{a}^{i-1})) > v_1 - m/2. \quad (11)
\]

Similarly, since player 2 must also have an incentive to deviate from $P_1$ at time $t + t_2$, it must be that

\[
U_2(q(s/\hat{a}^{i-1})) > v_2 - m/2.
\]
Now, between periods $t_1$ and $t_2$, player 2 is playing according to $P_1$. It is easy to check that this implies player 1’s average payoff during this period is less than $v_1^2$, even if player 1 continues to deviate. Thus, we have

\[
U_1(q(s/\tilde{a}^{t_1-1})) \leq (1 - \delta) \bar{g}_1 + \delta (1 - \delta^{h-t_1-1})v_1^2 + \delta^{h-t_1}U_1(q(s/\tilde{a}^{t_1-1}))
\]

\[
\leq m/2 + (1 - \delta^{h-t_1})[v_1 - m] + \delta^{h-t_1}U_1(q(s/\tilde{a}^{t_1-1})),
\]

with the last inequality following from the definitions of $\delta_*$ and $m$. Combining this result with Eq. (11) implies

\[
m/2 + (1 - \delta^{h-t_1})[v_1 - m] + \delta^{h-t_1}U_1(q(s/\tilde{a}^{t_1-1})) > v_1 - m/2,
\]

which rearranges to become $U_1(q(s/\tilde{a}^{t_1-1})) > v_1 - m$. Thus, the deviation $s$ must satisfy

\[
U(q(s/\tilde{a}^{t_1-1})) > v_1 - m, v_2^2 - m).
\]

But by the definition of $m$, this implies $U_1(q(s/\tilde{a}^{t_1-1})) \notin F$; hence, no such deviation can exist. $\blacksquare$

This result demonstrates that all payoffs on or near the Pareto frontier can be supported by SSN strategies that exploit the contradictory interests of the two players in this region. However, given that these payoffs can be supported, it may be possible to use them as punishments to support other equilibrium strategies. Indeed, if an equilibrium strategy exists which is significantly worse for a player than the current continuation value, that player will not wish to deviate and suffer such punishment. This logic is made precise in the following theorem:

**Theorem 12.** Suppose $v \in V^\delta$ for $\delta \in (\delta_v,1)$. Then for any $r \in R$ which is not Pareto inferior to $v$, there exists a $\delta_r \in (\delta_v,1)$ such that $r \in V^\delta$ for all $\delta \in (\delta_r,1)$.

**Proof.** Without loss of generality, let $r_2 > v_2$. Again, select an integer $T$ and $\delta_r \in (\delta_v,1)$ to satisfy

\[
T > \bar{g}_1/r_1,
\]

\[
\sum_{i=0}^{T} \delta_r^i = \frac{1 - \delta_r^{T+1}}{1 - \delta_r} \geq T,
\]

\[
(1 - \delta_r^T) \leq \frac{1}{2} \min \left\{ \frac{r_2 - v_2}{r_2 - g_2(M')}, \frac{r_2 - v_2}{g_2} \right\}.
\]

By hypothesis, for any $\delta \in (\delta_r,1)$ there exists an SSN $s_0$ with $V(q(s_0)) = v$. Then define the simple strategy $s$ to support the payoff $r$ as follows:
\( P_0. \) Play \( a(r) \) repeatedly. If player 1 deviates alone then start \( P_1. \) If player 2 deviates alone or with player 1, then start \( s_v. \)

\( P_1. \) Play \( M^1 \) for \( T \) periods. Then play \( a(r) \) thereafter. If player 1 deviates alone, start \( P_1. \) If player 2 deviates alone or with player 1, then start \( s_v. \)

Clearly, \( U(q(s)) = r. \) Now check one-stage incentives. Since the form of \( P_1 \) is precisely the type of punishment used in Theorem 10, Lemma 11 applies and Player 1 has no incentive to deviate individually. Now consider player 2's incentive to deviate. Her payoff for a one-stage deviation is bounded by

\[
(1 - \delta)g^2 + \delta v_2 \leq (r_2 - v_2)/2 + v_2 = (r_2 + v_2)/2,
\]

so that player 2 never deviates from a path with a continuation value of at least \((r_2 + v_2)/2. \) Her payoff playing \( P_1 \) with \( t \leq T \) punishment stages remaining is given by

\[
(1 - \delta')g(M^1) + \delta' r_2 = r_2 - (1 - \delta')(r_2 - g(M^1))
\]

\[
\geq r_2 - (r_2 - v_2)/2 = (r_2 + v_2)/2.
\]

Therefore, player 2 never has an incentive to undertake a one-stage deviation.

Thus, condition 1 of Theorem 8 is satisfied, so that the proposed strategy is an SSN unless there is some deviation which can sequentially block in which both players deviate infinitely often. However, after player 2 deviates once, the strategy profile \( s_v \) is played, so that any such strategy would also sequentially block \( s_v, \) contradicting the fact that \( s_v \) is an SSN. Thus, the constructed strategy profile is indeed an SSN yielding payoff \( r. \)

The technique used to prove the above result may obviously be iterated to further expand the set of constructed equilibrium payoffs. Eventually, such a procedure will cover the entire set \( R, \) as the next theorem demonstrates:

**Theorem 13.** Suppose \( g \) is a competitive, two-player game. Then for any feasible, individually rational payoff \( r \in R, \) there exists a \( \delta_r \in (0,1) \) such that for all \( \delta \in (\delta_r,1), \) the payoff \( r \in V^g; \) that is, \( r \) is obtained by some Sustainable Social Norm for \( G^g. \)

**Proof.** Since \( g \) is competitive, there exist distinct points \( v^1, v^2 \) in \( \text{Par}(R), \) with \( v^1_1 > v^2_1. \) Theorem 10 thus covers the case \( r \geq (v^1_1,v^2_1). \) Next, by applying Theorem 12 with \( v = v^1 \) or \( v = v^2, \) equilibria can be constructed for any payoff in \( R \) except for those where \( r \leq (v^1_1,v^2_1). \) Thus, consider this final case:
For any point $r^1 \in R$, where $r^1 \preceq (v^1_1, v^1_2)$, there exists another point $r^2 \in R$, such that $r^1$ and $r^2$ are Pareto incomparable. Without loss of generality, $r^1_1 > r^2_1$. Define $b > 0$ as follows:

$$b = \frac{1}{2} \min_{i,j} \{r^1_i - r^1_j, v^1_i - v^1_j\}.$$  

Now define the sequence of points $(r^3, r^4, \ldots)$ according to

$$r^{2k-1} = \arg\max\{r_1 \mid r \in R, r_2 = r^{2k-2}_2 - b\},$$  

$$r^{2k} = \arg\max\{r_2 \mid r \in R, r_1 = r^{2k-1}_1 - b\}.$$

By the convexity of $R$, the quadrangle $(r^1, r^2, v^2, v^1)$ is contained in $R$. (See the construction in Fig. 2). This implies that if $r^{n-1}$ and $r^n$ are both less than or equal to $(v^1_1, v^1_2)$, then

$$r^{n+1} \geq r^{n-1} + (b, b).$$

Therefore, for some finite $n$ we must have $r^n \succeq (v^1_1, v^1_2)$. Hence, $r^n$ can be supported as an equilibrium payoff according to Theorem 10. Further, the sequence was defined so that $r^{n-1}$ is Pareto incomparable with $r^n$, and so can be supported using $r^n$ according to Theorem 12. This can be iterated along the entire sequence so that $r^1$ is also supportable as an equilibrium outcome. \[\blacksquare\]
The logic of the above construction yields a surprising contrast between the form of a Sustainable Social Norm and a Subgame Perfect Equilibrium for an infinitely repeated game. With Subgame Perfection, whenever a punishment is imposed, it is sufficient to impose the most extreme punishment available. However, for an SSN, it is important that the punishment be mild, and that the player enforcing the punishment be well rewarded. This leads to a sequence of punishments that eventually brings the players out to the Pareto frontier, where no cooperation between them is possible. Thus the players know that the coalition must eventually break down, but this knowledge unravels and blocks current attempts to renegotiate from an inferior outcome. By exploiting this result, a leader or social norm can thus sustain any feasible and individually rational outcome.

6. CREDIBILITY OF DEVIATIONS

In this section, I present an alternative definition of equilibrium that accounts for the fact that certain deviations may not themselves be credible for a coalition. For example, in the Prisoner's Dilemma game, though joint defection is not a Strong Nash Equilibrium since the players have an incentive to deviate and agree to cooperate, this deviation is not itself a Nash Equilibrium and hence seems unenforceable. Bernheim et al. (BPW) (1987) have introduced a notion of Coalition-Proof Nash Equilibrium, a strategy profile which is not blocked by any credible deviation. Furthermore, a deviation from a strategy profile \( s \) by a coalition \( C \) is credible only if it is an equilibrium of the game \( G(s_{-C}) \) in which the players not in the coalition have their actions fixed according to \( s \). The formal definition given here differs somewhat from that provided by BPW, and is related to the definition given by Greenberg (1987, 1990).

**Definition 14.** A strategy profile \( s \) is a *Coalition-Proof Nash Equilibrium* for a one-stage game \( G \) if and only if there does not exist a coalition \( C \subseteq N \) and a deviation \( s_C \) such that \( s_C \) is a CPNE for the game \( G(s_{-C}) \) and

\[
U_C(s_{-C}, s_C) > U_C(s),
\]

where \( G(s_{-C}) \) is the game in which the actions of players not in \( C \) are taken as given according to \( s \).

Recall that the definition of a Sustainable Social Norm required that a strategy profile specify a Strong Nash Equilibrium in the induced one-stage game. If credibility of a deviation is also considered a necessary condition to upset a proposed equilibrium, then a more natural requirement would be that a strategy profile specifies a CPNE in the induced first-period game. This yields the following:
DEFINITION 15. A strategy profile $s$ is a *Coalition-Proof Social Norm* (CPSN) for the finite-stage game $G$ if

1. for any history $h^t$ with $t > 0$, the strategy $s/h^t$ is a CPSN for the subgame $G(h^t)$, and
2. the strategy $s_1$ is a CPNE for the one-stage game $G_1$.

This definition can also be reformulated in a way that makes clear the restrictions on the permissible set of deviations in the game. In fact, the next theorem demonstrates that a CPSN differs from an SSN precisely because it requires that deviations also be credible for a coalition, where credibility is embodied by the notion of CPSN itself:

**Theorem 14.** A strategy profile $s$ is a CPSN if and only if for any history $h^t$ there does not exist a deviation $\hat{s}_c$ that is a CPSN for $G(h^t,s_{-c})$ and that sequentially blocks $s_1/h^t$.

**Proof.** The theorem holds for one-stage games from the definition of CPNE. Suppose it holds for games with fewer than $T$ periods. First suppose $s$ is not a CPSN for the $T$-period game $G$. Then either $G(h^t)$ is not a CPSN, in which case the theorem holds by induction, or $s_1$ is not a CPNE for the first-stage game induced by $s$. But then there must exist a credible deviation $d_c$ from $s_1$, and it is easy to check that such a deviation is indeed a CPSN.

Next suppose $\hat{s}_c$ is a CPSN for $G(h^t,s_{-c})$ and sequentially blocks $s_c/h^t$. Consider the last stage in which $\hat{s}_c$ deviates from $s$. It is easy to check that in this stage, the one-stage game induced by $s$ is not a CPNE, so that $s$ cannot be a CPSN. 

Hence, requiring that deviations be credible increases the power of suggestion in games, or more formally:

**Corollary 15.** If a strategy profile $s$ is an SSN for $G$, then $s$ is a CPSN for $G$.

**Proof.** If $s$ is an SSN, then $s$ must be a CPSN, since there is no deviation that sequentially blocks $s$ by Theorem 1.

BPW also extend their definition of CPNE to extensive form games with the concept of a Perfectly Coalition-Proof Nash Equilibrium. The following definition corresponds to theirs for stage games, and is similar to the characterization developed by Asheim (1988):

**Definition 16.** A strategy profile $s$ is a *Perfectly Coalition-Proof Nash Equilibrium* (PCPNE) if and only if for any history $h^t$ there does not exist a deviation $\hat{s}_c$ that is a PCPNE for $G(h^t,s_{-c})$ and such that

$$U_c(h^t,q(s_{-c}))/h^t)) > U_c(h^t,q(s/h^t)).$$
This definition makes clear that a key conceptual difference between a CPSN and a PCPNE is that the former concept requires that the deviation sequentially block the proposed strategy profile, while the latter only requires that the deviation be ex ante profitable. As I have argued earlier, this weaker requirement may not be adequate in environments in which future expectations and focal points are not controlled by the players themselves.

Finally, I note that it is possible to construct examples demonstrating that both SSN and CPSN are neither subsets nor supersets of the set of PCPNE for general multistage games.

7. Conclusion

This paper provides a definition of equilibrium in stage games in which social norms, mediation, or a historical status quo serves as an important device for coordinating players' expectations and behavior. This concept of a Sustainable Social Norm is applied to both finite- and infinite-stage games. For finite-stage games, I show that any SSN outcome could be supported by a strategy with a particular structure, termed a Scapegoat Strategy. Such a strategy holds a particular member of a deviating coalition accountable for the group's action, and then punishes this player. The definition is also extended to infinite-stage games. I show that for two-player supergames, if players are sufficiently patient, then a leader can sustain any feasible and individually rational outcome.

One criticism of the definition of SSN provided here is that coalitions may be given too much power, since only the incentive to deviate is evaluated, but not the credibility of the deviation itself. Thus, an alternative definition is provided, called a Coalition-Proof Social Norm, that generalizes the notion of a Coalition-Proof Nash Equilibrium. This weaker notion of an SSN restricts the set of permissible deviations by requiring that they not only sequentially block a proposed strategy profile, but also be credible or "equilibrium-like" themselves.

The theory developed in this paper differs significantly from most other work on renegotiation in games. In particular, Bernheim et al. (1987), Bernheim and Ray (1989), Farrell and Maskin (1989), and Asheim (1988) have developed equilibrium notions requiring that the set of equilibrium outcomes be Pareto incomparable. This is sensible if a coalition can always renegotiate to a superior outcome from an inferior one. As demonstrated here, however, this may not be possible if the inferior outcome corresponds to a social norm, since then interim as well as ex ante incentives must be provided to keep the coalition intact.

There is some relation between SSN and Pearce's (1987) definition of renegotiation-proofness, in which players recognize that current renegoti-
COALITIONS AND SUSTAINABLE SOCIAL NORMS

ation may destroy the credibility of future punishments. In this way, old "agreements" may impact future play even after they have been renegotiated. SSN is also related to the model of intraplay communication developed by Blume (1990). Blume introduces an explicit bargaining round between stages of play, and considers the role of the bargaining "threat point" in determining the equilibrium outcome. The power of the leader in this paper is akin to the ability to determine the threat point in Blume's formulation.

Of course, much work remains to be done. Most importantly, this paper only attempts to characterize the set of outcomes sustainable by some exogenous social norm or mediator. A complete theory of social norms, however, should endogenize the creation of such norms. For a theory of institutions, the selection of the "leader" or mediator must be modeled. It is hoped that the analysis of this paper provides a useful starting point for such a theory.

REFERENCES


