Bringing Game Theory Back to Earth: Thinking, Feeling, and Talking

by

Julian Christopher Jamison

B.S., Mathematics (1994)
M.S., Mathematics (1994)
California Institute of Technology

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology

June 1998

Author .....
Department of Economics
May 15, 1998

Certified by .....
Lones A. Smith
Associate Professor of Economics
Thesis Supervisor

Certified by .....
Peter A. Diamond
Institute Professor
Thesis Supervisor

Accepted by .....
Peter Temin
Elisha Gray II Professor of Economics
Chairperson, Department Committee on Graduate Students

Bringing Game Theory Back to Earth: Thinking, Feeling, and Talking

by

Julian Christopher Jamison

Submitted to the Department of Economics on May 15, 1998
in partial fulfillment of the requirements for the degree of Doctor of Philosophy

ABSTRACT

The Nash Equilibrium concept, in which all players optimize given their beliefs and beliefs are correct in equilibrium, is central to game theory. It has long been informally justified by assuming that players may have an opportunity to engage in communication, or cheap-talk, before playing a game. Chapter I presents a formal model of this idea, in which credible announcements concerning future play in the game contribute to belief formation. In this model, it is further shown that if these announcements are made strategically, then such communication will lead to play of an efficient equilibrium in the action game, shedding light on the equilibrium selection problem.

Chapter II extends this idea to repeated games, for which a new question naturally arises. If players can always renegotiate to a Pareto superior continuation equilibrium between rounds of the stage game, threats of punishment may no longer have any bite. In this case, it may be impossible to support the superior equilibrium itself. This circularity gives rise to a notion of sets of internally renegotiation-proof equilibria. By comparing the external stability of these sets, a new concept of renegotiation-perfection for infinitely repeated games can then be defined, both axiomatically and constructively. It is unique, and agrees with the standard renegotiation definition for finitely repeated games.

Almost all equilibrium concepts take utilities – true final preferences over outcomes – as given and work from that starting point. Typically, however, in both applied settings and theoretical examples, only individual gross payoffs are known. In reality, among other factors, players may care not just about their own payoffs but also about the utilities of the other players (e.g., due to altruism). Chapter III presents a flexible formal framework in which to model this intuition, and determines the solubility of the resultant fixed-point problem. Several examples, including both well-known games and various experimental games, illustrate the potential applicability of these synergistic utilities.

Thesis Supervisor: Lones A. Smith
Title: Associate Professor of Economics

Thesis Supervisor: Peter A. Diamond
Title: Institute Professor

Acknowledgements

"I do not acknowledge anybody's advice or encouragement as being responsible for the good qualities of my work. I would be a liar if I did so. If I had a message for the great men of the world, it would not be one of thanks."

Thusly did Evariste Galois, one of the single greatest mathematicians in history, preface his memoirs. They were written while he was in prison for expressing certain reservations about the French king, Louis-Philippe. The year after he was released, 1832, Galois was killed in a duel. He was then 21 years old.

If I am not Galois' equal in ability, then I am at least his superior in longevity. Furthermore, and perhaps more important then either, I have been far more fortunate both in the generosity and support of those around me, and in the opportunity to here express my gratitude for those kindnesses.

I begin by thanking the Greeks, discoverers of reason and truth, without whom none of this would have been possible. Next, Pierre de Fermat, who has long been one of my heroes and inspirations. Finally, Benjamin Franklin, who shares my birthday, and the Founding Fathers deserve credit for creating a nation in which I can do precisely what I wish to and get paid (well, somewhat) for it.

Almost all of my formal teachers have had an impact on my life and on my desire to learn and to know, from nursery school to graduate school. From high school I can recall four teachers in particular who helped me to begin to realize the vast intellectual horizons and possibilities that exist. They are Amy Merrill, Lee Morris, Gary Thorpe, and especially Charles Dye.

In college I was able to start filling in those large blank sheets with information, and for once the product lived up to its promise. Completing a difficult mathematical proof is not like completing the last step in a process. It is instead the coming to such a total understanding that you realize the utter and absolute impossibility of the statement to be anything but true. This revelatory jump, from almost certain to truly and deeply certain, is unlike anything else I have ever experienced. Giving me the background to occasionally get to that point, and making me work hard enough to where it could happen, I owe to my math professors, especially those named David (Gabai and Wales) and Thomas (Mrowka and Wolff).

I learned my college economics from Kim Border and Charlie Plott, both of whom's classes I recall quite distinctly. It was also under each of them, separately, that I began to do original research. During my senior year, I took a topics course in co-operative game theory from Jean-Pierre Benoit. This was to be my first introduction to what became my chosen field: game theory.

In my graduate career, I have learned the most, and received the largest amount of valuable advice (both general and specific), from my advisor, Lones Smith. We may not have agreed on everything, but I have always found his thoughts enlightening, honest, and of tremendous help. The same could be said for my second advisor, Peter Diamond. They were, for me, what thesis advisors should be. Other faculty over the past years have contributed both to my persona as an economist and to this dissertation, and deserve to be mentioned. They include Abhijit Banerjee, Alberto Bisin, Colin Camerer, In-Koo Cho,

John Geanakoplos, Bengt Holmstrom, Michael Kremer, James Mirrlees, Whitney Newey, Michael Piore, Ariel Rubinstein, Paul Samuelson, Peter Temin, Jean Tirole, and, last but certainly not least, Glenn Ellison.

All three chapters of this manuscript, besides having benefited from the comments of many of the people listed above (and, indeed, many of those listed below), have been presented at the MIT Theory Lunch and have received valuable feedback as a result. Furthermore, an earlier version of chapter I was given at the seventeenth Arne Ryde Symposium in Lund, Sweden in August of 1997. Chapter II was presented as a seminar at Caltech in December of that same year. All remaining mistakes are, of course, errors and should be blamed entirely on myself.

Turning to the more personal debts that I am fortunate enough to owe, I begin with my friends and classmates. Once more, the appropriate time horizon stretches back for over twenty years. I applaud both my good luck and my good judgment concerning my "best friends" from elementary school – Jesse Davidson, Steven Schoenecker, Matthew Williams, and Michael Zurer. I had known Matt, whose family lived down the block from my family, from the age of four. My memories of him are particularly strong and poignant; of all the many wonderful people with whom I have had the privilege of interacting, he would truly have been near the very top in all ways.

High school is an impressionable time, and there are many people who come to mind fondly as I think back. I acknowledge all of them, but must limit myself to here mention only Eric Goralnick, David Hollander, Sarah Hurwitz, Ming-hsun Liu, and Sabrina Serrantino. Likewise in college: the percentage of interesting and stimulating students at Caltech is remarkably high. I was, however, closest to Elizabeth Barton, Bang Phi Dang, Ted Kanamori, Michael Mulqueen, and Lior Pachter. They got me through the heavy workload, both with direct help and with entertaining diversions consisting mostly of extended and occasionally even constructive conversations.

Perhaps the greatest strength of the MIT graduate economics program is the student body. They have been my TA's and I theirs. They have been my classmates, my research colleagues, and my friends. One even puts up with the indignity of living with me. So many of them, past and present, have contributed to my life and to my work that it would be acutely unfair to name some but not others. Let me put it this way: if you are reading this, and think that it might include you but are not positive, I assure you that it does. If you are positive, then you're right to be.

I would like also to take a moment to thank several people who, often through some small act, gave me a bit of confidence at one time or another when I needed it. They each went out of their way when they needn't have, and I do remember. They are K.C., L.D., S.H., L.M., T.R., K.S. (two of them!), and S.S.

My family has made itself both intellectually and emotionally indispensable to me, and I would not be here without them, in more ways than one. I thank Pat and Dell, Jack and Ian, Kathleen and Genevieve, Kay and Richard, Jim and Larry, and the others out there for everything over the years. I am truly fortunate to have received, in the one arena within which I had no choice, exactly what I would have chosen. My siblings, Eliot and Leslie, know who they are. (Hint: Check the last sentence.) Although doomed to be forever younger than I, they have provided an incentive structure without par and if I were a void, they would be the first to fill it. Thanks for being there.

I literally cannot say too much about my parents, Joanne and Dean. From first to last, from breadth to depth, from inspiration to frivolity, I have looked up to them. They have been teachers to coaches to audiences to everything in between, which is to say parents. Thank you indeed, and thank you again.

Finally I come to Tanya. I've said it to her before, and I'll say it again: I am the lucky one. You are the silly one. And all the other good things.

Nothing is too wonderful to be true.

Michael Faraday

This dissertation is dedicated

to my parents, who are wonderful and who made me what I am

and

to Tanya, who is wonderful and who makes me happy

Table of Contents

Introduction

Game theory, the study of optimizing behavior in strategic situations, had its start in the earlier part of this century with von Neumann and Morgenstern. It has played an increasingly central role in economic analysis in the last fifteen years. There are many fruitful applications of game theory still waiting to be pursued, and we can be sure that it is not merely a fad. That is, game theory will remain an indispensable tool in the arsenal of the economist, no matter how applied the subject matter. There are also, however, many unresolved problems in the theoretical aspects of the field. One general area of research is in pushing the depths (or the heights, if you prefer) of current models and theories. This is exemplified by the equilibrium refinements literature, but also includes active work in epistemology, among other areas. The second main line of research lies in extending the boundary of game theory, whether by modifying the theories themselves or by developing new theories. One example of this agenda is the current work on bounded rationality. It is to this latter category that this thesis belongs.

Many of the ideas that eventually turned into this dissertation were conceptualized while listening to courses in game theory, either for the first time or later (when there is more leisure to consider the material as a whole). All economic models involve drastic simplifications; the question is always what important assumption to unsimplify next. Two aspects of game theory that have typically gone unmodeled, and there are many more, are communication by the players outside of the formal game setting (where do beliefs come from?) and utility interactions among the players (where do payoffs come from?). The first chapter of the thesis is devoted to the former question. It can thus be thought of as one possible model of *talking*. The second chapter extends this to repeated games: players exercise foresight and must decide what can be achieved, given that their future selves will be performing the same calculation. Hence, it deals with an implication of *thinking*. The third chapter returns to the second question posed earlier. Preferences depend not only on material outcomes but also on emotional interactions, some of which

are precisely due to differences in these outcomes. It, then, incorporates *feeling* into game theory, albeit incompletely.

Ken Binmore (1994; see reference in Chapter III) states that he once had a paper rejected from a political science journal by both referees on the grounds that "the author does not understand that the purpose of studying the Prisoners' Dilemma is to explain why cooperation in the game is rational". It is my belief that game theory is an extremely broad analytical paradigm, and that it can answer (and, indeed, explain) objections like those attested to in the quotation above, as well as many more. It is up to the theorists to make it happen.

I. Valuable Cheap-Talk and Equilibrium Selection

1. Introduction

Samuel Goldwyn once said that oral contracts aren't worth the paper that they're written on. Nevertheless, it is not unreasonable to suppose that such a dire viewpoint is not always justified. In particular, so-called self-enforcing agreements, those which no party has any incentive to break given that all others comply, may well be carried out even if they are not binding in a formal sense. This is in fact the defining characteristic of the standard Nash Equilibrium concept, and thus one of the common justifications for this concept is that if players are allowed to communicate before playing, they could hardly reasonably agree on anything not satisfying this criterion. We assume throughout that there is no recourse to court-enforceable contracting, or equivalently that any such interactions have already taken place. Unfortunately, while intuitively pleasing, this justification for the use of Nash Equilibrium has found formal models in short supply.

On a related but distinct track of reasoning, it is natural to wonder why agents would ever agree on an inefficient outcome, given that they had a chance to talk in the first place. That is to say, why would players agree ahead of time to an outcome of a game if there were another outcome, also an equilibrium, that gave strictly greater payoffs to all of them? Once again, the challenge has lain in a realistic, but necessarily simplified, formal model of the communication process. Among other problems, this result appears to be incompatible with the arguments outlined above, in which Nash Equilibria in general are justified.

This type of communication is often called *cheap-talk*, which may be roughly defined as nonbinding, non-payoff relevant pre-play communication. Although cheap-talk has indeed received attention as a potential solution to these questions surrounding the equilibrium concept, it has in practice found the most use in the study of signaling games, in repeated environments (often in connection with learning), and in particular applied settings. These are of course all important topics, but they leave the original ambiguities unresolved. This paper, then, returns to the goal of a more comprehensive

model of pure cheap-talk and an exploration of its relationships with equilibria and equilibrium selection.

The model of cheap-talk developed below involves an unlimited communication session, called a *conversation*, before the play of a standard game. Players make announcements of what actions they plan to take in the upcoming game, and together these form one possible prediction of what they may in fact do. On the other hand, there is also a common prior forecast, given exogenously, of what each player will do; this forecast is updated as the conversation proceeds. An announcement is defined to be credible only if it is close to a best response to one or the other of these predictions about the rest of the players. Otherwise it is has no external justification and so is deemed unbelievable and disregarded. The conversation continues indefinitely in this manner, possibly but not inexorably toward some limit.

The first main result of the paper is that if the conversation converges toward a limit, then this limit must be a Nash Equilibrium of the stage game[1]. Conversely, any Nash Equilibrium is a possible limit of the conversation. This can be interpreted as saying that meaningful communication before a game can only lead to Nash outcomes. Since the cheap-talk is the initial interaction between the players, we assume that they are unaware of the communication-stage strategies of their opponents. Any strategy in this phase that is weakly dominated by another is clearly not optimal; anything else is potentially the preferred choice and is therefore, given the lack of information, one possible optimal choice[2]. The second result of the paper then states that optimal play in the conversation leads to an efficient outcome, and that any efficient outcome is a possible result of such strategic conversation[3]. We interpret this as saying that rational, or thoughtful, speech leads to efficiency. This completes the connection between cheap-talk (as modeled here), Nash Equilibria, and Pareto optimality.

This latter conclusion contrasts with previous "babbling" results, in which it is impossible to select among the set of Nash Equilibria because all communication is ignored. The main reason for the difference is that those results look for equilibria of the

---

[1] Technically, as the players are slightly indifferent (arbitrarily little), the limit is an ε-Nash Equilibrium.
[2] This is discussed in further detail in section 3.
[3] The notion of efficiency used here is *stable efficiency*, a concept that is equivalent to Pareto efficiency in generic two-person games.

extended communication game as a whole, for instance assuming that the full strategies of all players are known. Here we take a more primitive view, especially since we are in part attempting to justify the equilibrium concept in the first place. Naturally, although we do not impose beliefs about the cheap-talk stage, we still must make some assumption about beliefs entering the action game. Another method that will destroy the babbling equilibria is to assume an arbitrarily small but positive cost to sending messages; this is a restriction on the environment rather than on the structure of equilibrium or on belief formation. While this is plausible in reality, it is strictly speaking no longer a model of cheap-talk, even if the total sum spent on sending messages is always lower than the smallest payoff differential in the game.

The paper continues in section 2, which provides a brief survey of some of the relevant literature. In section 3, some motivation is given for the specific assumptions made in this conversational model of cheap-talk. Section 4 lays out the formal model, and states and proves the two main results of the paper. Several examples are detailed in section 5 in order to illustrate both the cheap-talk process and the implications of the theorems. Finally, toward the end of the paper, section 6 discusses possible extensions and concludes.

2. Literature

The concept of cheap-talk was introduced into the economics literature by Crawford and Sobel (1982) and Farrell (1987). Since that time a sizable literature has developed related to this topic, with such examples as Farrell and Gibbons (1989), Forges (1990), Farrell (1993), Aumann and Hart (1993), and Blume and Sobel (1995). A recent survey appears as Farrell and Rabin (1996). The paper which is perhaps closest to the present one is Rabin (1994). It models a finite instead of an infinite opportunity for communication, but also seeks a notion of optimality rather than equilibrium in the analysis of the extended game. The specific form of cheap-talk assumed is different from that presented below, in particular with respect to the element of choice between strategies against which to credibly best respond. The results can be framed in terms of the two central questions posed here, but are generally less conclusive in one direction or

the other. Both papers remain in the full rationality paradigm of classical game theory and previous work on cheap-talk, as opposed to, say, the evolution literature.

There are a number of papers that study a more limited class of games. For instance, Matsui (1991) applies cheap-talk to common interest games. His notion of *cyclic stability* yields efficiency in this context. Canning (1997) studies signaling games of common interest, although the messages do not necessarily constitute cheap-talk per se. He finds that off-path beliefs are vital to the question of whether or not efficiency is eventually realized; randomly drawn off-path beliefs encourage experimentation and lead to efficiency. Finally, Sandroni (1997) studies two-person repeated coordination games, without cheap-talk. He introduces the concept of *blurry beliefs*, which is a less restrictive (i.e. more fully rational) belief dynamic than those of evolutionary game theory, although it is stronger than anything used here. He shows that if the belief classes of the players satisfy *reciprocity*, then cooperation will be achieved.

This leads in to a fairly large class of papers that study repeated games and the emergence of Nash Equilibria, without introducing cheap-talk. This class includes Crawford and Haller (1990), Young (1993), and Kalai and Lehrer (1993). Finaily, there have been some experimental studies of communication and equilibrium selection in various coordination games; see e.g. Cooper *et al* (1992) and Cachon and Camerer (1996). The results can be summarized (and oversimplified) as finding that two-way pre-play communication greatly increases the chances of observing efficient equilibrium outcomes. Pertinently, this holds even if the efficient equilibrium is not risk-dominant.

3. Motivation

This section provides some intuition and (possibly) justification for the structure of the model which follows; it can be safely omitted by the impatient reader. There is an action game to be played, about which the players are assumed to have full information (in order to abstract away from any signaling incentives during the conversation). All players begin with common forecasts about what actions they will each take in the upcoming game. These can be interpreted as vague initial ideas about how the game might be played, arising perhaps from societal conventions or from focal points (hence

the assumption that they are common and known). They are not beliefs in the formal sense, although they will be updated throughout the conversation[4]. Since *a priori* nothing can be absolutely ruled out by any of the players, the prior forecasts are totally mixed[5]. Needless to say, the forecasts are not in any way binding: players ignore what they themselves are "expected" to do, although they take into account the influence of this expectation on their opponents[6].

In the conversation stage, before playing the action game, players send public messages to each other. Since we are attempting to understand what communication can achieve, we assume that there is an unlimited (but countable) opportunity to send these messages. For simplicity and without loss of generality, the messages are taken simply to be announcements of own actions in the game. One could assume instead that players announce mixtures over their possible actions, but this is an unnecessary complication. Essentially, given infinite riskless communication this slight limitation on the flexibility of messages imposes no loss in the long run. Implicitly, we are assuming that players can understand one another and that they take messages at face value (not in a strategic sense but in a linguistic sense). If the message "action L" is sent, everyone understands that to mean "action L" and not "action R". Thus, there is a *natural language* for speech; the players share enough common history or cultural affiliation that they are able to talk and understand one another in a previously unencountered situation.

Naturally, not all announcements should be considered seriously. We need to define a notion of *credibility* or believability. The first requirement is that an announced action should be *self-committing*, in the sense that if it were believed and best responded to, the original announcer would still be willing to carry through with it (within the confines of the action game). This is equivalent, then, to being in the support of some Nash Equilibrium of the action game. At the beginning of the conversation any self-

[4] The players don't have beliefs about the full strategies of their opponents, only ideas about what might actually occur in the game. Thus the forecasts are distributions over actions, not distributions over mixed strategies (themselves distributions over actions). This is not crucial to the conclusions reached.

[5] This is not strictly necessary for the results.

[6] The author has performed the analysis under the seemingly weaker assumption that all that is known about the prior forecasts is that they place a certain minimum weight on each action , but the results carry over. Since this assumption adds complexity but is no sounder in justification (why cannot the entire distributions be known if the minimum weights are?), it has been left out.

committing action is credible, so that players have a chance to guide the discussion[7]. In general there will be some tradeoff between allowing the players leeway to influence the conversation at the beginning, but requiring them at some point to pay attention to what the others are saying and to reflect that in their own announcements. It is important that, unlike in the deterministic best response dynamics of evolutionary models, players have choice over what to say – this is the hallmark of a conversation. It is this choice, along with the lack of payoffs until the action game at the very end, that differentiates this model from an evolutionary learning model.

The forecast is very slowly updated by each credible announcement. We can think of the prior forecast as the result of a long but finite fictitious history of credible announcements, with each new stated action adding to the average[8]. On the other hand, the initial forecast can be ignored and only the actual credible announcements counted toward an average: this is a player's *appearance*. In general, we recursively define an announced action to be credible if it is a best response (within ε) to either the current forecast of an opponent's behavior or to the current appearance of an opponent[9]. If there are more than two players, either the forecast or the appearance may be used for each. The intuition here is that a player can either say something like, "This is what I think you are going to do, and if so then I would plan to do such-and-such," or something along the lines of, "Okay – for the moment I'll take you at your face value, and in that case I'll want to do so-and-so." Of course he or she need only consider announcements that were credible in the first place.

At any time, a player can make any announcement desired, but only those that are credible will have an impact on the discussion. Since all players know the prior forecasts and all previous announcements, they can calculate which of these announcements were actually credible and hence also which announcements on their own part will be perceived as credible. If at any point there is but a single action that is credible for a particular player, it must be that this player can only seriously be considering that action (at that time). So in effect it does not matter whether or not he or she actually announces

[7] We could require rationalizability at this point instead; it makes no difference.
[8] Recall that the average of a multiset of actions is equivalent to a mixed strategy.
[9] We assume that players only care about payoffs up to some arbitrarily small constant ε, either because they cannot perceive finer differences or because they are indifferent over this range.

that action; it is known to all that it is being considered and hence it should count toward the forecast and appearance of that player, regardless of what may or may not be announced. This argument implies that we may assume without loss of generality that all players make credible announcements during each round of the conversation[10]. Finally, we assume that at each point in time any player can start over, i.e. declare a clean slate and begin their appearance from scratch. This is the equivalent of declaring that the conversation has broken down from his or her perspective and, among other implications, allows the players to attempt to coordinate. Although it may seem like an overly strong possibility, in fact the appearance is a powerful commitment device and so giving up on it involves a significant loss[11]. In any case, of course, the option is available equally to all of the players. This completes the description of the cheap-talk conversation.

One last remaining question about the credibility concept concerns the infinite durability of credible announcements. That is, a credible announcement always "counts", even if it is no longer credible. The reason for this is that any credible announcement indicates evidence of a desire for that action if possible, and there is no reason to think that the desire will change or that the action may not once again become plausible. In effect, each announcement has a small impact, building toward the whole impression, rather than fads of currently credible actions. In fact, if only those actions that are credible at the moment are averaged into the appearance, one can observe swings back and forth at each communication stage of what is and is not believable. Furthermore, in this setting eventually only one pure strategy will be credible, and so it is essentially impossible to converge to a mixed strategy.

Once the conversation is complete, we have a countably infinite sequence of announcements for each player, with an associated sequence of appearances (the average credible announcement to date). This latter sequence may or may not have a limit[12]. If the limit does exist for a given player, then because of the infinite horizon and the nature of the updating process, the forecasts made by the other players about this player will also

[10] We make the standard assumptions on the action game so that a best response always exists.
[11] In particular, continually starting over inhibits convergence, in which case the player has no influence on the ultimate course of the discussion. This is never optimal, as shown below.
[12] If no credible announcements were made after some finite stage, this is taken to mean that the limit does not exist. However, as above, we may assume that this does not occur.

converge, and to the same point. In this case, we specify that the beliefs held by the other players about this player, entering the stage game, are also this same point in strategy space. In this way the conversation is a model of belief formation. If the appearance does not converge, then the appropriate forecast won't either, and beliefs are left open for the moment. Of course it may be true in general that appearances have a limit only for some (possibly empty) subset of the group of players.

If the appearances of all players converge, then we say that the conversation itself converges. But in this case, every player continues to make credible announcements, and hence in the limit these announcements must be near best responses to the actions stated by the other players, and hence to the limits of the other players. Since these latter are by definition the beliefs of the given player entering the action game, his or her limit must be near a best response to his or her beliefs, and is therefore one optimal strategy to pursue in the action game. So we may assume that it is indeed chosen, validating the beliefs of the other players. Of course, since this is true for all players, the limits must be mutual best responses and thus play is at a Nash Equilibrium. This is Theorem 1 below.

We next turn our attention to the question of optimality in the cheap-talk stage of the extended game. Stepping back for a moment, we consider the question of whether or not to participate in the conversation at all, given the opportunity. Since there is a natural language with which to communicate, any player can initiate a conversation. Whether or not they participate, other players will hear and be influenced by the announcements of this player. So if they do not also make announcements, this player (or players) will have free reign to drive beliefs toward the equilibrium of their choosing (by announcing it *ad infinitum*). Since this is at least weakly bad for other players, it cannot hurt them to also join in the conversation and attempt to guide the discussion in a direction favorable for them. For instance, in the Battle of the Sexes game, a player one conversing with himself will continuously announce the equilibrium that he prefers. Entering the action game, the other player believes these announcements and best responds to them, so that play will in fact be at that equilibrium. It would have been a good idea in this case for player two to at least try to promote her favored outcome, i.e. to participate in the conversation. Thus we may assume, without any loss of generality, that all players converse.

Players do not know the cheap-talk strategies employed by their opponents (if they did, we should instead be modeling what occurred before this conversation in order for that knowledge to be gained), so must consider all of them to be possible. Thus if a cheap-talk strategy for one player never does better (in terms of the payoffs ultimately realized in the action game, of course) than a competing strategy, and does worse against at least one possible strategy profile of the opponents, then the original strategy should be discarded as suboptimal. Anything that is not weakly dominated is optimal[13]. This is intentionally a broad definition; it is meant to be as loose as possible and yet at least minimally capture the requirements of optimality. Theorem 2 below proves that if all players employ communication strategies that are optimal in this sense, then the conversation must converge to a stably efficient equilibrium of the game. This class of equilibria, defined below, are essentially those Nash Equilibria for which no coalition can break away and, on their own, force the other players to follow them to some other equilibrium that is preferred by the members of the coalition. In two-person games with distinct payoffs (a property that holds generically), this is equivalent to Pareto optimality.

4. Model

Consider a game **G** with n players and finite action spaces S_i for $i = 1,\ldots,n$[14]. Payoffs are given by u_i for $i = 1,\ldots,n$. It will be simplest to think of **G** in normal form. **G** is played exactly once, though **G** itself may be a repeated game. Before this happens, there is a **conversation** C(**G**), defined as follows. Each player begins with a totally mixed prior **forecast** $\pi_i = \pi_i^1 \in \Delta(S_i)$ about his or her behavior. The forecasts are common knowledge to all the players. At each round $t = 1,2,3,\ldots$ of the conversation player i announces $m_i^t \in S_i$. The announcements are made simultaneously by all players in each round[15].

[13] Naturally, since full rationality is assumed, we could iterate the process, but there is no need.
[14] The assumption of finiteness can be weakened.
[15] Sequential announcements lead to a forced asymmetry regarding who speaks when. The effects of this generalized first-mover advantage are irrelevant for the present discussion.

Let $NE(G) \subseteq \underset{i=1}{\overset{n}{\times}} \Delta(S_i)$ be the set of Nash Equilibria of **G** and define $E_i \subseteq S_i$ by $E_i = \{s_i \in S_i | \exists \sigma \in NE(G) \text{ with } s_i \in \text{supp}(\sigma_i)\}$. These constitute the self-committing actions for player *i*. At $t = 1$ any $m_i^1 \in E_i$ is said to be **credible**. If m_i^1 was credible, then we define $\pi_i^2 = (T\pi_i^1 + m_i^1)/(T+1)$, for some very large *T*. This captures the slow updating process of prior forecasts by credible announcements. In a similar fashion, the **appearance** is given by $p_i^2 = m_i^1$. If the initial announcement was not credible, then the forecast is not updated and the appearance is undefined. Recursively, we now define m_i^t to be credible when $m_i^t \in \varepsilon BR_i(\underset{j \neq i}{\times} q_j^t)$ with $q_j^t = \pi_j^t$ or $p_j^t \ \forall j$, where $\varepsilon BR_i(\sigma_{-i})$ denotes $\{s_i \in S_i | \max_{s_i' \in S_i} u_i(s_i', \sigma_{-i}) - u_i(s_i, \sigma_{-i}) < \varepsilon\}$ for some arbitrarily small $\varepsilon > 0$. If m_i^t is not credible[16], then $\pi_i^{t+1} = \pi_i^t$ and $p_i^{t+1} = p_i^t$. If m_i^t is credible, then we define $\pi_i^{t+1} = ((T+t-1)\pi_i^1 + m_i^1)/(T+t)$ and $p_i^{t+1} = ((t-1)p_i^t + m_i^t)/t$.

Say that player *i*'s appearance converges if player *i* never entirely stops making credible announcements and if $\lim_{t \to \infty} p_i^t$ exists. If this happens, it is clear that $\lim_{t \to \infty} \pi_i^t$ also exists and is the same – call it b_i for belief about player *i*. If the limit exists for all players, then the conversation converges. In this case, we assume that beliefs after the conversation and entering **G** are given by $\mu_i = \underset{j \neq i}{\times} b_j$.

Definition: an **acceptable equilibrium** (of **G**) is a profile $\sigma \in \underset{i=1}{\overset{n}{\times}} \Delta(S_i)$ such that $\sigma = b$ for some belief vector *b* resulting from a convergent conversation starting at some prior forecasts π; the set of acceptable equilibria is denoted $AccE(G)$

Theorem 1: (a) $NE(G) \subseteq AccE(G)$

(b) $AccE(G) \subseteq \varepsilon NE(G)$

[16] Unless player *i* has only one possible credible announcement, as discussed in section 3.

Proof: (a) Let $\sigma \in NE(G)$ and consider prior forecasts π very close to σ. By definition of Nash Equilibrium, any $s_i \in \text{supp}(\sigma_i)$ is in $\varepsilon BR_i(\pi_{-i})$. Now let the players announce actions in the support of σ in such a way as to match as nearly as possible the actual distribution prescribed by σ. These will initially all be credible as stated. Of course the forecasts will change over time, but since the updating process is slow and the cheap-talk announcements are matching the given distribution, the forecasts will always stay near σ. Hence the actions in the support will remain credible forever. In this manner, $\lim_{t\to\infty} p_i^t$ exists $\forall i$ and moreover $\lim_{t\to\infty} p_i^t = \sigma_i$. Thus σ is indeed an acceptable equilibrium.

(b) If $\sigma \in AccE(G)$ and so is the limit of a convergent conversation, it must be that all $s_i \in \text{supp}(\sigma_i)$ are credibly announced infinitely often during cheap-talk[17]. Since in the limit both the forecasts and the appearances are arbitrarily near σ, each such s_i must be in $\varepsilon BR_i(\sigma_{-i})$, and therefore $\sigma_i \in \varepsilon BR_i(\sigma_{-i}) \forall i$. □

Among other things, this result justifies the possibility that players both rationally and self-consistently hold the beliefs after a convergent conversation that are given by the model. Theorem 1 in some sense clarifies the relationship between cheap-talk (as it has been modeled here) and Nash Equilibrium. If the communication is meaningful, i.e. if the cheap-talk has a limit, then it must lead to a Nash outcome. Of course there is no guarantee that the conversation will converge, and it is quite possible that it will not[18]. Furthermore, no Nash Equilibrium, even if inefficient, can yet be ruled out. Something stronger than an acceptable equilibrium is required.

We next turn to defining the appropriate efficiency concept in this setting.

Definition: Call $\sigma \in NE(G)$ **directly attainable** from $\sigma' \in NE(G)$ by the coalition S if σ_S is Nash in the induced game fixing all players outside of S to play as in σ', and if also $\forall i \notin S$ we have $u_i(\sigma_i, \sigma_S, \sigma'_{-i,S}) > u_i(\sigma'_i, \sigma_S, \sigma'_{-i,S})$.

---

[17] In particular, since the conversation converges, there must be some round after which nobody ever cleans their slate and starts over.

[18] Consider, as one example, fictitious play in the Rock-Paper-Scissors game.

This is a strenuous definition: the first condition asks that the members of S be able to "jump" to σ from σ', and the second that once they have done so they can force the rest of the players to follow them.

Definition: Call $\sigma \in NE(G)$ **attainable** from $\sigma' \in NE(G)$ by the coalition S if there is a chain of equilibria, each directly attainable by S, leading from σ' to σ; if also $\forall i \in S$ $u_i(\sigma) > u_i(\sigma')$; and if finally there is no similar such chain (for any coalition) leading away from σ

These are once again fairly strict requirements. The second one states that all members of S must strictly prefer the new equilibrium, and the third states that the new equilibrium itself is immune to these sorts of deviations.

Definition: A Nash Equilibrium of **G** is **stably efficient** if nothing is attainable from it; the set of these equilibria is denoted $StEff(G)$

By considering the grand coalition of all players, it is clear that an equilibrium exhibiting stable efficiency will tend to be efficient. In games with distinct payoffs, no singleton coalitions can ever attain alternate equilibria (this follows from the first condition of the first definition), and hence in two-person games stable efficiency is generically equivalent to efficiency. It is clear that stably efficient equilibria always exist (since whatever is attained must itself be stably efficient). In most games, efficiency and stable efficiency will coincide, but when they do not it is important that we use the latter concept. Stable efficiency is related to the coalition-proof concept introduced by Bernheim, Peleg, and Whinston (1987) but is more farsighted in that it looks at the full implications of a coalitional deviation; it turns out that neither definition is a refinement of the other.

Recall that a cheap-talk strategy is **optimal** if it is not weakly dominated.

Definition: an **agreeable equilibrium** (of **G**) is a profile $\sigma \in \underset{i=1}{\overset{n}{\times}} \Delta(S_i)$ such that $\sigma = b$ for some belief vector b resulting from a convergent optimal conversation starting at some prior forecasts π; the set of agreeable equilibria is denoted $AgrE(G)$

Theorem 2: (a) $StEff(G) \subseteq AgrE(G)$

(b) $AgrE(G) \subseteq \varepsilon StEff(G)$

Proof: (a) Consider $\sigma \in StEff(G)$ and let the prior forecasts π be very close to σ. Since the forecasts favor σ so heavily, the only way that another equilibrium can ever be reached during the conversation is if it is directly attainable, or the result of a chain of directly attainable equilibria. Thus all of the players know that these are the only feasible outcomes and in fact (see strategies below) they can be reached in a conversation. But since σ is stably efficient, it is not possible for any player (as a member of any coalition) to be sure that by deviating to one of these alternates a superior payoff can be achieved. It must be the case that either not all members of the coalition will profit by the switch (in which case of course those who don't will not participate in the deviation) or if they do that then there is another coalition who can profitably and successfully deviate away from this new point. Of course it is possible that one's payoff will be increased by attempting to switch equilibria, but there will always be circumstances in which it is not profitable. Thus there is no strategy that weakly dominates the strategy "emulate σ", which is always available due to the prior forecasts. This implies that one optimal strategy for all players is to follow σ, and the result of this will be that the conversation converges with σ as the result. There may be other optimal strategies and there may be other possible results to the conversation, but this is sufficient to show that $\sigma \in AgrE(G)$, as desired.

(b) Suppose that a conversation is converging toward an equilibrium σ that is not stably efficient (even up to ε-indifference). If there is just one coalition that can attain a superior equilibrium for itself, they can pursue the following strategies: Erase my appearance and start over. Announce the action that leads to the first equilibrium along the chain. If the other members of the coalition have all done likewise, then we will all be able to credibly repeat this announcement in the next round, since they are mutual best

responses given the forecasts near σ for the other players. If the other members have not done this, start over again and try once more. If eventually we coordinate, then continue to make these announcements indefinitely. At some point our forecasts and appearances will be very close to this new equilibrium and the only credible choice for the other players will be to switch to it as well (this follows from the definition of directly attainable). Continue in this fashion until we finally reach our ultimate goal, where the process will conclude (by the argument in part (a) above).

Of course this attempted deviation will not always work, but it is safe in that either it works (i.e. all members of the coalition coordinate) and a higher payoff is with certainty realized, or it does not and the conversation stays at σ instead. So it weakly dominates the "emulate σ and stay where you are" strategy. Since this is true for all members of the coalition, optimality implies that they will all attempt in some way to force the switch to the preferred attainable equilibrium, and will eventually coordinate (since they always have the opportunity to start over). So σ was not in fact an agreeable equilibrium.

Similarly, if there were several coalitions that could attain superior equilibria, each member of each coalition can start over at each round and attempt to coordinate with his or her coalition. Any player who is a member of several coalitions, or who has a choice between attainable equilibria, can randomize between these possibilities. If the player puts almost all weight on his or her individually preferred outcome among all these choices, and spreads $o(\varepsilon)$ weight across the others, then this will be ε-optimal but will at the same time guarantee that with probability one coordination takes place at some point. This weakly dominates "emulate σ" because either the conversation converges to σ anyway (though this never actually happens with optimality), or another coalition coordinates (which couldn't be helped), or one of the attempted coalitions coordinates first (which increases payoffs). So once again, no optimal conversation will remain at σ and it couldn't have been agreeable. □

The intuition behind part (a) is particularly simple in two-player games. In this case, given a strong prior forecast, either player can insist on the original equilibrium σ for longer than the other player can credibly hold out against it (by definition of Nash).

So both players must optimally be able to get at least their payoff from σ. But since σ is efficient, this means that both players get exactly this payoff under any optimal strategies, and thus staying at σ itself is as good as anything else. The examples in the next section serve to illustrate the mechanisms behind both the definitions and the proof of the theorem. It should be pointed out that in most specific cases very little of the complex machinery developed above is necessary or applicable; the process is often hopefully quite natural and intuitive.

5. Examples

The most obvious example of an equilibrium selection problem is posed by the following coordination game:

	A	B
A	2,2	0,0
B	0,0	1,1

Of the three Nash Equilibria in the game, only one is efficient. Theorem 2 implies that the efficient equilibrium (A,A) is the only possible outcome after rational non-binding communication between the players, no matter the prior forecasts. This is easy to see if either of the forecasts puts significant weight on A. In that case the other player can credibly repeatedly announce A as a best response, and in this manner eventually force the only credible announcement by either player to be A. Since this yields the highest possible payoff, it is optimal and the conversation will converge to A.

If instead the prior forecasts are both heavily skewed toward B, then each player can reason as follows: 'If I announce B, we will be stuck there forever and I will get a payoff of 1. If I announce A, there is some chance that my opponent will announce B, in which case we will get stuck and I will receive 1. However, there is also some chance that my opponent will announce A. If we both continue to do this, these will remain credible announcements (since they each best respond to the other's appearance) and we will converge to the efficient equilibrium, delivering me a payoff of 2 instead. I can always go back to announcing B and force that equilibrium (or start over altogether), so

there is no risk of ending up at the really inefficient mixed equilibrium. Since there are no instantaneous payoffs lost from miscoordination along the way, the only possible optimal strategy is for me to announce *A*.'

Both players are rational, so they will in fact both announce *A* at all rounds of the cheap-talk communication and the conversation will end up converging to the efficient equilibrium. Given that the forecasts were heavily toward *B*, it may be a long time before the two players have truly convinced each other of their intention to play *A*, but they have all the time in the world and every reason to make use of it. If we looked instead at the pure coordination game in which (*A*,*A*) also yields payoffs of 1 to each player, the analysis is slightly changed. If the prior forecasts lean toward either of the symmetric and efficient pure equilibria, the conversation will converge in that direction. But if the priors miscoordinate just right (e.g. they are completely uniform for both players), it will be necessary for both players to randomize their initial announcement. If they coordinate at that point, fine. If not, they simply clean their slate, start over, and try again. At some point they must both choose the same action (this is why it is necessary to randomize rather than to try to coordinate in some deterministic pattern) and then they're done.

A less clear-cut example with a unique efficient equilibrium may be found in the following version of the "stag-hunt" game:

	S	R
S	5,5	0,4
R	4,0	3,3

Here the unique efficient equilibrium involves choosing a risk-dominated action, making it perhaps more difficult to reach. Allowing communication, however, will afford the players an opportunity to convince each other that it is perfectly safe to play action *S*. Aumann (1990) has argued to the contrary that cheap-talk may not help in this game. His reasoning is that since each player would prefer the other to take action *S*, they should each attempt to convince the other to choose it. The way to do this is by claiming that you yourself are also going to pick *S*. Therefore, hearing the other player announce *S* should be discounted as purely manipulative and ignored.

It seems that this argument is not consistent, at least when there is an unlimited chance to communicate. Rational players know that they will eventually agree on a Nash Equilibrium; there is zero probability of suckering the other player or miscoordinating. At this point it comes down to a choice among equilibria. Knowing this perfectly in advance, if a player announces *S* it must be because he or she is hoping to eventually end up at the efficient equilibrium, i.e. to end up playing *S*. It is, after all, the best response at that point. In any case, the data clearly support the idea that allowing pre-play messages increases the probability of observing the efficient but risk-dominated equilibrium; see Cachon and Camerer (1996).

We turn our attention next to the Battle of the Sexes, which is not at all a game with common interests:

	F	B
F	2,1	0,0
B	0,0	1,2

In this case it is not immediately obvious that even with communication can efficiency necessarily be achieved. If the prior forecasts favor either one of the pure equilibria, then the player who prefers that equilibrium will be able to credibly "insist" on it and it will be the ultimate limit of the conversation. If the forecasts are balanced, however, neither player can be assured of getting their preferred outcome. Insisting on it whenever possible may lead the conversation to converge toward the inefficient mixed equilibrium, which is worse for both players. So this strategy is not optimal. If instead the players "yield" to the other player with some extremely small probability at each round, this will always achieve within ε of any other strategy, and since it always leads to one of the efficient equilibria, it weakly dominates the strategy that insists forever on getting its way. The players are behaving optimally and can achieve efficiency with certainty.

As a final example, we turn to games with three players in order to explain some of the added complexity that arises. First, consider the following game in which the matrix player's payoffs are listed last:

	L	R
U	0,0,10	-5,-5,0
D	-5,-5,0	1,1,-5

A

	L	R
U	-2,-2,0	-5,-5,0
D	-5,-5,0	-1,-1,0

B

This game has two pure Nash Equilibria, namely (*U,L,A*) and (*D,R,B*), only the first of which is efficient. The second equilibrium is directly attainable from the first through a coalition of the row and column players, but it is not fully attainable because they enjoy a lower payoff in this equilibrium. Thus the first equilibrium is stably efficient (and hence the second, dominated, one cannot be) and will be the result of rational communication. Nevertheless, since the row and column payoffs would be higher at the intermediate point along the chain fixing the matrix player at *A*, the original efficient equilibrium is not coalition-proof. Now modify the payoffs slightly:

	L	R
U	2,2,10	-5,-5,0
D	-5,-5,0	1,1,-5

A

	L	R
U	-2,-2,0	-5,-5,0
D	-5,-5,0	3,3,0

B

Only the equilibrium payoffs have been changed, but the analysis has been affected greatly. Both pure equilibria are now efficient, but for exactly the reasons outlined above only the second one, (*D,R,B*), is stably efficient and can be the result of cheap-talk. On the other hand, the original equilibrium is now coalition-proof, showing the discrepancy between the two concepts.

One of the (unavoidable) limitations of this model is that it can say nothing about zero-sum games, except that communication can only converge to a Nash Equilibrium. Other games in which all equilibria are efficient, and so for which Theorem 2 is vacuous, are games with a unique Nash Equilibrium. These include Matching Pennies, Rock-Paper-Scissors (where many of the convergence problems of fictitious play show up), and the game-theoretic standby of the Prisoner's Dilemma. Of course we cannot expect that

simple communication would lead to cooperation, a strictly dominated strategy. We have assumed throughout that there is only a single (though unlimited) chance for the players to talk. If **G** is a repeated game, and the players have a full conversation between each stage, then optimal speech should lead to efficient outcomes all along the extensive form game tree, both on and off the equilibrium path. This gives rise to the difficult problem of finding renegotiation-proof equilibria[19].

6. Conclusion

Coordination games of various forms, from actual rendezvous games to super-modular games and complementarity, have received increasing attention in the game theory literature. Most equilibrium selection in such games, however, has been relatively informal, appealing to such concepts as focal points, initial conditions, or competition (essentially an evolutionary argument). Cheap-talk, i.e. costless and non-binding pre-play communication, has presented an intuitively pleasing method for formally attacking the equilibrium selection problem. The model of *conversations* presented here attempts to provide one possible resolution to this question, as well as to the even older question of justifying the Nash Equilibrium concept.

The model assumes that players meet for the first time and communicate in order to allay their uncertainty about the future actions of their opponents. Since they have no knowledge of the cheap-talk strategies used by the other players, we do not look for an actual equilibrium of the extended game. Instead, we look for all outcomes that could reasonably occur as the result of rational communication on the part of the players. Messages are defined to be credible in the context of a particular conversation. If at the end of a conversation a player has put forward a consistent and credible appearance, this is assumed to in fact be the other players' belief about his or her future actions. From this base, it is proved that meaningful communication (i.e. in which there is convergence) must end up at a Nash Equilibrium. This is a partial justification for the Nash concept. It is then proved that optimal communication, i.e. in which all players make strategic and rational announcements, leads to the deselection of inefficient equilibria.

---

[19] See, for example, Jamison (1998).

A strength of the paper is that it gives a decisive answer to these two issues within the context of a single model. It also applies to games with more than two players or that don't necessarily exhibit common interests. There are, however, several drawbacks to the model. First, the results do not prove that convergence must take place, only that if it does then it takes a certain form. Secondly, since by no means all applications allow the possibility for pre-play communication, this cannot be a general justification for the Nash concept. Finally, the model does put restrictions on the belief formation process, in that it requires some very small faith to be put in credible announcements, at least over the long run. Note that this is not a departure from full rationality; traditional models have simply left this process unmodeled. There are also a number of possible relevant extensions of this model, notably to correlated equilibrium and to introducing a stochastic element in the conversation.

Calvin Coolidge once wisely said, "It is better to remain silent and be thought a fool than to speak and prove it." But that applies only to fools: the moral of this paper is, "It is worse to remain silent and only be supposed rational than to speak and confirm it."

References

Aumann, Robert J. (1990). "Nash Equilibria Are Not Self-Enforcing," in *Economic Decision-Making: Games, Econometrics and Optimisation*, J.J. Gabszewicz et al eds., Elsevier: Amsterdam, pp. 201-6.

Aumann, Robert J. Hart, Sergiu (1993). "Polite Talk Isn't Cheap," working paper.

Bernheim, B. Douglas. Peleg, Bezalel. Whinston, Michael D. (1987). "Coalition-Proof Nash Equilibria: Concepts," *JET* **42**(1), pp. 1-12.

Blume, Andreas. Sobel, Joel (1995). "Communication-Proof Equilibria in Cheap-Talk Games," *JET* **65**(2), pp. 359-82.

Cachon, Gerard P. Camerer, Colin (1996). "Loss-Avoidance and Forward Induction in Experimental Coordination Games," *QJE* **111**(1), pp. 165-94.

Canning, David (1997). "Learning and Efficiency in Common Interest Signaling Games," in *The Dynamics of Norms*, C. Bicchieri et al eds., Cambridge University Press: Cambridge, pp. 67-85.

Cooper, Russell. de Jong, Douglas V. Forsythe, Robert. Ross, Thomas W. (1992). "Forward Induction in Coordination Games," *Econ Let* **40**(1), pp. 167-72.

Crawford, Vincent P. Haller, Hans (1990). "Learning How to Cooperate: Optimal Play in Repeated Coordination Games," *Econometrica* **58**(3), pp. 571-95.

Crawford, Vincent P. Sobel, Joel (1982). "Strategic Information Transmission," *Econometrica* **50**(6), pp. 1431-51.

Farrell, Joseph (1987). "Cheap-Talk, Coordination and Entry," *Rand J Ec* **18**(1), pp. 34-9.

Farrell, Joseph (1988). "Communication, Coordination, and Nash Equilibrium," *Econ Let* **27**(2), pp. 209-14.

Farrell, Joseph (1993). "Meaning and Credibility in Cheap-Talk Games," *GEB* **5**(4), pp. 514-31.

Farrell, Joseph. Gibbons, Robert (1989). "Cheap-Talk Can Matter in Bargaining," *JET* **48**(1), pp. 221-37.

Farrell, Joseph. Rabin, Matthew (1996). "Cheap Talk," *JEP* **10**(3), pp. 103-18.

Forges, Francoises (1990). "Universal Mechanisms," *Econometrica* **58**(6), pp. 1341-64.

Jamison, Julian C. (1998). "Renegotiation Perfection in Repeated Games," chapter II of this thesis.

Kalai, Ehud. Lehrer, Ehud (1993). "Rational Learning Leads to Nash Equilibrium," *Econometrica* **61**(5), pp. 1019-45.

Matsui, Akihiko (1991). "Cheap-Talk and Cooperation in a Society," *JET* **54**(2), pp. 245-58.

Rabin, Matthew (1994). "A Model of Pre-game Communication," *JET* **63**(2), pp. 370-91.

Sandroni, Alvaro (1997). "Reciprocity and Cooperation in Coordination Games: The Blurry Belief Approach," Northwestern working paper.

Young, H. Peyton (1993). "The Evolution of Conventions," *Econometrica* **61**(1), pp. 57-84.

II. Renegotiation Perfection in Repeated Games

1. Introduction

In sharp contrast with the welfare theorems of general equilibrium theory, traditional equilibrium concepts in game theory exhibit no particular tendency to pick out Pareto efficient outcomes. The various Folk Theorems suggest that coordination may be possible in repeated games, but that a great many other strategy choices can also be supported as equilibria[1]. This displays a disturbing lack of predictive power. However, even were it supposed possible to reach an efficient equilibrium in a one-shot game, repeated games pose another quandary: coordination generally requires the threat of punishments, and these are in turn often inefficient by their very nature. Thus if it is always possible to renegotiate to an efficient equilibrium, punishments may no longer be credible and the original equilibrium itself breaks down. Equilibria that are immune to such problems are commonly called renegotiation-proof.

For the majority of this paper, it will simply be assumed that the group of players as a whole can (and hence will) switch to a Pareto superior outcome if it is possible. This is, of course, an assumption. For instance, in a pure coordination game with Pareto-ranked outcomes, none of the standard equilibrium definitions shows a preference for the efficient action choice. It is necessary to appeal to a theory of focal points or cheap-talk in order to predict this outcome, and both of these areas are still very much in the process of being worked out.

In finitely repeated games, what is meant by renegotiation-proof is fairly well accepted. The definition and full characterization results can be found in Benoit and Krishna (1993). The idea is that in the last period only an efficient Nash equilibrium can possibly be played. So in the penultimate period, only those subgame perfect equilibria that direct efficient NE in every contingency are credible; otherwise, they would be

[1] See, e.g., Benoit and Krishna (1985) and Fudenberg and Maskin (1986).

renegotiated away from. Within this class of feasible SPE, the efficient ones (within the set) are renegotiation-proof. It is now possible to work backwards step by step, in exactly the manner by which SPE are found, to deduce which equilibria are renegotiation-proof from the perspective of the first period. As may be expected, in general these constitute a strictly smaller class than that of all subgame perfect equilibria. Also, they turn out to be neither generically efficient nor generically inefficient.

The case of infinitely repeated games has proven more difficult to resolve. It was first seriously studied in the contemporaneous papers of Bernheim and Ray (1989) and Farrell and Maskin (1989). The problem is that backward induction no longer applies. More precisely, it is necessary to determine the set of renegotiation-proof equilibria at the same time that the set of credible continuation equilibria is determined. This is because the continuation game is always identical to the original game. For example, if in a particular game the question of feasibility had been previously resolved for all equilibria but one, the problem would be trivial for that one. If the proposed equilibrium was not dominated (in payoff space) by any of the feasible equilibria, and if all of its continuation equilibria were known to be feasible, then it too would be renegotiation-proof. In any other case, it would not be. Unfortunately, such information is not available for the other equilibria unless the proposed equilibrium is already known to be credible or not. It is exactly this circularity which makes a definition hard to pin down.

It is perhaps less difficult to recognize a given definition as capturing what is meant by renegotiation-proofness. That is to say, there are certain properties that any such class of equilibria ought to satisfy. These are formalized in five axioms stated below, which can be thought of as Rationality; Consistency; Internal Stability; External Stability; and Optimality. A new definition, which is termed *renegotiation perfection*, is then constructed and is shown to satisfy the axioms. Loosely speaking, this definition differs from much of the previous literature in that it works "outward" instead of "inward". That is, it begins with the set of efficient one-shot Nash equilibria. These are certainly renegotiation-proof, *if* nothing better can be agreed upon. It then proceeds outward (in payoff space), with each stage serving as the starting point for the next, until it is possible to go no further. Finally, this concept of renegotiation perfection is shown to be equivalent to the concept of Pareto perfection for finite games, described above.

An important implicit assumption has been made throughout the preceding discussion. This is what might be called stationarity: it is assumed that whether or not an equilibrium is renegotiation-proof depends only on the structure of the game, and not, for instance, on the history of play[2]. In particular, no preference is given to the current equilibrium, and any deviations in the past are considered irrelevant information (at least for the choice among equilibria). Instead, we preserve the original motivating factor behind the study of renegotiation-proofness, that "bygones are bygones" and at any stage of the game any truly renegotiation-proof equilibrium should be considered credible. In some situations, e.g. when there is importance attached to the status quo, this may not be the correct model to use, and some authors have correspondingly taken an alternate approach to the problem of renegotiation[3]. This idea is discussed further in section 5 below, but for the majority of this paper only the standard theory is considered.

One final point, on nomenclature, is worth making before proceeding with the paper. The renegotiation literature has perhaps been somewhat unfortunately named, in that it is not a theory of renegotiation, but rather of the implications deriving from the *possibility* of renegotiation. Indeed, since all of the games considered have perfect information and no stochastic element, all actions are correctly foreseen and hence no actual renegotiation ever takes place. Be that as it may, we continue to use renegotiation-proof as a generic term for stability with respect to the possibility of unanimous deviation to an alternate and preferred feasible equilibrium. It is roughly synonymous with Pareto perfect or dynamically consistent. The specific definition developed in this paper will be called renegotiation perfect[4].

This paragraph outlines the paper, beginning with this paragraph. It is followed by section 2, which develops the axiomatic formulation of renegotiation perfection. Section 3 then introduces the constructive definition and equates the two. Next, in section 4, finite games are discussed. Section 5 presents a more detailed look at the previous literature, and of exactly where the present concept lies in relation to it. To conclude this outline, section 6 concludes.

---

[2] Of course, the actions may still be history-dependent.

[3] See, for instance, the work of Abreu and Pearce (1991,1993).

[4] Blume (1994) also uses this term, though with the nominative form renegotiation-perfectness. This is not to say that the concepts are related, but merely that there is an unfortunate scarcity of applicable names.

2. Axiomatic Definition

For the purposes of this section, it is assumed that we are given a game G and a set $RP(G)$ which constitutes the set of all renegotiation-proof equilibria of G. These are assumed to be, in some sense or other, the true dynamically stable equilibria. We ask what properties this set of equilibria ought to satisfy, if they are in fact to be what we mean by renegotiation-proof. The first axiom is very simple: each renegotiation-proof outcome should itself be an equilibrium, i.e. it should be self-enforcing. Denoting the set of Nash equilibria of G by $NE(G)$ [5], we can write this as

A.1 $RP(G) \subseteq NE(G)$

The second axiom is equally uncontroversial. If an equilibrium requires the use of continuation equilibria which are not themselves renegotiation-proof, then they will not be credible and will thus be unable to effectively enforce the original equilibrium. This renders it equally uncredible, so it cannot possibly be renegotiation-proof either. Intuitively, there would be in this case no reason not to deviate if it were myopically profitable, since it would then be possible to renegotiate away from the prescribed punishments, perhaps even back to the original equilibrium. Formally, let $\sigma \in \Sigma(G)$, where $\Sigma(G)$ is the set of all strategy profiles in G. Then after any history h which begins a subgame G', σ directs play by the profile $\sigma' = \sigma|_h \in \Sigma(G')$. Hence we can write

A.2 If $\sigma \in RP(G)$ and $\sigma' = \sigma|_h$ then $\sigma' \in RP(G')$

Thus a renegotiation-proof equilibrium should only direct continuation strategies which are themselves renegotiation-proof. Combined with **A.1**, this also implies that all renegotiation-proof equilibria will be subgame perfect. In fact, this requirement is very

[5] Of course, we could use correlated equilibrium as our basic concept instead. However, the results are qualitatively similar so we will concentrate on Nash equilibria, both to focus better on the renegotiation aspects and to facilitate comparison with the existing literature.

similar to subgame perfection in that it makes the same requirement on all continuation equilibria that is made on the original equilibrium, getting to credibility.

The first two axioms have ensured that $RP(G)$ is self-enforcing as an entity. It is now time to address the various requirements raised by the possibility of renegotiation. We first introduce a definition that is key to the idea of renegotiation, and which will be used extensively throughout the paper:

Definition: If B and B' are two sets in $\mathbf{R}^n$, then B **confounds** B', written $B \rhd B'$, if there exist $b \in B$ and $b' \in B'$ such that $b >> b'$, that is if some element of B is strictly more efficient than some element of B'

This is equivalent to $B \rhd B'$ iff $B \cap [B' + \mathbf{R}^n_{++}] \neq \varnothing$. We abuse notation very slightly and say further that if $A, A' \subseteq \Sigma(G)$, A **confounds** A' ($A \rhd A'$) whenever $u(A) \rhd u(A')$. Here $u(A)$ is the set of expected payoffs associated with A, so $u(A) \subseteq \mathbf{R}^n$ if G is an n-player game. Thus a set of strategy profiles (or equilibria) confounds another if any element of the first Pareto dominates an element of the second. In terms of renegotiation, we know that players will switch from the dominated strategies in A' to the preferred outcome in A if it is at all possible. Since this is true for at least one element of A', it may not be generally stable as a whole.

Returning to our list of properties, we know that every element of $RP(G)$ is considered feasible by definition. Hence if the players ever find themselves at an equilibrium which is dominated by an element of $RP(G)$, they will want to renegotiate to the latter, and they will be able to do so because it will be considered credible. So no equilibrium which is dominated by $RP(G)$ can possibly be renegotiation-proof. In particular, since all elements of $RP(G)$ itself are indeed renegotiation-proof, we have

A.3 $RP(G) \not\rhd RP(G)$

Thus no two elements of $RP(G)$ are Pareto rankable (we might say $RP(G)$ forms a Pareto-antichain). Sets which satisfy **A.1-A.3** will form the building blocks for the rest of

the discussion in this section and the next. They are not new to the literature, and were first developed for infinite games independently by Bernheim and Ray (1989) and Farrell and Maskin (1989). Following the terminology of the latter paper:

Definition: A set $A \subseteq \Sigma(G)$ is **weakly renegotiation-proof** (WRP) if it, replacing RP, satisfies **A.1-A.3**

The reason that we are not done is that there may be many WRP sets in a given game G, and some of them may be preferred (by all players, under all circumstances) to others. If there were a maximal, or best, WRP set then we could say that it was $RP(G)$, but unfortunately matters are generally not so simple. Yet we must somehow ensure that $RP(G)$ is truly the best that can be accomplished. There are two dangers against which it is necessary to guard: first that there is nothing credible which is better than $RP(G)$, and second that $RP(G)$ itself is credible. We already know that $RP(G)$ is internally stable, but now we are asking that it also be credible from an external viewpoint. Distinguishing between these two external stability requirements is one of the clarifying aspects of the present approach.

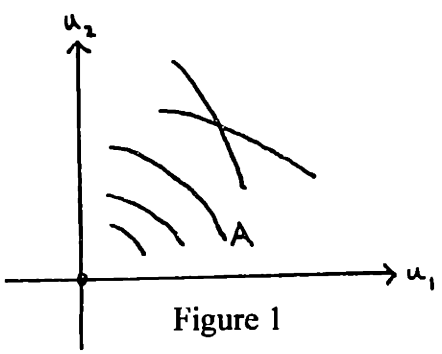

Figure 1

As an example, consider Figure 1. We assume that all of the sets depicted are in fact WRP sets. In this case, the two Pareto efficient sets confound each other. Since it is not possible for both sets to be renegotiation-proof (since the combination of the two would not even be WRP), neither can be. Loosely speaking, they can not be externally stable. We expect that the set labeled A, since it has no such problems itself and is best among all of the sets that do not, is the renegotiation-proof set, and indeed it will be chosen. In what follows, this intuition is formalized.

We begin with the question of maximality, or optimality. Essentially, what we require is that given $RP(G)$, there is no other WRP set which is stable and is not confounded by $RP(G)$. We will find the best that can be agreed upon, with $RP(G)$ as an initial point of reference, and ask that there never be an incentive to move to it. It is necessary first to introduce some further notation. If B is any class of WRP sets, and A and A' are in B, then define the minimal confounding chain length of A over A', $m_B(A,A')$, to be $\min\{k \mid \exists A_1, A_2, \ldots, A_{k-1} \in B \text{ with } A \rhd A_1 \rhd \ldots \rhd A_{k-1} \rhd A'\}$. Of course this may be infinite if A never even indirectly confounds A'. We next let $m_B(A)$ be $m_B(A,A)$, the minimal cycle length of A over itself. Since A is WRP, $m_B(A)$ is at least 2 for all B. It turns out that 3-cycles are a special and important case, falling into two groups. If A is in a 3-cycle, the other two elements in the cycle may or may not confound each other (see figure 2). If it is ever the case that they do not, i.e. A is in a "pure" 3-cycle, then we define $m_B(A)$ to be 3-, and otherwise we define it as 3+. In other words, if you first remove all 2-cycles from B, and find that $m(A)$ is now at least 4, then the original $m_B(A)$ was 3+. If it is still 3, then it was originally a 3-.

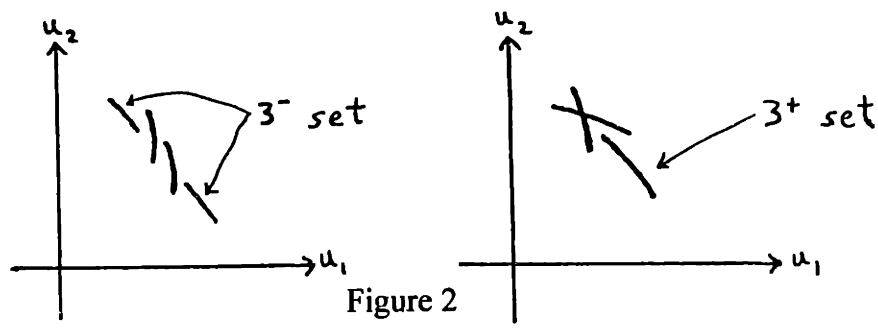

Figure 2

If $m_B(A) = 4$, so $A \rhd A_1 \rhd A_2 \rhd A_3 \rhd A$, then it must be that $A_3 \rhd A_1$. Otherwise, since $A_1 \not\rhd A_3$ (else $m_B(A)$ would have been 3), $A' = A_1 \cup A_3$ would be WRP and such that both $A \rhd A'$ and $A' \rhd A$, contradicting $m_B(A) = 4$. Note that we have assumed that B is closed under unions [of unrankable elements]. By a similar argument, if $m_B(A) = 5$ and $A \rhd A_1 \rhd A_2 \rhd A_3 \rhd A_4 \rhd A$, we must have $A_3 \rhd A_1$, $A_4 \rhd A_1$, and $A_4 \rhd A_2$. So consider the following procedure to find the maximal stable element in B: first get rid of all type 2's, i.e. all A such that $m_B(A) = 2$. They are the least stable; since there is no

way to distinguish one of the pair over the other, it is impossible to pick either. Then, in what remains, discard all type 3 WRP sets, for a similar reason. Note that this means that exactly those sets A with $m_B(A) \geq 3+$ survive past this round. This makes sense for type 3's, because they are precisely the sets that can be distinguished (by one extra degree of stability) from their compatriots in the cycle. It would be possible to continue the procedure for higher types, but this would be vacuous exactly because of the argument above. It was shown that any minimal cycle of length at least 4 had smaller cycles within its elements, so none of those elements have survived, and the cycle no longer exists.

Thus the intuitive iterative procedure leaves us with $B' \subseteq B$, where [the possibly empty] B' is given by $\{A \in B \mid m_B(A) \geq 3+\}$. As there are no cycles of any length in B', it makes sense to speak of maximal elements. We define the *solution* of B, $s(B)$, to be the union of all maximal elements in B'. Since they are all unranked with respect to each other, this is naturally also a WRP set, if it exists. Essentially, $s(B)$ is what should be chosen if B is the class of feasible WRP sets. Note that $s(G)$ always exists, where $s(G) = s(WRP(G))$ and $WRP(G)$ is the class of all WRP sets for the game G[6]. It would not, however, be consistent to think that $RP(G)$ ought simply to be $s(G)$. This is because we know that $WRP(G)$ is not truly the class of feasible WRP sets. In particular, $s(G)$ is not a consistent solution concept since there may be WRP sets which are better than it but which were knocked out only by WRP sets inferior to it. If it were truly renegotiation-proof, these latter sets would no longer be feasible, and at that point there would be no reason not to switch to one of the former sets, which are preferred (see figure 3). So $s(G)$ is not necessarily stable.

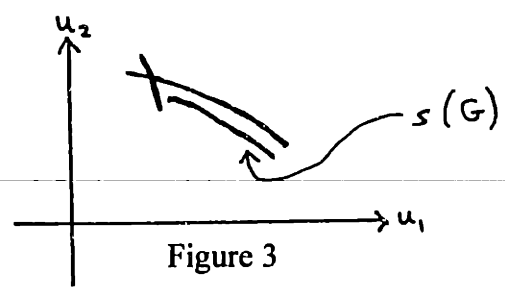

Figure 3

[6] It exists because repeating any one-shot NE constitutes a singleton WRP set, which must survive any cycle and is thus never eliminated.

Let us now suppose that we are given some arbitrary WRP set A. If "placed" there, we may ask if there is any credible incentive to move to another WRP set, and if so to where. Define the class of WRP sets that are potentially feasible outside options, given that A is the initial starting point, by

$$F(A) = \{A' \in WRP(G) \mid A' \not\subseteq A \text{ and } A' \not\triangleright A\}.$$

If $F(A)$ has a solution, i.e. if $s(F(A))$ exists, then we know by the argument above that it is a credible alternative, and of course there is an incentive to move to it. So if it exists, we will call it the *distinguished element* for A, $DE(A)$. This is where, starting at A, one would end up. If it does not exist, is there any other way in which an element of $F(A)$ can credibly be picked out? Yes – if it can "hold out longest" against being confounded by A, then it has an argument for being the optimal alternative. That is, given that $s(F(A))$ does not exist, if there exists $A' \in F(A)$ such that $m(A, A') > m(A, A'')$ for all $A'' \in F(A), A'' \neq A'$, then this A' is a credible and profitable deviation. So in this case we define it to be $DE(A)$. Essentially, we have (as the players would like to) done everything possible to distinguish a particular optimal element of $F(A)$, the set of feasible alternatives from A. If neither of the above cases holds, we define $DE(A)$ to be simply A. This gives us a complete definition of $DE(A)$, the best credible WRP set from an initial reference point of A.

What does all of this imply for renegotiation-proofness and $RP(G)$? Well, since $RP(G)$ is renegotiation-proof, there cannot possibly be a credible alternative to it which would be preferred. Otherwise the players would renegotiate to exactly that alternative, and $RP(G)$ would not have been the optimal choice. So if we define the WRP components of A to be $C(A) = \{A' \in WRP(G) \mid A' \subseteq A\}$, then one implication of stability for $RP(G)$ is

A.4 $\forall A \in C(RP)$ if $DE(A) \not\subseteq RP$ then $\exists A' \in C(RP)$ s.t. $A' \triangleright\triangleright DE(A)$

Here we are using $A \triangleright\triangleright A'$ to mean that $A \triangleright A'$ and $A' \not\triangleright A$, i.e. the former set totally confounds the latter. Hence the axiom simply requires that for any WRP set in $RP(G)$

the optimal credible alternative is either also in $RP(G)$ or is dominated by it. Again, if this were not the case, then there would be no reason not to switch to the alternative, and $RP(G)$ could not have been renegotiation-proof in the first place. It is a minimal hurdle that must at least be passed. Note that the axiom implies in particular that $DE(RP) = RP$ itself, which of course is exactly what we want.

Axiom four refers to the optimality of $RP(G)$ because it addresses the question of whether there are any WRP sets which do at least as well as $RP(G)$. It ensures that there are none which are themselves credible or stable. We have still not completed our description of $RP(G)$ for two reasons: first, there may be several sets which satisfy **A.1-A.4**, and more importantly, we have not checked the external stability of $RP(G)$ itself. These turn out to be the same question, and we already have all the tools necessary to answer it. In particular, since $RP(G)$ should not have any stability problems, any set that is worse than $RP(G)$ should not be able to satisfy **A.4**, because there would be a superior and credible alternative. Somewhat counterintuitively, this is saying essentially that $RP(G)$ should be minimal among the class of sets which satisfy the first four axioms. If something better than $RP(G)$ satisfies the axioms, that is not a problem because axiom four then implies that whatever it is cannot itself be stable. In this sense, we are looking at the stability of $RP(G)$ from both ends at once. Formally,

A.5 If $RP(G) \triangleright A$ then A does not satisfy **A.1-A.4**

We have thus completed an axiomatic description of renegotiation-proofness. The axioms have required that our renegotiation-proof set $RP(G)$ be self-enforcing and that there is no possible proposed deviation which is itself credible, either internally or externally. It is perhaps worth listing the axioms once more (if the notation can be kept in mind!), along with monikers suggestive of each axiom's role:

A.1 (Rationality) $RP(G) \subseteq NE(G)$

A.2 (Consistency) If $\sigma \in RP(G)$ and $\sigma' = \sigma|_h$ then $\sigma' \in RP(G')$

A.3 (Internal Stability) $RP(G) \not\rhd RP(G)$

A.4 (Optimality) $\forall A \in C(RP)$ if $DE(A) \not\subseteq RP$ then $\exists A' \in C(RP)$ s.t. $A' \rhd\rhd DE(A)$

A.5 (External Stability) If $RP(G) \rhd A$ then A does not satisfy **A.1-A.4**

We are now ready for our central definition:

Definition: A set of strategy profiles $RP^*(G) \subseteq \Sigma(G)$ is called **renegotiation perfect** in G if it satisfies **A.1-A.5**

Of course, we have not yet shown existence or uniqueness of renegotiation perfect sets. The question of existence will be postponed until section 3, where it will be easy to answer after giving a constructive definition of $RP^*(G)$. The question of uniqueness can be answered now. Basically, if there are two competing renegotiation perfect sets, then neither can confound the other because of axiom five. But in that case, either is feasible from the point of view of the other, and since both are known to be externally stable themselves, they will each be a distinguished element for the other. This is of course an impossible situation if they are not equal to each other. This argument is made formal in the following

Proposition 1: In any game G, any $RP^*(G)$ satisfying **A.1-A.5** is unique.

Proof: Suppose that B also satisfies **A.1-A.5**. If $B \rhd RP^*$ then by **A.5** RP^* cannot satisfy **A.1-A.4**, a contradiction. Similarly it is impossible that $RP^* \rhd B$. Let us suppose for the moment that $B \not\subseteq RP^*$.

Since $B \not\subseteq RP^*$ and $RP^* \not\rhd B$, we know that $B \in F(RP^*)$ by definition. In fact, $B \in F(A)$ for each $A \in C(RP^*)$. Now if $m_F(B) = 2$, then there exists $B' \in F(RP^*)$ such that $B \rhd B'$ and $B' \rhd B$. So replace the confounded sections of B with their B' analogues, creating a new WRP set that is "equivalent" to B. That is, let $\hat{B} = \tilde{B} \cup B''$, where $\tilde{B} = \{A \in C(B) \mid A \not\rhd B'' \text{ and } B'' \not\rhd A\}$ and $B'' = \{A' \in C(B') \mid A' \rhd B \text{ and } B \rhd A'\}$.

Then $\hat{B}$ is a WRP set and $\hat{B} \in F(RP^*)$. Furthermore, since we know B satisfies **A.4**, $\hat{B}$ also will by construction. Thus $\hat{B}$ satisfies **A.1-A.4**, but $B \rhd \hat{B}$. This contradicts **A.5** for B, and so we have $m_F(B) \neq 2$. But an exactly analogous argument shows that it also cannot be the case that $m_F(B) = 3$-. Therefore B is eligible to be the solution $s(F(RP^*))$, and in particular $s(F(RP^*))$ exists. This is however impossible, as we know that $DE(RP^*) = RP^*$.

Hence it must have been the case that $B \subseteq RP^*$ in the first place. In an entirely symmetric fashion, we must have $RP^* \subseteq B$, and so $B = RP^*$. □

Since it is now possible to legitimately speak of the single renegotiation perfect set for a game G, we can extend the concept to cover individual strategy profiles through this

Definition: A strategy $\sigma \in \Sigma(G)$ is a **renegotiation perfect equilibrium** of G if it is in the unique $RP^*(G)$

It is worth mentioning, before proceeding to the next section, that it is possible to think of the axioms in a social, rather than an individual, setting[7]. This is not to suggest that we are suddenly switching to a cooperative framework, merely that it may be worth looking briefly from a different perspective. In this scenario, then, axiom one becomes individual rationality. Axiom two can be thought of as dynamic consistency; nothing that is now feasible ever becomes infeasible, nor vice-versa. Axiom three describes social stability – once the choice set has been agreed upon, there is no reason to deviate. This also corresponds to a notion of incentive compatibility, i.e. optimality among the credible alternatives. Axiom four is of course social rationality, stating that there is nothing better that could be agreed upon. And finally, axiom five refers to a type of social consistency, requiring that it be possible to distinguish the renegotiation-proof set in some other than arbitrary manner.

[7] This line of reasoning was suggested to me by Lones Smith.

3. Constructive Definition

We are now ready to [re-]define renegotiation perfection through an explicit procedure. Since much of the notation and conceptual groundwork was developed in the previous section, this will not be a lengthy process. However, despite the fact that we are referring to terminology from above, it is better to attempt to step back from the axioms, at least from the last two, at this point. The construction should be able to stand on its own; only afterwards will we bring the two together.

So, starting anew, consider a stage game G and its infinitely repeated counterpart $G^{\infty}(\delta)$[8]. Let $RP_0 = [WEff\ NE(G)]^{\infty}$, where $WEff\ NE(G)$ is defined to be the set of weakly Pareto efficient Nash Equilibria of G, and for $\sigma \in \Sigma(G)$, σ^{∞} simply refers to the strategy profile in $G^{\infty}(\delta)$ which plays σ after every history. Then since σ^{∞} is WRP if σ is Nash, RP_0 is a WRP set. Also, as long as G has a Nash equilibrium, RP_0 will be non-empty. Now define $W_1 = \{A \in WRP(G^{\infty}) \mid RP_0 \not\triangleright A\}$. This is similar to $F(RP_0)$, except that it is not the set of feasible deviations, but rather the entire set of feasible possibilities since it includes RP_0 itself. Note that, from the definition of W_1, $m_{W_1}(RP_0)$ is infinite because RP_0 cannot possibly be a part of any cycle. Hence $s(W_1)$ exists.

Now we define RP_1 to be $s(W_1)$, the solution to the set of possibilities. Recall what this means: first you delete all unstable pairs that confound each other. Second, you delete any unstable triples, or 3-cycles, that remain. At this point there are no more cycles among any elements, and the solution is simply the union of maximal elements. As above, we are assured that it exists.

Iteratively, we let $W_i = \{A \in WRP(G^{\infty}) \mid RP_{i-1} \not\triangleright A\}$ and $RP_i = s(W_i)$. Because of RP_{i-1} itself, we know that these always exist. If for some i, RP_{i-1} is in fact the solution $s(W_i)$, then $RP_i = RP_{i-1}$ and the process terminates, that is it remains constant forever. It is clear that this must in fact occur at some point. Otherwise the limit of the RP_i sets

[8] From now on the discount factor is suppressed, for ease of notation and because it is not conceptually relevant. Of course, it is still implicitly there and it would affect the actual results.

would be WRP and would have been the solution somewhere along the way (any set which knocks it out of eligibility would also have knocked out sets arbitrarily close to it). Essentially, if the limit were okay the process would have jumped immediately to it. So, if the process stops at *i*, we may define

$$RP^*(G^\infty) = RP_\infty = RP_i .$$

Thus we have completed a second definition (for infinitely repeated games) of the renegotiation perfection concept. Once again, we call a strategy profile a renegotiation perfect equilibrium if it is in $RP^*(G^\infty) = RP_\infty$. It may be illustrative to think about how this definition works in some of our previous examples. For instance, in figure 1 it arrives immediately at the best non-cyclic WRP set, as we expected and knew it would. In the example of Figure 3, however, it requires two stages, eventually working out and pinpointing the WRP set that confounds both of the other two.

The intuition behind the construction is very straightforward. The players are trying to figure out what is the best outcome to which they can credibly all agree. A repeated one-shot NE is certainly a safe starting point. That is, *if* nothing better can be found, it will be stable. Of course, only efficient equilibria need be considered. Given that it is now known that they can always do at least that well, anything that is confounded by this set should not be thought of as credible and can be dropped from consideration. The solution (in the formal sense defined earlier) to what remains is then the obvious choice. The players can now agree to doing minimally that well. If nothing better can be found, it at least will be stable. But now, anything that is confounded by this new potential choice is no longer credible. This may free some even better sets which were previously constrained only by sets that are no longer feasible. To be consistent, they should now be reconsidered. And so on. At each stage, the players agree on a new fallback position and then ask if they can do any better[9]. It is in this sense that the process may be considered as working outward toward efficiency, rather than inward due to feasibility constraints.

[9] Naturally, this is meant only as a heuristic description. We are not attempting to formally model such a discussion or process.

Naturally, we must justify the use of the same name and notation for the two separate definitions of renegotiation perfection given above. This is validated by the following

Theorem 1: RP_∞, as constructed above, satisfies **A.1-A.5** for G^∞.

Proof: Since RP_∞ is equal to RP_i for some i and RP_i is itself a WRP set, **A.1-A.3** clearly hold. To show that **A.4** is satisfied, take any WRP set in RP_∞, i.e. take $A \in C(RP_\infty)$, and consider $A' = DE(A)$, the distinguished element starting from A. The first possibility is that $A' = s(F(A))$, the solution from A. If $A' \not\lhd RP_\infty$ and $A' \not\subseteq RP_\infty$ then $A' \in F(RP_\infty)$ and so $A' = s(F(RP_\infty))$ because $F(RP_\infty) \subseteq F(A)$ by construction. But this is impossible, since $F(RP_\infty)$ has no solution by definition of RP_∞. So either $A' \subseteq RP_\infty$, in which case we're done, or $A' \lhd RP_\infty$. If the latter, then there exists some $A'' \in C(RP_\infty)$ such that $A'' \rhd A'$. If $A' \rhd A''$ as well, then we cannot possibly have $A' = s(F(A))$ because it would have immediately been knocked out; note that RP_∞ is WRP so A'' is certainly in $F(A)$ (of course $A'' \not\subseteq A$ since $A' = s(F(A))$ and $A'' \rhd A'$). Thus $A' \not\rhd A''$, and this yields exactly the second condition of **A.4**.

The second possibility is that $m(A, A') > m(A, A'')$ for all $A'' \in F(A), A'' \neq A'$. The argument in this case is similar to the one given in detail above. Essentially, since A' is feasible, if it were not totally confounded by something in RP_∞, then it would have part of RP_∞, as required. The third and final possibility in the definition of $DE(A)$ is that $A' = A$, in which case certainly $A' \subseteq RP_\infty$. Hence we have shown that RP_∞ always satisfies **A.4**.

Now let A be WRP and such that $RP_\infty \rhd A$. If $A \not\rhd RP_\infty$, then $RP_\infty \in F(A)$ and we can define j to be the smallest number such that $RP_j \in F(A)$ and $RP_j \rhd A$. Then since $RP_j = s(W_j)$, either $RP_j = s(F(A))$ or there exists A' such that $RP_{j-1} \rhd A' \rhd RP_j$ and so (since this sort of cycle must hold for any such A') RP_j "holds out longest" with respect to A. In either case $DE(A) = RP_j \neq A$ and thus A does not satisfy **A.4**. If on the

other hand $A \rhd RP_\infty$ (so they confound each other), let j minimally satisfy $RP_j \rhd A$ and $A \rhd RP_j$. Since $RP_j = s(W_j)$, we have that $RP_{j-1} \rhd A$. By the definition of j, $A \not\rhd RP_{j-1}$ and hence $RP_{j-1} \in$ / $A)$. But now we are once again in the situation considered above, and so once again we know that A fails **A.4**, and RP_∞ therefore satisfies **A.5**. □

Corollary: If G has a Nash equilibrium, then $RP^*(G^\infty)$ exists.

With the preceding corollary, we complete the description of the renegotiation perfect set $RP^*(G^\infty)$. It should be noted (see section 5 below) that many previously introduced concepts of renegotiation-proofness failed to exist in a variety of standard games. Of course this feature is always undesirable to some extent, but it seems particularly so in the present context. The whole idea behind renegotiation is that the players will "cooperate" (in a non-binding sense) and pick out an optimal equilibrium from those available. If they are given this opportunity, they will certainly decide on something or other. It is thus reassuring that the only requirement for there to be a renegotiation perfect equilibrium is that there be a Nash equilibrium.

4. Finite Games

In the axiomatic definition of renegotiation perfection, arbitrary stage games were allowed. The stages were not even required to be identical; in fact the beginning of a stage was associated only with an opportunity to renegotiate, rather than with any structure inherent to the underlying game itself. However, since repeated games are a particularly natural and relevant environment, we concentrate on them. The axioms also apply to either finite or infinite horizon games, whereas in section 3 we concentrated only on infinite games. Here we address the question of renegotiation perfection in finitely repeated stage games.

Although the constructive definition was stated in terms of infinite games, it is clear how to "extend" it to the finite case: if G is to be repeated m times, begin instead with $RP_0 = [WEff\ NE(G)]^m$ and continue from there exactly as before. The same process

of seeing where it would be possible to get to, given that such-and-such a proposal has been reached so far, applies equally well[10]. Furthermore, it is clear that the theorem goes through in this case and that the axioms are still satisfied. However, there is now a third definition against which to compare renegotiation perfection. This is Pareto perfection, the version of renegotiation-proofness developed by Benoit and Krishna (1993) for finite games.

To briefly recap their definition, let $P^1 = WEff\ NE(G)$. Next let $Q^2 \subseteq \Sigma(G^2)$ be the set of all subgame perfect equilibria of G^2 which use only continuation equilibria in P^1; with two periods remaining, nothing else would be credible, given that otherwise renegotiation to some element of P^1 would take place. Since efficiency must hold at this stage as well, define $P^2 = WEff\ Q^2$. Iterate from this point, until reaching P^m, the set of Pareto perfect equilibria of G^m. Note that the construction of renegotiation perfection that was outlined in section 3, although iterative as well, is entirely different from this process in that it considers the entire length of the game at all times. Since it was designed to also apply in infinite games, this should hardly be surprising. Despite this somewhat radical difference in appearance, the two concepts are in fact identical, as we now show:

Theorem 2: For any game G, $RP^*(G^m) = P^m$.

Proof: It suffices to show that P^m satisfies **A.1-A.5**. **A.1** is clear, while **A.2** follows from the construction since $P^m \subseteq Q^m$ and thus only uses continuation equilibria which are themselves Pareto perfect in the ensuing subgame. Because $P^m = WEff\ Q^m$, no element of P^m can possibly strictly Pareto dominate another, and so **A.3** is also satisfied.

Since in a finite game the continuation subgame is never equal to the original game itself, there is never a need to compare an equilibrium with its own continuation equilibria. Therefore any element of Q^m is a [singleton] WRP set in G^m, and so it is always possible to reach the Pareto frontier of Q^m. In essence, there are no cyclicity

[10] In fact, it is easy to see that in this finite case the process always stops after only one step.

problems. Thus for any $A \in C(P^m)$, $DE(A) = (WEff\ Q^m) \setminus A \subseteq P^m$ and hence **A.4** is also satisfied by P^m.

Finally, if $P^m \triangleright A'$ then there must be some singleton $A'' \in C(A')$ which is not on the efficient frontier. In this case, $DE(A'') = WEff\ Q^m$ which is of course not confounded by anything. Then if A' satisfies **A.4**, it must be that $WEff\ Q^m \subseteq A'$ and so $A' \triangleright A''$. But this means that A' confounds itself, which contradicts **A.3**. Thus we conclude that P^m satisfies **A.5**. □

This theorem shows that renegotiation perfection can be interpreted as a true generalization of Pareto perfection to include infinitely repeated games. Since finite games don't raise any of the specters involved in comparing an equilibrium with its own continuation paths, they are conceptually considerably easier to deal with. Indeed, the concept of renegotiation-proofness is much less controversial in this case, so being able to capture the finite version is a good check on any proposed definition for infinite games.

5. Literature

The literature on renegotiation-proofness in infinitely repeated games began with the simultaneous (and neighboring) papers of Bernheim and Ray (1989) and Farrell and Maskin (1989). The latter paper introduces the concept of a weakly renegotiation-proof (WRP) set and equilibrium. It then gives a fairly full characterization of WRP sets, and of some examples in various games. Bernheim and Ray introduce an equivalent concept, which they call *internally consistent* sets. They then strengthen this to *consistency*, but this has the drawback that there may be multiple consistent sets, and they may confound each other. Thus the status quo is given extra weight; there is a lack of stationarity.

Farrell and Maskin also strengthen the WRP concept, to what they call strongly renegotiation-proof (SRP). A set is SRP if it is WRP and if it is not confounded by any WRP set. Any SRP set is clearly renegotiation-proof, since no profitable deviation is even WRP. However it is too strong a concept; why should a set necessarily be

disallowed solely because of another set which is not itself renegotiation-proof? Indeed, SRP often fail to exist. If they do exist, there is little argument against their being renegotiation-proof, so it is a relief to note that any SRP set is in the renegotiation perfect set developed in this paper. This should be clear from either of the definitions. It also implies that all of the characterization results about SRP sets apply equally to renegotiation perfect sets.

Abreu and Pearce (1991) and Pearce (1991) also study renegotiation in infinite games, but with quite a different approach. They do not require internal stability for an equilibrium to be renegotiation-proof. The reason that this is consistent is that stricter conditions are placed on proposed deviations than on current equilibria. So punishments may still be credible, since it will no longer be possible to reach the original equilibrium. This is a large departure from the stationarity assumption throughout the current paper, but it may be reasonable in some situations. Existence also poses some problems.

Bergin and MacLeod (1993) synthesize much of the literature to that point, on both the finite and infinite cases. They provide a complete axiomatic framework, which unfortunately but inevitably involves a lot of notation. There is a brief section on the explicit modeling of communication. They also introduce a new concept, *recursive efficiency*, which again gives some precedence to the status quo.

The case of finitely repeated games is treated in Benoit and Krishna (1993). They give a full characterization of the renegotiation-proof equilibria, and discuss behavior as the time horizon lengthens indefinitely. Much of the work along these lines is closely related to the coalition-proofness concept, developed by Bernheim, Peleg, and Whinston (1987). It is clear that some of the same questions arise in this model as in renegotiation models, especially if the full Pareto dominance requirement is weakened. Wen (1996) studies renegotiation in finite games with more than two players.

Blume (1994) discusses an explicit model of communication and [possibly costly] bargaining over continuation payoffs. Although much of the focus is on the finite case, the model is extended to cover the infinite case as well. Unfortunately, existence can again be a problem in the infinite case.

There is also an entire literature in contract theory on renegotiation. One of the distinctions between it and the game theory literature, of course, is that the principal is

generally choosing the contract. The effect of allowing renegotiation is therefore always negative (from the principal's point of view), even if renegotiation only occurs with the consent of both parties. This is because avoiding it becomes simply another constraint in the principal's optimal contract problem. This result is consistent with the game theory result that it may be impossible to reach the Pareto frontier if renegotiation is allowed. Another distinction is that the contracts literature often deals with information problems, both in terms of asymmetric information between the players and in terms of unrealized information (such as stochastic output processes). This latter type of problem can mean that renegotiation may actually take place in equilibrium. Finally, finite horizons (often just two periods) are the norm.

Somewhere in the background of the entire literature on renegotiation lies a model of communication. The model is almost never formally introduced, but certain results from it are assumed and constitute the driving force behind renegotiation. These results are basically as follows: before each stage of a repeated game, the players are given the opportunity to speak freely. At this point they all know what the current equilibrium directs them to play. However, if they can all believably agree to play a different equilibrium from this point on, letting bygones be bygones, then we assume that they are in fact able to switch and to play it. Believability requires three attributes – first, that it in fact be an equilibrium; second, that it Pareto dominate the current directed play (else why believe the player who is losing out?); and third, that it be renegotiation-proof itself.

It is (even *post facto*) worthwhile to briefly turn our critical eye toward these assumptions. For instance, why do we demand Pareto dominance in order to overturn the status quo? Since the original equilibrium was not a contractual obligation, it is not the case that any single participant can unilaterally revert to it, or force it to occur. At least it is not obvious that this should be the case. Perhaps, instead, some sort of bargaining would take place whenever there was an opportunity to renegotiate. In this case, the bargaining strengths of the players would determine how well they fared in the discussion process, and some players might well end up worse than when it started. Abreu, Pearce, and Stacchetti (1993) use this approach. The most obvious drawback is that the analysis becomes more difficult, and that more assumptions need to be made about the explicit nature of the process. It is possible to imagine situations in which either of the two

assumptions is more appropriate, so neither approach is superior; we start here by studying the simpler baseline case in which unanimous consent is required to shift among equilibria. Jamison (1998) is one formal model of cheap-talk that can be used to support the renegotiation-proofness literature. In a repeated game, it makes sense for the prior forecasts assumed in that model to be determined by what the current equilibrium directs. This gives potential justification for the consideration of Pareto improvements only. Naturally, this is not the only possible formal model of communication, so this is not meant to be the final word.

6. Conclusion

This paper has presented a new definition, renegotiation perfection, for ,peated games. The problem it attempts to address is of what happens if players are given the opportunity to communicate between stages of the game. It is natural to think that they will, if possible, "renegotiate" to a Pareto superior equilibrium. But given that this is the case, punishments may no longer be credible. Rational players will be able to foresee this problem and will avoid equilibria which cannot be supported. What, in the end, is the set of feasible equilibria?

Whatever this set is, it should satisfy several basic properties. First, no one player should be able to profitably deviate (Rationality). Second, since only renegotiation-proof equilibria are credible, only they should be allowed as continuation paths (Consistency). Third, an equilibrium is not feasible if it is dominated by a renegotiation-proof equilibrium (Internal Stability). Fourth, there should be no stable distinguished equilibrium which is at least as good as all renegotiation-proof equilibria (Optimality). And fifth, the renegotiation-proof set itself should always appear as a stable alternative (External Stability). Axioms four and five can be thought of as saying that you should always be satisfied if you're there, and you should never be satisfied if you're not yet there. A set is called renegotiation perfect if it meets these criteria.

Such a set may be constructively defined by a simple procedure. Beginning with the set of repeated efficient one-shot Nash equilibria as an initial reference point, ask, "If we knew we could do at least this well for sure, how much better could we do?" Repeat

this process until it eventually stops. In this fashion it is possible to work outward, one step at a time, making absolutely sure of the way as you go. The result will be self-consistent insofar as nothing was ruled out arbitrarily and everything was considered equally. This construction satisfies all of the properties listed above. It is shown that renegotiation perfect sets exist (and are non-empty). They also capture many of the properties which already appear in the literature on renegotiation. For instance, they match the Pareto perfect concept of renegotiation-proofness in finite games, and they encompass strongly renegotiation-proof sets in infinite games, when the latter exist.

There are, naturally, several directions in which to proceed from here. Topics such as bargaining power and coalitions would have to appear on any list. Possibly the most interesting and fruitful investigations, however, are in the field of stochastic games. It is in this context that actual renegotiation in equilibrium can be expected to take place, so that we may eventually have not just a theory of renegotiation-proofness, but rather an actual working theory of renegotiation. Therefore it provides an excellent testing ground to determine the implications and perhaps also the relative merits of any competing theories of renegotiation-proofness. It may be hoped that for these endeavors there is now at least a basis from which and with which to work.

References

Abreu, Dilip. Pearce, David (1991). "A Perspective on Renegotiation in Repeated Games," in *Game Equilibrium Models II*, R. Selten ed., Springer-Verlag: Munich.

Abreu, Dilip. Pearce, David. Stacchetti, Ennio (1993). "Renegotiation and Symmetry in Repeated Games," *JET* **60**(2), pp. 217-40.

Benoit, Jean-Pierre. Krishna, Vijay (1985). "Finitely Repeated Games," *Econometrica* **53**(4), pp. 890-904.

Benoit, Jean-Pierre. Krishna, Vijay (1993). "Renegotiation in Finitely Repeated Games," *Econometrica* **61**(2), pp. 303-23.

Bergin, James. MacLeod, W. Bentley (1993). "Efficiency and Renegotiation in Repeated Games," *JET* **61**(1), pp. 42-73.

Bernheim, B. Douglas. Peleg, Bezalel. Whinston, Michael (1987). "Coalition-Proof Nash Equilibria, I: Concepts," *JET* **42**(1), pp. 1-12.

Bernheim, B. Douglas. Ray, Debraj (1989). "Collective Dynamic Consistency in Repeated Games," *GEB* **1**(3), pp. 295-326.

Blume, Andreas (1994). "Intraplay Communication in Repeated Games," *GEB* **6**(2), pp. 181-211.

Farrell, Joseph. Maskin, Eric (1989). "Renegotiation in Repeated Games," *GEB* **1**(3), pp. 327-60.

Fudenberg, Drew. Maskin, Eric (1986). "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information," *Econometrica* **54**(3), pp. 533-54.

Jamison, Julian C. (1998). "Valuable Cheap-Talk and Equilibrium Selection," chapter I of this thesis.

Pearce, David (1991). "Renegotiation-Proof Equilibria: Collective Rationality and Inter-temporal Cooperation," Yale working paper.

Wen, Quan (1996). "On Renegotiation-Proof Equilibria in Finitely Repeated Games," *GEB* **13**(2), pp. 286-300.

III. Games with Synergistic Utility

1. Introduction

Frank (1987) states that, "Our utility-maximization framework has proven its usefulness for understanding and predicting human behavior. With more careful attention to the specification of the utility function, the territory to which this model applies can be greatly expanded." This is a particularly germane observation with respect to game theory. Theorists tend simply to assume that they are given the full and correct final preferences of players in a game, and that their object is to analyze the resulting strategic interactions. Where these preferences come from, and especially what differences might arise between the payoff to an individual and his or her ultimate preference over outcomes, has generally not been considered to be within the purview of game theory. However, as Frank pointed out, this necessarily limits the scope of the theory. For instance, it is probably not an exaggeration to say that all game theorists feel that no rational player should ever knowingly play a strictly dominated strategy. And yet this is exactly what robustly occurs in the one-shot Prisoner's Dilemma. The fault lies not with the theory, but with the inattention as to its application.

This paper attempts to provide a general, formal, theoretical link between the base payoffs in a game, and the resulting final utilities or preferences. The discrepancy is due to the fact that players care about the utilities of the other players in the game, e.g. due to altruism. The main reason to formalize this link is to provide applied and experimental economists with a model for this pervasive interaction, so they are not forced to come up with new (and ad hoc) formulations every time it is relevant. There is also a second reason, the stock-in-trade of theorists: to understand the process better. The jump from payoffs to final utilities goes on all the time in almost all games, so we should have a model (or, better yet, several competing models) of how it happens and what it implies.

We introduce a general definition of games with synergistic utility. Synergistic utility functions capture the idea that utility increases in one's own payoff, and may

increase or decrease in others' utilities. Sufficient technical conditions are imposed for the concept to be well-defined, but otherwise the formulation is general enough to allow maximal variety in specific applications. All players are fully rational (including being expected-utility maximizers) and no new equilibrium concepts are introduced. A specific example, the linear synergistic utility function, is introduced and analyzed in greater detail. Several applications of the theory are given, including: cooperation in the Prisoner's Dilemma, overproduction in Cournot oligopoly, extended play in the centipede game, and interior solutions in the dictator game.

The paper proceeds to section 2, in which some of the related literature, both applied and theoretical, is discussed and compared with the synergistic utility concept. In section 3, the formal model, including the central definition, is given. Next, section 4 illustrates the theory with examples both of different synergistic utility functions and of their application to different games of interest. Section 5 addresses several topics from game theory, such as incomplete information, in the context of synergistic games. Finally, section 6 briefly concludes.

2. Literature

The literature relating to altruism and interdependent preferences is wide and diverse, with each paper seemingly taking its own course. The first broad category can be considered to be the various applications of altruistic-like tendencies in specific situations. This includes, in the OLG macroeconomics literature, the famous paper of Barro (1974) on Ricardian equivalence, and the subsequent paper by Kotlikoff et al (1990) which disputes the finding. The models in these papers have "dynasties" in which ancestors care about their descendants' consumption as well as their own. Bisin and Verdier (1996) study the Prisoner's Dilemma in the context of cultural transmission, modeling altruism with the addition of a positive constant. All of these papers model altruism in one direction only, i.e. there is no feedback effect between the players. In labor economics, Rotemberg (1994) studies relations in the workplace. He determines under what conditions cooperation can be obtained and when this benefits the employer, but defines altruism only insofar as an employee's utility is the sum of payoffs to the

group. He states "Cooperative outcomes for *either* individual in the Prisoner's Dilemma obtain only when *both* individuals feel altruistic toward each other." As we shall see, this contradicts the conclusions of a synergistic utility model, in which an altruistic player may desire to cooperate even when facing a non-altruistic opponent.

Altruism within the family has been studied since Becker (1974) and his 'Rotten Kid Theorem'. He models interdependent utilities using a basic additive form. Bruce and Waldman (1990) compare this line of work to the Samaritan's Dilemma and Barro-Ricardian equivalence in a similar framework. Other work applying some degree of altruism includes Coate (1995), who studies insurance with rich and poor agents, and Collard (1975) in a general equilibrium framework. It is to be emphasized that this is only a small sample of the work that employs altruism or interrelated utilities in some form or other. In addition to the various subfields of economics already mentioned, these types of models have been used in areas ranging from law to philosophy to political science.

The second general class of papers are those on evolution and biology, which are also closely tied to the theoretical psychology literature. Frank (1987) is in this vein when he studies the commitment problem. He finds that if one can choose to be a guilty type (perhaps through an evolutionary process) and show it, one can commit credibly. This can be of great benefit, for instance in the provision of public goods. Bergstrom (1995) studies genetically predetermined behaviors, which is to say there is no free choice on the part of the players. He finds that cooperation in the Prisoner's Dilemma can be a stable outcome when players have preferences taking into account the payoffs (not the utility) of others and genetic propagation occurs through imitation of successful strategies. This is, once again, only a sample of the papers which consider this sort of evolutionary fitness paradigm. They are distinguished from the present work in that the latter is concerned with rational players in a non-dynamic setting, but it is interesting to note that some of the conclusions reached are similar.

A large number of experimental economics papers have looked at a number of different games and found results that diverge from those predicted by the basic equilibrium concepts. Dawes and Thaler (1988) study experiments with public goods, ultimatum games, and the Prisoner's Dilemma. They discuss altruism in general as an

explanation but do not suggest a model. Palfrey and Rosenthal (1988) also study public goods provision, with altruism consisting of a single lump-sum addition to payoffs (from "doing the right thing") when a player contributes. Cooper et al (1992) consider altruism in the setting of cheap talk and coordination games. One of the complications that arises from explaining the data in these and other games in this way is that it requires not only positive emotional interactions, such as altruism, but also negative interactions, such as spite (or at least retribution). For instance, it is otherwise impossible to rationalize rejected offers in the ultimatum game. Levine (1995) creates a relatively simple theory with utility linear in one's own and one's opponent's payoffs (with a possibly negative weight on the opponent). He pins down the parameters of his model by matching data on ultimatum and centipede games. He then tests the model, with some success, on public goods games and on market games. The main distinctions between his theory and the synergistic utility theory are that his players care about the payoffs, rather than the utilities, of their opponents, and that he includes a reciprocity factor, so that how a player cares about others depends on how they care about him. It turns out that much of the observed behavior can be explained without introducing this additional slight complexity, and that synergistic utilities can also rationalize some behavior (e.g. in the dictator game) that Levine's model, as it stands, cannot.

This leads naturally to the final group of related papers, those from the game theory literature. Geanakoplos, Pearce and Stacchetti (1989) introduce the concept of psychological games (and psychological equilibrium), in which utility is a function not only of actions but also of beliefs over actions. Among other things, this allows utility to depend on reactions of pleasure or anger, although only with respect to expected actions in a particular game. Players do not explicitly care about the welfare of their opponents, though as always it can in theory be incorporated into their preferences. This is an extremely powerful and all-encompassing structure, but because of this there is very little in the way of a common backbone from which to deduce or to explain results observed across a variety of different games. Rabin (1993) specializes this idea by introducing a *fairness equilibrium*, a more inherent concept which begins with a *kindness function* between the two players. Because of the special nature of the equilibrium concept, his results depend on the absolute level of the base payoffs and apply only to two-person

games. Nevertheless, he is able to draw several fairly general conclusions. Sally (1995) has a similar but somewhat more extended approach, building on the "psychological distance" between players. He develops the *sympathetic equilibrium* concept, and finds that it is sometimes possible to choose cooperation in the one-shot Prisoner's Dilemma. As in Rabin's paper, reciprocity is the starting point and again, essentially because of reciprocity, it is unclear how to extend the results to more than two players. Returning to the traditional equilibrium concepts, Bergstrom (1989) is perhaps closest to the present paper. His fun note consists of examples rather than a general model, but it does present the idea that a player's utility could be a linear function of his own payoff and the other players' utilities. One distinction with the synergistic utility concept is that he takes a fixed-point rather than a limit-point approach. He is able to explain cooperation in the Prisoner's Dilemma, although this approach does lead to some rather counter-intuitive conclusions in other situations.

3. Model

One way to introduce an altruism-like aspect in a formal game-theoretic model is to add a positive constant to payoffs following a "good" action, such as contributing in a public goods game or cooperating in the Prisoner's Dilemma. This is plausible in some circumstances, but does not capture the positive or negative benefits that a player may receive depending on the welfare of his or her opponents[1]. These can be captured most simply be adding a proportion of the opponents' payoffs to that of the player in question. This approach, however, has an inherent inconsistency: if the benefit, for instance, arises not just from doing good, but instead from being glad that a fellow player is happy, then it should be the other player's utility and not payoff that matters[2]. That is, rational players will be farsighted and will think through more than one step of the process. In general, then, final utilities will be a function of one's own payoff and of the [final] utilities of the other players.

---

[1] Throughout the paper "opponent" will be used interchangeably with "other player", whether or not the particular relationship happens to be adversarial.

[2] One caveat is that this may not apply as fully in a corporate setting.

It is not unreasonable to ask why utilities should not be a function of own *utility* and others' utilities. The short answer is that this too is self-inconsistent: preferences are utilities, they are not over utilities. As an example, consider an altruistic player with an indifferent (entirely self-concerned) opponent. The opponent will necessarily always have final utility equal to base payoff. If the altruist has utility equal to a weighted average between own payoff and the other's utility, her final utility will lie somewhere in between the two original payoffs. If, however, her utility is a weighted average between own *utility* and the other's utility, her final utility must equal that of her opponent no matter what her original payoff. In fact, it is not uncommon under these assumptions that the final utilities of both players will depend only on their altruism types and will be wholly independent of their original payoffs, an undesirable feature[3].

One final matter that should be clarified before proceeding to the formal model is the interpretation of the base payoffs. They are already objects in utility space, so they should not be thought of as monetary payoffs or profits. Rather, they can be considered to be the utility resulting from that outcome if it were in a one-person setting, or in a setting where the effects of that outcome on other players are unknown. Alternately, they are the utilities of thoughtless players, to whom it has not yet occurred that there are other players or what implications that might entail. We assume, as ever, that they already include any positive feelings from simply doing good or being fair, or on the flip side any negative feelings directly arising from an act of, say, betrayal. What they do not include are preference changes due to the realized utility of one's opponents in a particular outcome of the game[4].

We are given a game **G** with I players and payoffs v_i. A **synergism type** for a player i is an element θ_i drawn from a type-space Θ. Denote the vector of synergism types for the I players by θ. Let f be a real-valued function taking as arguments I real numbers (interpreted as welfare measures for oneself and one's opponents, respectively) and as parameters the elements of Θ. Hence $f : \mathbf{R}^I \times \Theta \to \mathbf{R}$. So f is the same for all players, but each has a separate synergism type. The base payoff for player i is $v_i = u_i^0$.

[3] The author has worked considerably with this alternate model and is more than willing to share the results of these pursuits with anyone who is interested.

[4] Note that we are assuming, as we must, the possibility of interpersonal comparison of utility.

Following the motivation above, we define $u_i^1 = f(v_i, v_{-i}; \theta_i) = f(v_i, u_{-i}^0; \theta_i)$ and in general $u_i^n = f(v_i, u_{-i}^{n-1}; \theta_i)$. At each round, players recalculate their opponents' utility levels and then adjust their view of their own utility in response, continuing ad infinitum. Finally, let $u_i^\infty(v_i, v_{-i}; \theta_i) = \lim_{n \to \infty} u_i^n$. Of course this may not exist in general.

Definition: Given Θ, a function $f : \mathbf{R}^I \times \Theta \to \mathbf{R}$ is a **synergistic utility function** if

(i) f is everywhere both continuous and strictly increasing in its first argument

(ii) f is everywhere both continuous and either strictly increasing, decreasing, or constant in each of its other real arguments

(iii) there exists $\theta_E \in \Theta$ such that for all vectors $\mathbf{v}$ in $\mathbf{R}^I$, $f(\mathbf{v}; \theta_E) = v_i$

(iv) for all $\theta \in \Theta$, $f(\mathbf{0}; \theta) = 0$

(v) for all $\theta \in \Theta$ and all $\mathbf{v}$ in $\mathbf{R}^I$, $u^\infty(\mathbf{v}; \theta)$ exists (as defined above)

In words, then, requirement (i) states that utility must be increasing in one's own payoff. Requirement (ii) asks that utility, if it is affected by someone else's payoff, always be affected in the same direction. This could be weakened, but imposes no untoward restrictions[5]. Requirement (iii) imposes that there exist a traditional game-theoretic type, i.e. one who has utility equal to own payoff regardless of the other players in the game[6]. Requirement (iv) is a moderately weak normalization that rules out adding arbitrary constants to the utility: you can't get something for nothing. And finally, requirement (v) insures that utilities exist in all cases and are well-defined.

Definition: If **G** is a game with payoffs v_i, then we say (**G**, f, θ) is a **game with synergistic utility** (a synergistic game) if it is identical to **G** except that utility is given by $u_i = u_i^\infty(v_i, v_{-i}; \theta_i)$ for all i, and f is a synergistic utility function

[5] Note, however, that it does not allow sufficient flexibility for very much reciprocity. This is by design: we see how much can be accomplished in as simple a setting as possible.

[6] *E* stands for economist or egotist, as the reader prefers.

Proposition 1: If $(\mathbf{G}, f, \theta)$ is a synergistic game, then $u_i = f(v_i, u_{-i}; \theta_i)$ for all i

The proposition says that the limit utilities, which necessarily exist, satisfy a fixed-point property. The proof follows straightforwardly from the definitions and the continuity of f. One can imagine defining synergistic utilities directly as solutions to the fixed-point equation, but this has several factors against it. First, the motivation for synergistic utilities, that players update their own welfare by taking into account the welfare of the other players, leads directly to the limit process. Secondly, the fixed-point solution may exist even if the limit does not[7]. For example, suppose that we have two altruistic players of the same type; in particular we assume $f = v_i + 2u_{-i}$ for both[8]. If $v_1 = v_2 = 1$ then the limit diverges, as would be expected (utilities go to infinity as each player gets happier and happier contemplating the situation). The fixed-point solution, on the other hand, yields $u_1 = u_2 = -1$, which appears unreasonable. Thus the limit is central to the definition, but Proposition 1 may provide a short-cut in explicit calculations.

Proposition 2: In a synergistic game, utilities u_i are continuous in payoffs $\mathbf{v}$

Proof: Let $\mathbf{v} \in \mathbf{R}^I$ have associated synergistic utilities $\mathbf{u} \in \mathbf{R}^I$. Take any sequence $\{\mathbf{v}_n\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} \mathbf{v}_n = \mathbf{v}$. We wish to show that $\lim_{n\to\infty} \mathbf{u}_n = \mathbf{u}$. If not, there exists $\varepsilon > 0$ such that $B(\mathbf{u}, \varepsilon) \cap \{\mathbf{u}_n\}_{n=1}^{\infty} = \varnothing$. From the definition of synergistic utility, $\mathbf{u} = \lim_{m\to\infty} \mathbf{u}^m$ and hence there exists M such that $d(\mathbf{u}, \mathbf{u}^m) < \frac{\varepsilon}{2}$ for all $m \geq M$. But since f is continuous, we know that $\mathbf{u}^1 = \lim_{n\to\infty} \mathbf{u}_n^1$, and iterating $\mathbf{u}^2 = \lim_{n\to\infty} \mathbf{u}_n^2$, ... so that in particular $\mathbf{u}^M = \lim_{n\to\infty} \mathbf{u}_n^M$. Thus we can choose N with the property that $d(\mathbf{u}^M, \mathbf{u}_n^M) < \frac{\varepsilon}{2}$ for all

[7] In general, of course, there may be several fixed-point solutions, while there is necessarily at most one limit point. This is another reason to choose the limit definition, although in synergistic games as defined multiplicity won't be a problem.

[8] Note that since f is simply a function of bound variables, whether we write the other players' welfares as v or u is a matter of clarity and convenience only.

$n \geq N$. But now $d(\mathbf{u}_N^M, \mathbf{u}) \leq d(\mathbf{u}_N^M, \mathbf{u}^M) + d(\mathbf{u}^M, \mathbf{u}) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$, implying $\mathbf{u}_N^M \in B(\mathbf{u}, \varepsilon)$. This is a contradiction, and so we're done. □

Proposition 2 gives us another general property of synergistic utility functions, but this is about as much as can be said in complete generality. It may be helpful at this point, in part to clarify the definitions, to consider some examples of potential synergistic utility functions. We say potential because for the moment we ignore condition (v), and we leave Θ unspecified. The most obvious is probably the linear formulation $f = av_i + \sum_{j \neq i} b^j u_j$. Here $\theta = (a, \mathbf{b})$ and $\Theta \subseteq \mathbf{R}^I$. On the other hand, $f = av_i + b(u_{-i})^2$ is impermissible, for instance, because it violates (ii). The effect of an increase in the other player's utility on one's own should be independent of the absolute levels involved. Thus, $f = av_i + b(u_{-i})^3$ is once again acceptable. Cobb-Douglas formulations, more common in macroeconomics, look like $f = (v_i)^a (u_{-i})^b$ and require that "consumptions" be non-negative. However, upon taking logs, this is equivalent to the original linear form[9]. All of the above satisfy condition (iii) by choosing $a = 1$ and $b = 0$, and satisfy condition (i) if $a > 0$. Examples of applications of these utility functions to particular games, along with an additional nonlinear formulation, are given in Section 4.

To apply the theory in a specific situation, one must choose an appropriate (f, Θ) pair and show that this pair yields a synergistic utility function. We do this now for the two-player linear case, though it is easy to extend it to more players.

Proposition 3: Let $\Theta = (0, \infty) \times (-1, 1)$ and $\theta = (a, b)$. Then $f(v_i, u_{-i}; \theta) = av_i + bu_{-i}$ is a synergistic utility function.

Proof: We have the recursive equations $u_1^n = a_1 v_1 + b_1 u_2^{n-1}$ and $u_2^n = a_2 v_2 + b_2 u_1^{n-1}$, or

$$\begin{bmatrix} u_1^n \\ u_2^n \\ 1 \end{bmatrix} = \begin{bmatrix} 0 & b_1 & a_1 v_1 \\ b_2 & 0 & a_2 v_2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_1^{n-1} \\ u_2^{n-1} \\ 1 \end{bmatrix}.$$

[9] Note that we cannot then independently choose the cardinalization for taking expected utilities.

We may write this as $u^n = \mathbf{M}u^{n-1}$, and hence $u^n = \mathbf{M}^n u^0$, where $u^0 = \begin{bmatrix} v_1 & v_2 & 1 \end{bmatrix}'$. Then multiplying out the powers of $\mathbf{M}$ yields

$$\mathbf{M}^{2m} = \begin{bmatrix} (b_1b_2)^m & 0 & (a_1v_1 + b_1a_2v_2)\sum_{i=0}^{m-1}(b_1b_2)^i \\ 0 & (b_1b_2)^m & (a_2v_2 + b_2a_1v_1)\sum_{i=0}^{m-1}(b_1b_2)^i \\ 0 & 0 & 1 \end{bmatrix}.$$

But by assumption $|b_1b_2| < 1$ so

$$\lim_{n\to\infty}\mathbf{M}^n = \frac{1}{1-b_1b_2}\begin{bmatrix} 0 & 0 & a_1v_1 + b_1a_2v_2 \\ 0 & 0 & a_2v_2 + b_2a_1v_1 \\ 0 & 0 & 1-b_1b_2 \end{bmatrix}.$$

Now $\lim_{n\to\infty} u_i^n$ is simply the i^{th} row of the 3rd column of the matrix above so it too exists (and in fact this gives an explicit formula for it). Naturally, this is the same solution one would find from solving the system of two fixed-point equations. It is clear that conditions (i)-(iv) also hold. □

Note that the perverse example mentioned earlier, which had $b = 2$, is not allowed in this scenario. Nonlinear synergistic utility functions will have their own requirements for Θ [10]. Turning to another question that can be answered given a specific synergistic utility function, it is well-known that positive linear transformations of any player's payoffs leaves the strategic structure (i.e. the preferences over final outcomes) of a game unaffected. This result carries over to synergistic games as much as possible (it is clear that multiplying only one player's payoffs by some constant may substantively change utilities in an interdependent setting).

Proposition 4: In a linear synergistic game, preferences over outcomes are unaffected if

(a) all player's payoffs are multiplied by the same positive constant, or

(b) any or all players have a constant added to their payoffs

[10] For example, we might imagine that more generally one would require the derivative of f with respect to opponent's utility to be bounded by 1.

Proof: (a) Since f is linear in v_i (or in fact more generally whenever f is homogeneous of degree one in v_i), utilities all along the limiting sequence, and hence also final utilities, will also be multiplied by this constant. So then, by the standard result, preferences remain the same.

(b) Adding a constant to one player's payoffs affects all players, but only to the extent of adding some constant to each of their payoffs. Although this constant may be different for each player, it is the same for a given player across his or her outcomes. This is clear from the explicit formulas found in the proof of Proposition 3. But now, once again, the standard result applies. □

Although this result doesn't hold in general for all synergistic games, it will hold in other particular settings. We now turn our attention to illustrating the theory with a spectrum of examples.

4. Examples

The proof of the pudding lies in the taste, and the believability of synergistic utilities lies in its potential applications. For the time being, we confine ourselves to the linear synergistic utility function analyzed above, $f = av_i + bu_{-i}$. We first define three types of players to give some idea of the range of possibilities. Although unnecessary, it is convenient to choose them such that $a + |b| = 1$; this keeps the magnitude of the utilities directly comparable to those of the base payoffs[11]. The first type is the one required by part (iii) of the definition, $\theta_E = (1,0)$. This type always has final utility equal to base payoff regardless of the other players. The second type is an altruist, denoted by S for socialist: $\theta_S = (\frac{1}{2}, \frac{1}{2})$. This type approximately treats the two players equally. Finally, we define an unfriendly type: $\theta_J = (\frac{1}{3}, -\frac{2}{3})$. In the game theory literature, this general type has been called spiteful, but that is perhaps too strong a condemnation for these

[11] Most of the previous literature has instead chosen (in its own context) $a = 1$.

preferences. Rather, this player simply enjoys doing better than his or her opponent; the notation is thus Jones, for "keeping up with the Joneses"[12]. Note that since we apply the theory to single games, it is possible to switch types over time or in differing situations or against different players. The model does not require them to be intrinsic. Also, it is fairly easy to see how to come up with multi-player analogues for these types.

The basic Prisoner's Dilemma can be written as:

		Player 2 C	Player 2 D
Player 1	C	0,0	-9,3
	D	3,-9	-6,-6

Here *C* stands for cooperate and *D* for defect, as usual. Of course the unique Nash Equilibrium is (*D*,*D*). If two type *E*'s (economists) play against one another, the payoffs remain as they started and the game is unchanged. So the unique NE is also the same. We next consider an economist player 1 opposing a Jones player 2. *E*'s utilities are the same as ever, while *J*'s may then be calculated using *f* (it takes only one step in this case). We arrive at the following game form:

		type J C	type J D
type E	C	0,0	-9,7
	D	3,-5	-6,2

The unique NE is again for both players to defect. What is interesting, however, is that this outcome is no longer Pareto inefficient, as it was previously. The economist is so unhappy that it makes the Jones player happy. This depends, of course, on the exact payoff structure and type of player 2, but holds over a wide class. Consider next a socialist player 1 against a Jones type:

[12] A similar Jones type appears in the macroeconomics consumption literature, so this is conceivably an example of micro keeping up with the macro Joneses.

		type J: C	type J: D
type S	C	0,0	-3,3
	D	0,-3	-3,0

This game now has two pure NE, in both of which type *J* defects (unsurprisingly it turns out that types *E* and *J* always defect). Type *S* is completely indifferent, and is thus willing to cooperate. Of course this is knife-edge; types near to *S* will be pushed in one direction or the other, some of them always cooperating. The (*C*,*D*) equilibrium is [weakly] Pareto efficient in this case. We now change player 2 to a type *S* as well:

		type S: C	type S: D
type S	C	0,0	-5,-1
	D	-1,-5	-6,-6

Cooperation is a dominant strategy here for both players; it is also the optimal outcome in the game. This is the stereotype of altruistic cooperation in the Prisoner's Dilemma. The final combination of players that we consider is when player 1 is a type *E* once more:

		type S: C	type S: D
type E	C	0,0	-9,-3
	D	3,-3	-6,-6

The unique and strict NE is (*D*,*C*). The surprising observation here is that it requires less inherent altruism to cooperate with a type *E* than with a type *S*[13]. This result can be explained by noting that defection hurts a type *E* opponent more than it does a type *S*

[13] Contrast this once again with the quote from Rotemberg (1994) in Section 2.

opponent (who is consoled by the fact that one's own payoff has been improved). Hence a type S will have a stronger incentive not to defect when playing against a type E. Recall that we have tried to put aside any issues of reciprocity.

Turning next to an example of a continuous game, we consider Cournot duopoly. In the simplest case with linear unit demand and zero marginal cost, price $p = 1 - q$, where q is the total quantity produced. Payoffs are simply net profits, so $v_i = q_i(1-q)$. The unique Nash Equilibrium with standard (i.e. type E) players is for both to produce at $q_i = \frac{1}{3}$. It is plausible, however, to model the firms as type J. Perhaps it is a small market so that profits themselves are not important but beating the rival firm is critical for advertising. Or perhaps the managers are paid with yardstick competition incentives, so again what is important is to do better than the other firm. The symmetric equilibrium in this case is that both produce $q_i = \frac{3}{7}$. In the end of course neither firm actually does any better than the other, but each is willing to overproduce ("sacrificing" profits) in order to try to do so. Note also that this is much closer to the zero profit outcome of Bertrand competition, and in fact it converges to that outcome as the firms get more and more extreme in the Jones direction.

Experimental game theory has included extensive work not only with the Prisoner's Dilemma but also with other games such as ultimatum, dictator, centipede, and public goods games. As in the case of the Prisoner's Dilemma, the results are often quite disparate from those predicted by standard theories. For instance, no positive quantity should ever be rejected in an ultimatum game, yet this is often observed in experiments. This outcome can be explained using synergistic utilities: types similar to Jones will reject all offers up to some level (which will depend on the exact type chosen and on the type of the opponent). Of course altruism alone, without some sort of negative analogue, can never rationalize these rejections. Recall that it is possible to extend the theory to include reciprocity if desired, so a player's type need not be constant. As has been documented previously (see section 2), altruism can explain extended play in a centipede game or contribution in a public goods game. The point is that a simple theory, such as synergistic utilities, is sufficient to do this.

In the so-called dictator game, player one simply decides how to divide an amount of money (typically around \$10 in experiments) between him- or herself and an often anonymous opponent. Player two has no action other than to accept the split as dictated. Traditional equilibrium concepts predict that player one should keep the entire amount, and previous models of altruism have not altered this prediction. For instance, continuing with the types as defined above, if an altruistic type S opposes another type S, the optimal action is still to give nothing away. No linear model can predict an interior solution, although in practice this is what the data clearly support. We turn, then, to a nonlinear synergistic utility function. For simplicity we assume that player two is a type E, so that as always $u_2 = v_2$. For player one, we assume the altruistic formulation $u_1 = \sqrt{v_1 u_2}$. In this case the optimal allocation is an even split, i.e. \$5 for each player. This outcome is occasionally, though rarely, observed in experiments. If we assume instead the slightly less magnanimous utility $u_1 = \frac{1}{2} v_1 + \frac{1}{2}\sqrt{v_1 u_2}$, then we find $v_1^* \approx \$8.54$. In fact this agrees remarkably well with the observed average division. Naturally, this is meant only to illustrate the potential applicability of the theory, in addition to the fact that nonlinear functions do not simply provide generality but in fact may be necessary in practice.

5. Topics

Despite the fact that the game structure remains the same in synergistic games (only the payoffs have changed), there are several topics that take on new meaning in this context. For instance, cooperative games with transferable utility will be difficult to analyze since some players may actually prefer a smaller total surplus to divide (think of the type J above). As another example, evolutionary game theory has been a popular subject of study recently. In the present setting, it is possible to discuss the evolutionary strengths not just of different strategies but also of different synergistic types. What is unclear, however, is what to use as a measure of reproductive fitness. One could argue that players with the highest welfare (final utility) will be the most productive and successful. On the other hand, it may be that the determination of success is made by physical rather than mental well-being, so that base payoffs (food or money leading to direct consumption) should enter the calculation of the dynamics. A player might be

happy that his or her fellows do well, but this does not necessarily grant an increased chance of survival. The appropriate measure may depend on the particular situation. In the Prisoner's Dilemma example of section 4, note that altruistic players, type S in the notation there, fare relatively poorly under either system.

A related consideration, though more in the mode of full rationality, is the idea of segregation. Since players are of different types, they may prefer to play against one type of opponent rather than another, and thus selectively associate. Of course, they may not have the opportunity to make this choice, but if they do then it has long-term welfare (and hence possibly evolutionary) implications. Returning once again to the Prisoner's Dilemma example of the previous section, note that while types E and S always prefer an altruistic type S opponent, this is not necessarily true of type J players, who like to play type E's (since the latter end up so unhappy). So a plausible scenario is that S types play against themselves, while J's and E's pair off against one another. This leaves the self-centered economist types quite unhappy; their only hope is to run across extremely altruistic players, who will actually like to make them happy by cooperating (in effect, happily sacrificing themselves). Recall that all players are fully utility maximizing at all times.

There is no doubt at least some element of reciprocity in almost all human interactions. Synergistic utilities, as defined, make no account for this; a player's degree of altruism is independent of the attitudes of the other players. The work of Rabin (1993) and Sally (1995) depend explicitly on these added interactions, and similar constraints can be added to synergistic games. One method would be to require that players enter a game with their own individual synergistic type θ, but that then all of the players play the game using the average θ of the group (if Θ is such that this has meaning). Another possibility is to add a reciprocity player, type R, who takes on the θ of whomever he or she is playing. As always, this is difficult to implement with more than two players. The point is that altruism, jealousy, and so on make sense independently of any reciprocity arguments, so the simplest models of such behavioral tendencies will not include them as a building block. They may however be necessary in order to fully explain either our own introspective assessments or all empirically observed behavior.

Finally, games with incomplete information take on an added dimension if there is also the possibility of synergistic types. There is no reason in general to assume that all players know the type of each of their opponents, synergistic or otherwise. Fortunately, the entire game-theoretic apparatus developed to analyze this eventuality is still perfectly applicable. In particular, the Bayesian equilibrium concepts apply just as well here. As synergistic types are certainly payoff relevant, signaling will be an important component to playing extensive-form synergistic games. It may or may not be beneficial for a player in a given situation to reveal his or her synergistic type (consider, for instance, the discussion of segregation above). In fact, incomplete information aspects of synergistic games seem to be perhaps the most fruitful line for future theoretical research using this model.

6. Conclusion

Game theorists assume that the payoffs in a game indicate true preferences, which is to say that they already take into account welfare interactions between the players. But often in real-life situations, the only information available is about base payoffs, e.g. profits for firms or monetary payoffs in an experimental setting. It is useful to have a specific model of altruism and other emotional aspects in order to link these payoffs to the ultimate utilities in a game. The concept of synergistic utilities attempts this, by providing a simple framework in which to address these concerns in various applied contexts. Each player's utility is a function of his or her own payoff and of the other players' utilities. Standard equilibrium concepts are sufficient, and since the process is a transformation of payoffs only, the theory can be applied to arbitrary games, with any number of players. One special case, a linear formulation, was given and analyzed in more detail. Examples, such as how both cooperation in the Prisoner's Dilemma and positive gifts in the dictator game can be rationalized, followed.

The main distinction between the present work and previous literature lies in the simplicity of synergistic games. There is nothing new imposed on the game structure or analysis, since the only change made is in the numerical values of the payoffs. Nor is an idea of reciprocity inherent or necessary to the model. Nevertheless, many observed

behaviors can be explained within this paradigm. Note in particular that standard theories have done exceptionally well in predicting behavior in market situations. In these games, by definition, a player cannot influence the payoff of anyone else in the game (or at least is of this impression). Hence a player with synergistic utility will behave exactly as a standard player would, a robustness check on the theory. Surely there will be more such checks to come.

References

Barro, Robert J. (1974). "Are Government Bonds Net Wealth?" *J Pol Ec* **82**(6), pp. 1095-117.

Becker, Gary S. (1974). "A Theory of Social Interactions," *J Pol Ec* **82**(6), pp. 1063-93.

Bergstrom, Theodore C. (1989). "Love and Spaghetti, The Opportunity Cost of Virtue," *JEP* **3**(2), pp. 165-173.

Bergstrom, Theodore C. (1995). "On the Evolution of Altruistic Ethical Rules for Siblings," *AER* **85**(1), pp. 58-81.

Binmore, Ken (1994). *Game Theory and the Social Contract Volume 1: Playing Fair*, MIT Press: Cambridge, MA.

Bisin, Alberto. Verdier, Thierry (1998). "On the Cultural Transmission of Preferences for Social Status," *J Pub Ec*, forthcoming.

Bruce, Neil. Waldman, Michael (1990). "The Rotten-Kid Theorem Meets the Samaritan's Dilemma," *QJE* **105**(1), pp. 155-65.

Coate, Stephen (1995). "Altruism, the Samaritan's Dilemma, and Government Transfer Policy," *AER* **85**(1), pp. 46-57.

Collard, David (1975). "Edgeworth's Propositions on Altruism," *EconJ* **85**, pp. 355-60.

Dawes, Robyn M. Thaler, Richard H. (1988). "Anomolies: Cooperation," *JEP* **2**(3), pp. 187-97.

Frank, Robert H. (1987). "If *Homo Economicus* Could Choose His Own Utility Function, Would He Want One with a Conscience?" *AER* **77**(4), pp. 593-604.

Geanakoplos, John. Pearce, David. Stacchetti, Ennio (1989). "Psychological Games and Sequential Rationality," *GEB* **1**(1), pp. 60-79.

Kotlikoff, Laurence J. Razin, Assaf. Rosenthal, Robert W. (1990). "A Strategic Altruism Model in Which Ricardian Equivalence Does Not Hold," *EconJ* **100**, pp. 1261-68.

Levine, David (1995). "Modeling Altruism and Spitefulness in Experiments," UCLA working paper.

Palfrey, Thomas R. Rosenthal, Howard. (1988). "Private Incentives in Social Dilemmas: The Effects of Incomplete Information and Altruism," *J Pub E* **35**(3), pp. 309-32.

Rabin, Matthew (1993). "Incorporating Fairness into Game Theory and Economics," *AER* **83**(5), pp. 1281-302.

Rotemberg, Julio J. (1994). "Human Relations in the Workplace," *J Pol Ec* **102**(4), pp. 684-717.

Sally, David (1995). "On Sympathy," Cornell working paper.