

# Commitment and Weakness of Will in Game Theory and Neoclassical Economics

By Roger A. McCain

A decision-maker may be said to be rational if his decisions fulfill two conditions: first, they are consistent with his preferences in the light of his beliefs, and second, his beliefs are consistent with the available evidence. Either of these conditions may be fulfilled exactly or only approximately. Selten writes (1975 p. 320) "... game theory is concerned with the behavior of absolutely rational decision makers whose capabilities of reasoning and memorizing are unlimited ...." The same is true of neoclassical economics. But it is not an open question whether this "concern" is descriptive of actual human decision-makers. There is ample evidence to reject it. No rational decision-maker, choosing among theories, can choose one that is based on absolute rationality. Real human rationality is bounded<sup>1</sup>, and although boundedly rational decisions may approximate those that would be made by one of Selten's decision-makers, theories based on the postulate of absolute rationality can be quite misleading. (Akerlof and Yellen 1985, Akerlof 2001)

Nevertheless, absolute rationality retains some intellectual interest. On the one hand, the "possible world" inhabited by absolutely rational beings is a possible world that derives its interest from our interest in our own (bounded) rationality and the implications of our attempts to improve it. If we adopt a model of boundedly rational learning,

---

<sup>1</sup> Simon 1955, 1995. It should be stressed that bounded rationality (artificial intelligence) is a theory of rational, not irrational behavior.

absolute rationality can define attractors and stable points that may (or may not) be observed in the boundedly rational learning process. In a related way, noncooperative game theory (at least) can be a useful problem-finding tool. (McCain, 2009) Finally, to the extent that individual boundedly rational decisions do approximate absolutely rational ones, difficulties or ambiguities in the meaning of absolute rationality carry over to real rational behavior. This paper is concerned with one such ambiguity.

In “The Intimate Contest for Self-Command” (1980) and other publications, Thomas Schelling explored the implications of weakness of will for rational choice and behavior. Noting that weakness of will creates within an individual a conflict not altogether unlike the conflict of objectives assumed in noncooperative game theory, he drew on game theory to point out strategies by which an individual might overcome weakness of will and carry out commitments that could create such a conflict. Schelling subsumed weakness of will to *bounded* rationality, but regarded it as very widespread (if not universal) in imperfect human behavior. His examples make it clear that weakness of will is an obstacle to commitment. But what is the relation of absolute rationality to weakness of will? This is the ambiguity in the concept of absolute rationality with which we will be concerned.

- i. Weakness of Will

The issue does not arise in most of neoclassical economics because neoclassical economics excludes, by assumption, many of the circumstances in which it might arise. Consider the case of intertemporal inconsistency in choice. We adopt the neoclassical convention of expressing time preference by a discount rate. Most economic literature assumes that this discount rate per unit time is the same regardless of the delay before the

payment is made. This assumption of a uniform rate of time preference has no basis in empirical observation, but is made in order to reconcile the theory of rational choice, as it is understood in modern economics, with the assumption of time preference. The difficulty is that a non-constant rate of discount can result in what are called intertemporal inconsistencies in decision-making. What this means is that a rational, maximizing decision maker would make one decision at one point of time, but at a later point of time would rationally prefer the alternative he has initially, rationally rejected. (There has been some recent research on alternatives to constant rates of time preference, such as hyperbolic discounting, but it has been directed to a different issue.)

This can be illustrated by an example. Suppose that the decision-maker discounts any prospect delayed by more than six months at 18%, but that his rate of discount for prospects delayed six months or less is zero. Now the decision maker must choose at  $t_0$  between two alternatives. Alternative Alt 1 is a payment of \$5000 at  $t_0+1$  year. Alternative Alt 2 is a payment of \$10000 at  $t_0+5$  years, but Alt 2 has a cancellation clause: at any time during the first year, for a cancellation fee of \$100, the decision-maker can cancel his decision for Alt 2 and receive the payment of \$5000 at  $t_0+1$  year.

At  $t_0$ , the discounted present values are

Alternative Alt 1	\$4,237
Alternative Alt 2	\$4,371

Accordingly, the decision-maker chooses alternative Alt 2. However, at  $t_1 = t_0+6$  months and one day, the payoff for alternative Alt 1 is less than six months away, and so is not discounted, and is valued at \$5000. To obtain this payment, however, the decision maker must pay the cancellation fee of \$100. The net values discounted to  $t_1$  are

Alternative Alt 1                    \$4,900

Alternative Alt 2                    \$4,748

Therefore, the rational decision-maker reverses his decision.

This is a one-person game. Suppose we express these decisions as plans of action for the successive stages like the pure strategies as understood by von Neumann and Morgenstern<sup>2</sup>. (2004 originally published 1944) The decision maker has three pure strategies:

- 1) Choose Alt 1.
- 2) Choose Alt 2, then do not cancel.
- 3) Choose Alt 2, then cancel.

The payoffs of these strategies, discounted to  $t_0$ , are

- 1) \$4,237.
- 2) \$4,371.
- 3) \$4,127.

Why, then, does our rational decision-maker not simply choose strategy 2 and stick with it? Suppose that the decision-maker has a weak will, in Schelling's sense, and knows that he does. Then he can anticipate that if he chooses Alt 2, he will indeed cancel it after six months and in fact carry out strategy 3. Because of his weakness of will, strategy 2 simply is not available to him. That being so, in the spirit of Ulysses and the Sirens, (note Elster 1977) the rational but weak-willed decision-maker will choose strategy 1 and alternative Alt 1.

---

<sup>2</sup> "Imagine now that each player ..., instead of making each decision as the necessity for it arises, makes up his mind in advance for all possible contingencies ... We call such a plan a *strategy*." (von Neumann and Morgenstern, 2004, p. 79).

Strength of will may be very rare, but I can say from my own experience that it does exist. I recall that in 1960 my late father gave up smoking, after having smoked about 40 cigarettes a day for two decades. Many people have quit smoking, and I would say that all who succeed possess strong wills, even if they have taken steps (like avoiding places where they were in the habit of smoking) to reduce temptation. My father, however, carried an unopened pack of cigarettes in his pocket every day for a year after he quit. He explained to me that he wanted to prove to himself that he was not a slave to the habit – and he never started again.

This is not to say that intertemporal inconsistency does not exist. No doubt a strong-willed decision maker, having chosen strategy 2, will feel some subjective tension in the nature of regret or temptation during the time interval  $t_1$  to  $t_2=t_0 + \text{one year}$ . Does rationality require him to act on the temptation? Well – perhaps it does. Selten writes (1975 p. 328) that the decision maker “... should not be guided by his payoff expectations in the whole game but by his conditional payoff expectations,” at the moment the decision is made.

Weakness of will may also be a factor in interactive decisions. Consider the following two-person game in extensive form, shown as Figure 1. All decisions are close enough together in time that there is no need to discount payments to present value.

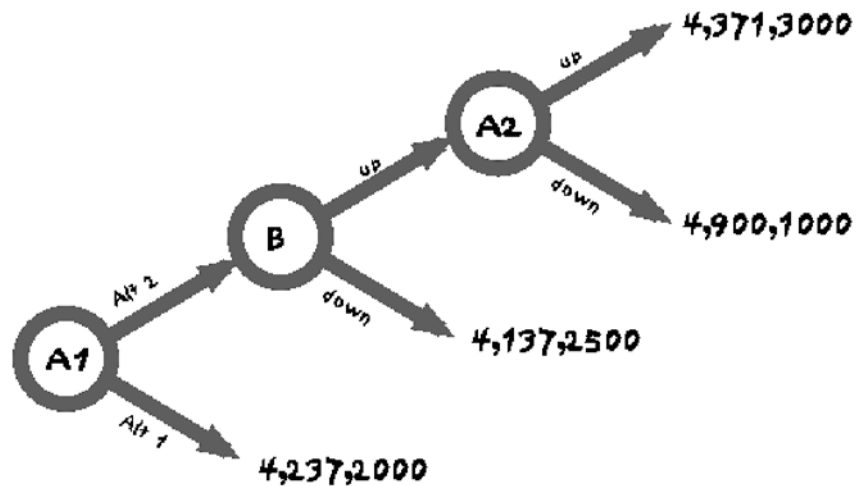


Figure 1. Two-Person Game in Extensive Form

First we note that the subgame perfect equilibrium for this game is for decision-maker A to choose alternative 1 for a payoff of 4,237. However, when we express this game in terms of von Neumann-Morgenstern contingent strategies, we have, for decision-maker A,

- 1') Choose Alt 1.
- 2') Choose Alt 2, then, if B chooses up, choose up.
- 3') Choose Alt 2, then, if B chooses up, choose down.

and for decision maker B,

- 4') If A chooses Alt 2 then choose up
- 5') If A chooses Alt 2 then choose down

If decision-maker A chooses strategy 3', then decision-maker B's best response is strategy 5', while if decision-maker A chooses strategy 2', and B knows this with certainty, then B's best response is strategy 4'. Taking this into account, the payoffs to be expected from these strategies would seem to be

1') 4,237

2') 4,371

3') 4,137

4') 3000

5') 2000

This being so, we ask again, why does decision-maker A not simply choose strategy 2'? There are two possibilities: i. Decision-maker B believes that decision-maker A has a weak will, and will not carry out strategy 2 but, having arrived at decision-point A2, will choose down. Decision-maker B therefore chooses strategy 5'; and this is known to agent A, who then chooses strategy 1' as his best response to strategy 5'. Thus, it seems, the subgame perfect equilibrium can be necessary because of the belief that A has a weak will. ii. The second possibility is that B believes A is dishonest and opportunistic and will choose "down" at decision point A2 regardless of any protestations to the contrary. Thus, the subgame perfect equilibrium can be necessary because of the belief that A is dishonest. But suppose that decision-maker A has a strong will, that is, a capability to choose strategy 2' and stick to it despite the temptation to choose "down" at decision point A2; and suppose that this is known to decision-maker B. Suppose decision-maker A also is honest, and this, too, is known to decision-maker B. Thus, decision-maker needs only announce "on my honor, I am choosing strategy 2'," and then B's rational decision is for strategy 4', and the cooperative solution Alt 1, up, up results.

Notice that, so far as A's decisions are concerned, decision sequence 1', 2', 3', with its payoffs, is identical to 1, 2, 3. The major difference is that it is decision-maker B's belief that A has a weak will, rather than the fact that A's will is weak, that puts Alt

2, up, up out of A's reach. Supposing B to be rational, what basis might he have for that belief? One possibility is that we *define* rationality as maximization *constrained by weakness of will*. Then we need only apply common knowledge of rationality to induce B's belief in A's weakness of will. I submit that this is indeed the concept of rationality noncooperative game theory and in neoclassical economics. In what follows choices that maximize payoffs subject to the constraint of weakness of will be called "perfectly rational choices," not because their outcomes are perfect (as the example shows) but because it is rationality in this sense that defines subgame perfect equilibrium.

But common knowledge of *perfect* rationality is not the only possibility, and we need to consider others. First consider the possibility that B believes A is dishonest. Then B will not believe any assertions by A that he will choose "up" at decision point A2 and accordingly B chooses strategy 5'. But i) A's honesty is of concern to B only if B believes A has strong will. If B believes A has a weak will then B's decision will not be affected by the further knowledge that A is honest or dishonest. ii) A can benefit by acting dishonestly only if B believes both that A has strong will and that A is honest. iii) Accordingly, we must consider a 4-step game in which A's decision whether to act honestly or dishonestly is the first stage. If A has a strong will he can commit himself to one or the other and carry out the commitment. iv) However, if B believes A has chosen to act honestly, then A's best response is dishonesty. v) Therefore, this first stage requires a mixed-strategy solution. vi) Since B is rational, he will be aware of this and will accordingly estimate the payoffs of strategies 4' and 5' as expected values reflecting the optimal mixed strategy for A, which is to act honestly with probability 2/5. Thus, B's belief that A will be dishonest with probability 1 is either irrelevant or irrational.

It seems that B's behavior, as assumed in subgame perfect equilibrium theory, can be rational only if B believes that weakness of will is a common trait of all human beings. This in turn can be considered a rational belief only if i) it is true, or ii) B's experience has been so idiosyncratic that it seems to B that the belief is true, although B is mistaken. We can eliminate ii) as inappropriate to be the basis of a general theory, and conclude that: for subgame perfect equilibrium theory, universal weakness of will is a necessary assumption. If both weakness of will and perfect rationality are common human characteristics, then there is little point in distinguishing between them: in that case we lose nothing by failing to distinguish between rationality and weakness of will. But the results of such an identification can be rather peculiar:

The results of the example of intertemporal inconsistency and of the two-person game from Figure 1 can both be stated in the following way:

- i) Define rationality as perfect rationality.
- ii) Suppose decision-maker A in fact adopts strategy 2 (or 2') and carries it out.
- iii) As a result of this choice, decision-maker A is better off.
- iv) Decision-maker A has acted irrationally.

Stated in just that way, perfect rationality is not a very intuitively appealing concept of rationality.

ii. Commitment

Game theory offers an alternative to perfect rationality. It is widely understood that game theory comprises two distinct traditions: noncooperative and cooperative game theory. Both are founded<sup>3</sup> on the assumption of absolute rationality. It is widely supposed that the difference between the two is a matter of assumptions about institutions: for cooperative game theory, agreements to correlate strategies are enforceable, while for noncooperative game theory they are not. However, a study of the origins of these two traditions suggests that there is more to it than that. Consider the following quotations: from von Neumann and Morgenstern: “As soon as [the game] involves three or more persons, the game is ruled by coalitions ... it is dependent on each player *ascribing to his opponent ... the desire to inflict a loss* rather than to achieve a gain.” (2004 originally published 1944, p. 507, p. 559) And from Nash: “Supposing A and B to be rational beings, it is essential for the success of the threat that A be *compelled to carry out his threat* T if B fails to comply. Otherwise it will have little meaning. For, in general, to execute the threat will not be something A would want to do, just of itself.” (Nash, 1953, p. 130; italics added in both cases). In case the contradiction between these two views is not apparent, here is the context: von Neumann and Morgenstern are arguing that any coalition (less than the whole) will face a unified opposition coalition that will threaten them with the greatest loss the opposition coalition can arrange, and this threat will be believed regardless of the cost to the opposition coalition. Nash is speaking of bargaining between two parties (which might be the opposed coalitions) and tells us that no threat

---

<sup>3</sup> In recent decades a growing literature within game theory has relaxed these assumptions, and it has become conventional to distinguish classical from behavioral game theory much as we distinguish neoclassical from behavioral economics. (Camerer, e.g.)

will be believed unless it is “something A would want to do, just of itself.” And both of these views are understood as rational behavior.

They differ, in particular, on the possibility of *commitment*. For Nash, commitment and rationality are incompatible; for von Neumann and Morgenstern, rationality subsumes commitment as one of its aspects. Cooperative game theory arises from the reasoning of von Neumann and Morgenstern, and follows their assumptions on rationality, while noncooperative game theory arises from Nash’s ideas and follows his. For cooperative game theory, we must amend Selten’s phrase to read “... [cooperative] game theory is concerned with the behavior of absolutely rational decision makers whose capabilities of reasoning and memorizing [and strength of will to keep their commitments] are unlimited ...” In what follows, rationality in this sense will be termed “ideal rationality.”

I am using the word “commitment,” here, in a different sense than Sen does in his famous essay “Rational Fools.” (1979) For Sen, commitment is one of two interpretations of unselfish behavior and can occur only when there is at least an expected-value sacrifice by the decision-maker of her own interest (pp. 327, 330), while I will use the word to refer to cases where, by making a commitment, the person increases her own payoffs. Nevertheless the two uses are related, particularly as they both may be limited by weakness of will. (p. 340) In his essay, Sen distinguishes between *commitment* and *sympathy* as bases for non-self-serving behavior. This distinction extends to the cases considered in this paper.

### iii. Empirical Issues

In any case, we have ample evidence that the noncooperative solution most relevant here, subgame perfect Nash equilibrium, does not agree well with empirical studies. To assess the implications of this evidence for ideal and perfect rationality would require a large review of the literature, which is beyond the scope of this paper. We may, however, consider one example: the widely studied ultimatum game.

The Ultimatum Game is a two-person game along the following lines: the two agents may be able to share a fixed amount, such as \$100. The first agent, the proposer, suggests a payment to go to the second agent, the responder. If the responder accepts the payment, he receives it, and the balance is paid to the proposer. However, if the responder rejects the payment, neither agent gets anything. The only perfect Nash equilibrium is one in which the proposer makes the smallest possible positive offer and the responder accepts it. However, experimental evidence disagrees with this prediction. If the proposer makes a very small offer, the responder is sometimes observed to reject the proposal despite sacrificing the small positive payment. Moreover, offers are often more than the minimum needed to avoid a rejection, and 50-50 offers are fairly common<sup>4</sup>.

This evidence bears against an hypothesis of perfect rationality. What does it say about the empirical reliability of the hypothesis of ideal rationality? The experimental studies that cast doubt on noncooperative game theory and thus on perfect rationality do not, for that reason, necessarily support any particular alternative theory. Indeed, it is difficult to tease empirical predictions for those games out of cooperative game theory,

---

<sup>4</sup> E. g. Guth , et. al., 1982, Henrich, et. al. 2005, and note also Roth (1995), Stanley and Tran (1998), Roth et. al. (1991), Andreoni and Blanchard (2006), Oosterbeek et. al. (2004)

because 1) there are several solution concepts in cooperative game theory, which may not agree in a particular case, and 2) the experiments were designed to test noncooperative hypotheses, and thus deliberately exclude conditions that may be assumed in cooperative game theory, such as communication and prior agreement on a common strategy and (to use Kuhn's, 1953, terminology) perfect recall.

It would seem that an agent who is capable of committing himself to a strategy such as 2 and 2' in the example would also be able to commit himself to play according to a rule, provided that such a commitment would increase his payoffs on the whole. Moreover, a rule such as "always punish opportunism" is one that might offer such a reward, especially if the agent could acquire a reputation for it; and rejections of low offers in the Ultimatum Game would be a particular case of that general rule. But it will not be enough to say that individuals will choose according to some rule. How will the rule be chosen<sup>5</sup>?

A parallel problem arises with rule-utilitarianism,<sup>6</sup> which holds that a person *ought* to act according to some rule, but that the rule should be one that produces the greatest utility on the whole. But that does not quite resolve the question. In choosing a rule to follow, one might ask "What would be the consequences of rule K, if everyone were to act according to rule K?" The rule with the best (total utility) consequences would then be the rule chosen. This has been called "utilitarian generalization." (Fuchs in West, p. 144) Alternatively, one might ask "What would be the consequences of rule K,

---

<sup>5</sup> Aumann's (1959) supergame analysis is suggestive. It is explored, and distinguished from the approach adopted here, in Appendix 2.

<sup>6</sup> On utilitarianism, see Sen and Williams, West.

if I were to act according to rule K in the given circumstances, regardless what others may do?” Often, the answers will be different.

Of course, utilitarianism is an ethical theory, which demands that the decision-maker consider the total utility (total payoffs in the game), and here we are concerned with self-regarding decisions that consider only the payoffs to the decision-maker himself. Nevertheless, the same issue arises. Consider, for example, a social dilemma. A decision-maker who asks the second question will ask himself, “Given the strategies chosen by my counterpart, what are the consequences for me of playing cooperatively?” But this leads directly back to the noncooperative dominant strategy equilibrium. On the other hand, if the decision-maker asks himself, “What are the consequences for me if everyone plays cooperatively rather than noncooperatively,” then the answer is that cooperative play has the better consequences. Similarly, self-regarding utilitarian generalization leads to the solution that maximizes total payoffs in Figure 1, and will, in general, lead to a Pareto-efficient set of decisions. Tentatively, then, we associate cooperative play, in a game that does not permit explicit agreement, with commitment to a rule chosen by self-regarding utilitarian generalization. Strong-willed agents who act according to rules chosen in this way will be called “nomists” in what follows.

Note that an ideally rational agent may not inevitably act according to a rule. Put otherwise, “act with perfect rationality” is a rule that an ideally rational agent might adopt. The converse is not true: perfectly rational agents, recall, are constrained by their weakness of will. In what follows, an agent who acts with perfect rationality, either by choice or by constraint, will be called an “opportunist.”

Consider, then, a population of ideally rational agents who will be randomly matched to play the ultimatum game. Each will consider the rule “act with perfect rationality” as a possibility. If all adopt this rule, each faces a maximal or minimal payoff with equal probability, depending on whether he is randomly selected as a proposer or responder. Each also considers a rule along the lines of “if I am the responder and am offered  $x\%$  or more, then accept, else reject; if I am the proposer, offer  $y\%$ .” Provided that  $y\% \leq x\%$ , this will yield the same expected value, namely 50, but will in general be less risky. If at least some agents are risk-averse, then  $x\% = 50\%$  will be the riskless behavioral rule. (For risk-neutral agents any  $x\%$  is equally acceptable, but this is a solution only if risk-neutrality is universal.) Thus the prediction for ideal rationality is that all agents will act as nomists and split the payoff equally<sup>7</sup>.

If the agents are in fact risk-neutral, or if the payoffs are von Neumann-Morgenstern utilities,<sup>8</sup> then any proportional division  $x\%$  will be equally acceptable. In such a case, nomist agents face a coordination game. If a particular agent chooses a proportion  $x\%$  or  $y\%$  that differs from the proportions chosen by others, this may result in a disadvantage: for example, if  $x\%$  is greater than the values of  $y\%$  chosen by others, then the agent is quite likely by following his rule to decline an offer and be left with nothing. Similarly, if an agent chooses a value of  $y\%$  that is less than most values of  $x\%$ , he is relatively likely to suffer. Suppose there is a social convention such that, in a situation of this kind, the offer should be  $x\%$ . Then action according to the social convention will be

---

<sup>7</sup> This also agrees with the Shapley Value and nucleolus for this game, as conventionally computed, despite the fact that the ultimatum game is a game of imperfect recall.

<sup>8</sup> Experiments can be so conducted that the payoffs are equivalent to von Neumann-Morgenstern utilities. See Cooper et. al and Roth and Malouf for a protocol for doing so. However, this protocol does not seem to have been followed in ultimatum game experiments, in general.

the unanimous choice of nomist decision-makers. However, all social conventions and all proportions of payment are equally likely.

Thus, we have two cases: risk-aversion and risk-neutrality. In the case of risk-aversion, allowing for errors, offers might be distributed more or less symmetrically around 50-50. But this is not what we observe. In the second case, we would expect to see some social convention emerge around a particular proportion of division, and so far this is in agreement with some evidence; but again we have no reason to expect that offers of less than 50% are more likely than offers above 50%. It seems, therefore, that neither ideal nor perfect rationality (cooperative or noncooperative game theory) is confirmed by the experimental evidence.

Common sense suggests that rationality, strength of will and honesty are distinct traits and that rational individuals may exist in positive numbers whose will is strong and is weak; and that within each category some are honest and some are crooked. Returning to the examples, suppose B believes that a very large proportion of all human beings have weak wills, but has no way to know which type A is. In that case, once again, he would estimate the payoffs of his choices as expected values, using the probabilities based on the frequency of weakness of will in the population and such other evidence as he may have. To fail to do so would be irrational or at best boundedly rational! A rationality that recognizes that there are players of different types in the game and attempts to estimate their proportions has been called “sophisticated rationality.” (Stahl and Wilson 1995). Such sophisticated rationality could, again, be linked either with a strong or a weak will, although it should be distinguished from either perfect or ideal rationality, since perfect rationality is linked to the assumption that all other agents are also perfectly rational, and

ideal rationality would in parallel be linked to the assumption that all other agents are also ideally rational.

We now have three concepts of rationality: perfect, ideal and sophisticated. Ideal rationality is not perfect, perfect rationality is not ideal, and neither is sophisticated (if some agents in the actual world have strong wills and others do not.) What does sophisticated rationality imply for the ultimatum game? Assume a mixed population of opportunists and nomists. Assume risk neutrality, and suppose also that there is a social convention known to all, that  $x\%$  should be offered in an ultimatum game. An opportunist would have to form a conjecture as to the frequency of nomists in the population. (See Appendix 1 for more detail.) Since these will reject offers of  $x\%$  or less, the opportunist may wish to offer at least  $x\%$ , in order to maximize his expectation of payoffs. For example, suppose one-third of the agents will reject an offer under  $x\%$ . If  $x\%$  is more than 35%, then opportunistic proposers will do better to offer a minimal payoff, as in the noncooperative equilibrium. Moreover, agents with strong wills will find that the rule “play with perfect rationality” dominates “offer  $x\%$ ” for the conventional  $x\%$ , so they will choose to be opportunists. But if  $x\%$  is less than 35%, then opportunists would maximize their payoffs by acting according to the rule also. Moreover, they will be able to carry out the rule without relying on strength of will, since following the rule in these circumstances will be the perfectly rational best response for proposers. As this example suggests, the maximum conventional division  $x\%$  that will be adopted by opportunists corresponds roughly to the proportion of the population who can be expected always to play the convention. If the agents are risk-averse rather than risk-neutral, this would favor larger offers (and individual risk-averse agents may offer more

than the conventional amount, especially if there could be some error in determining what the convention is). Nevertheless, there will be an upper limit beyond which the convention will not be followed by anybody. Thus 1) it is at least possible that ideally rational agents, if they are numerous enough, will induce others to act according to the same rule, and 2) this becomes less likely as the conventional offer becomes more generous to the responder.

Suppose, then, that the mixed population has a minority of strong-willed agents. Suppose also that agents of all kinds have less information, in that there is no social convention known to all and different opportunists, having different experiences, may rationally form different estimates as to the proportion of nomists they are likely to encounter. Then we would expect to see a high proportion of accepted offers less than or at most 50%, with 50% chosen by quite risk-averse agents, and some rejections of opportunists and even of some nomists who had guessed wrongly about the commitments of others and gotten unlucky in their matches. And this is what we observe.

The hypotheses that a significant minority of agents have strong wills, and that all are sophisticatedly rational, generate predictions that seem to agree with the evidence on the ultimatum game. In the experimental literature, these results have instead been attributed to reciprocity motives, (Hoffman et. al) an instance of what Sen calls sympathy. To offer an alternative explanation is not to refute the hypothesis of sympathy or reciprocity – both may be true, and may reinforce one another. However, it does indicate that alternate conceptions of rationality deserve equal attention as hypotheses of sympathy in the ongoing research in this field.

The discussion so far assumes absolute rationality in each case, with, at most, some allowance for errors due to imperfect information. However, computation of a perfect, ideal or sophisticated rational solution to a problem may require a great deal of cognitive effort, and cognitive effort is a very scarce resource for real human beings. Acting according to a rule of thumb that may not be “optimal” from any point of view is “boundedly rational.” Can “boundedly” rational decisions be ideal, perfect or sophisticated?

The answer is yes. Suppose that in fact the population comprises individuals both with strong and weak wills, and this is a known fact. Then only sophisticated rationality can be defended as consistent with rational belief. Suppose then that an agent faces a threat, and will attempt to judge the credibility of the threat. Suppose also that the individual believes, on evidence and experience, that most people (though not all) have weak wills. Then computing a perfect equilibrium for the game will be easier than computing a sophisticated solution, since the sophisticated solution requires us to know the perfect solution anyway (and in a particular case, such as Figure 1, the perfect solution may be very easy indeed). Then the rule that “only subgame perfect threats are credible” is boundedly rational. Suppose in addition (as seems very plausible) that most people are better able to exercise a strong will in some circumstances than in others, and that in particular, the agents are situated in a culture that values personal honor highly and regards oath-breaking as dishonorable. In such a society, we suppose, the probability that an oath will be carried out is very high. Then the ideally rational solution may be boundedly rational, in case oaths have been sworn. Finally, consider the game of “running the red light.” A stoplight is a correlated equilibrium solution to an

anticoordination game: its perfectly rational solution in the presence of traffic is to obey the light. However, if an individual is ideally rational, he may commit himself to running the stoplight, and carry the strategy out despite the temptation to stop at the last second. This could maximize his utility if the light has just changed and his intention is made very clear by speeding up. Having observed that about one in three drivers will do this (in Philadelphia) the sophisticated solution of delaying one's start into a green light (without trying to see whether the driver coming the other way has speeded up or not) is sophisticatedly and boundedly rational, although it is neither ideal nor perfect.

In one sense, ideal rationality may be more compatible with the evidence for bounded rationality than perfect rationality or sophisticated opportunism is, since boundedly rational agents are often supposed to act according to (heuristic, imperfect) rules. The agent whose rationality is bounded but ideal will simply choose rules that are cognitively easier to calculate and apply than the ones he might choose if he were perfectly rational. By contrast, the boundedly rational opportunist must calculate an appropriate response to whatever circumstances he may encounter. Rules may conserve cognitive effort and information and so come more easily to an agent with limited cognitive capacity and information than an opportunistic course of action does. This is not a new idea, and in fact recalls the rule-utilitarianism of Mill, though Mill's discourse had mainly to do with ethical rather than prudential decisions.

Suppose that the population includes individuals both with strong and weak wills, and at least some of those with strong wills are honest. How then will coalitions form? First, there will be no mutually beneficial coalitions comprising only the weak-willed. Such coalitions would accomplish nothing that would not be accomplished by a

noncooperative equilibrium. The typical coalition, then, will include at least a subset of strong-willed individuals who adopt threat strategies that encourage the others to keep their agreements and correlate their strategies so as to increase the value of the coalition. These strong-willed individuals may be known to the others as leaders, but more probably as officious busybodies, nosy parkers and snitches. It may be that the officious busybodies, nosy parkers and snitches will form a grand coalition and formalize some of their threat strategies as institutions such as property rights and enforcement of contracts. If, as seems likely, people are better able to act with strength of will in some circumstances than in others, we are likely to see cooperative arrangements more often in some social circumstances than others, for example among people of a common religious faith, and to see a good deal of noncooperative interaction among the coalitions that do occur. If a part of the population are both strong-willed and dishonest, they may be able to form some coalitions for their dishonest purposes, by means of committed threat strategies, even though dishonesty breeds distrust and distrust is an obstacle to cooperation, as in the game in Figure 1. For coalitions of this kind, the “irrationality” of gang vendettas would be seen as an expression of ideal (though not perfect) rationality. As this example and the stoplight example suggest, one should not suppose that ideal rationality is preferable to perfect or sophisticated rationality on any general normative grounds.

#### iv. Conclusion

This essay has argued that strength or weakness of will is often linked to concepts of rationality, leading to at least three distinct concepts of rationality, each of which may or may not be bounded. “Ideal rationality” links rationality to strength of will. The essay

has argued that “ideal rationality” is characteristic of cooperative game theory and is the substantive difference that distinguishes cooperative game theory from noncooperative game theory. “Perfect rationality” links rationality to weakness of will. This essay has argued that “perfect rationality” characterizes neoclassical economics and noncooperative game theory. The third concept of rationality is “sophisticated rationality,” which is consistent with the belief that the population includes both types with strong and with weak will. This belief leads toward a world very much like the world we seem to live in, a world not susceptible to analysis in terms either of perfect or of ideal rationality.

## Appendix 1. Optimal Offers in a Mixed Population

This appendix illustrates the reasoning behind the proposition, in the text, that the conventional offer must be less than a critical value determined by the proportion of nomists in the population. A conventional offer will always be accepted, since opportunists will accept any offer and nomists will accept a conventional offer. Thus the expected value is simply what is left to the proposer, i.e.  $100 - Y$  where  $Y$  is the offer. A minimal offer will be rejected by nomists, so that its expected value is  $(1 - X)(100 - Y)$ , where  $X$  is the proportion of nomists in the population. Figure A1 shows the expected values as they vary with the conventional offer, assuming that  $1/3$  of the population are nomists.

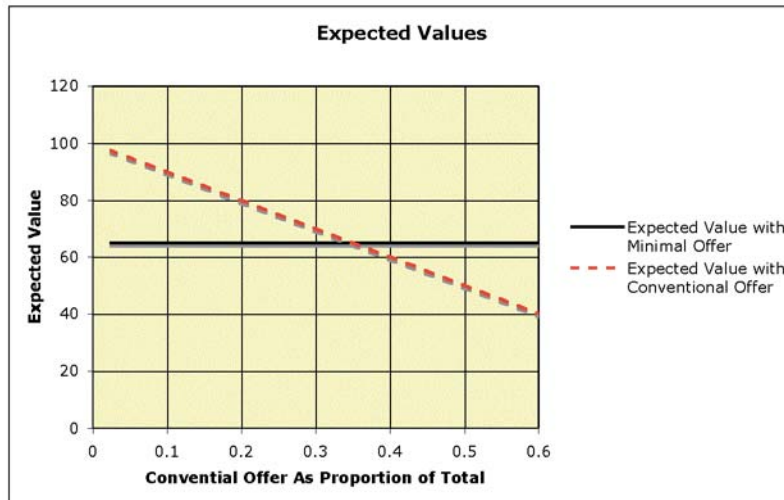


Figure A1. Conventional And Minimal Offer with  $1/3$  Nomists

## Appendix 2. Supergame Analyses of Three Examples

For a given game, Aumann (1959) defines the supergame as follows. Suppose the game is played for an infinite series of repetitions and the players choose their strategies for each individual play according to some rule. Then the supergame is the noncooperative game of choosing the rules by which each repetition will be played. In his Nobel lecture (2005) Aumann returns to the supergame concept, drawing on the extensive literature since 1970 on repeated play. Aumann identifies the cooperative solution of the original game with a *strong* equilibrium of the supergame. A strong equilibrium is a Nash equilibrium with the further property that no group of players (coalition) can improve its payoff by a coordinated shift to another Nash equilibrium. Thus for Aumann's "acceptable point" solution, the cooperative solution is a set of rules that define a Nash equilibrium for the supergame, such that no group can mutually benefit by shifting to another set of rules. This is one of very few solution concepts for cooperative games that specifies the solutions of the underlying game and that does not require an explicit agreement. In the 1959 paper, the infinite repetition is not supposed actually to occur, but is a conceptual device to identify cooperative solutions, and this is consistent with what I have called ideal, but not perfect, rationality.

As an example, consider the social dilemma in Table A1. Among the rules that the two players might adopt are 1) always play C, 2) always play D, or 3) Tit-for-tat. The well-known Tit-for-tat rule says "First play C, and always play C unless the other player has played D on the previous round. There are, of course, many other rules the agents might choose, but these three will suffice. In the literature of repeated play, the rules that will be chosen depend on the discount factor,  $\delta$ . This discount factor allows for the

probability that the play will be discontinued on a particular round (since nothing really lasts forever) as well as for time preference. For this game, if  $\delta > \frac{1}{2}$ , then the Tit-for-tat rule will deter the other player from playing noncooperatively.

Table A2 about here

Table A2 gives a supergame for the social dilemma in Table A1, considering only these three rules. With  $\delta > \frac{1}{2}$ , we have  $\frac{1}{1-\delta} 2 < Y < \frac{1}{1-\delta} 3$  and  $X < \frac{1}{1-\delta} 2$ . Thus, the supergame has two equilibria, one in which both agents always choose D and one in which both choose Tit-for-tat. Since the latter is Pareto-preferable to the former, only the Tit-for-tat equilibrium is strong, and thus it defines the cooperative solution in Aumann's sense.

Table A1. A Social Dilemma

Payoff order: A, B	Player B		
	C	D	
Player A	C	3,3	1,4
	D	4,1	2,2

A difficulty with this discussion is that there are many equilibria of a repeated game such as this, with discounted total payoffs at every level between the noncooperative and efficient outcomes. Aumann's focus on strong equilibria narrows the field greatly, however, since any equilibrium that is (like always D, always D) Pareto-dominated will not be strong, only efficient outcomes will be among the strong equilibria. Conversely, while there might be other (perhaps very complex) rules that would dominate Tit-for-tat if they were taken into account, the Nash equilibrium of the supergame,

whatever it may be, can do no worse than cooperative play. Finally, recall that the condition  $\delta > \frac{1}{2}$ , with *actual* repeated play of the game, are conditions necessary for *perfectly* rational agents to realize cooperative play. For *ideally* rational agents these conditions will not be necessary.

Table A2. Partial Supergame for A Social Dilemma

Payoff order: A, B		Player B		
		Always C	Always D	Tit-for-tat
Player A	Always C	$\frac{1}{1-\delta}3, \frac{1}{1-\delta}3$	$\frac{1}{1-\delta}, \frac{1}{1-\delta}4$	$\frac{1}{1-\delta}3, \frac{1}{1-\delta}3$
	Always D	$\frac{1}{1-\delta}4, \frac{1}{1-\delta}$	$\frac{1}{1-\delta}2, \frac{1}{1-\delta}2$	Y,X
	Tit-for-tat	$\frac{1}{1-\delta}3, \frac{1}{1-\delta}3$	X,Y	$\frac{1}{1-\delta}3, \frac{1}{1-\delta}3$

Now consider the two-person game of Figure 1 in the text. Suppose that the responder considers the grim trigger rule, “If A chooses strategy 3’ then choose 5’ on every subsequent play; otherwise play 4’”. Other rules we will consider are to always choose one of the strategies 1’, ..., 5’. If  $\delta$  is greater than 0.8, this trigger strategy will deter opportunism by player A. Table A3 shows a supergame in normal form for just these rules, with  $X < 4371$ ,  $Y < 2500$ . Equilibria of the supergame are 1’, 5’ as in the text and 2’, Grim. Of the two, only 2’, Grim is strong. Thus the supergame analysis again identifies what intuition picks out as the cooperative equilibrium, with the same qualifications as before.

Table A3. A Partial Supergame for the Game in Figure 1

Payoff order: A, B		B		
		4'	5'	Grim
A	1'	$\frac{1}{1-\delta}4,237, \frac{1}{1-\delta}2000$	$\frac{1}{1-\delta}4,237, \frac{1}{1-\delta}2000$	$\frac{1}{1-\delta}4,237, \frac{1}{1-\delta}2000$
	2'	$\frac{1}{1-\delta}4371, \frac{1}{1-\delta}3000$	$\frac{1}{1-\delta}4137, \frac{1}{1-\delta}2500$	$\frac{1}{1-\delta}4371, \frac{1}{1-\delta}3000$
	3'	$\frac{1}{1-\delta}4900, \frac{1}{1-\delta}1000$	$\frac{1}{1-\delta}4137, \frac{1}{1-\delta}2500$	X,Y

For the ultimatum game, trigger strategies do not seem helpful, since rejection of an offer is never a noncooperative equilibrium, unlike “D, D” in the social dilemma and “1’, 5’” in the game in Figure 1. Nevertheless, ideally rational responders might play in the supergame according to a rule “always reject offers below x%.” In that case, the proposer’s best response is “always offer x%.” Conversely, if proposers “always offer x%” then the responders have nothing to lose by playing “always reject offers below x%.” Thus a supergame equilibrium can always exist for any x%. Moreover, any shift from one to another such equilibrium will make one of the players worse off, so all are strong equilibria. Further, this does not depend on risk neutrality. Thus, as in the text, there is a coordination problem in the supergame, which could be solved by a social condition.

Suppose, however, that we move slightly away from absolute rationality by introducing “trembling hand” errors. (Selten 1975) A “trembling hand’ error is a small probability that any non-equilibrium strategy will be chosen despite the agent’s intention to play the equilibrium strategy. In the supergame, this means a small probability that the

other player will play a non-equilibrium rule. In the social dilemma supergame, the equilibrium at “Tit-for-tat, Tit-for-tat” is unaffected. Thus, the equilibrium is both strong and perfect. The same is true of the strong equilibrium in the supergame based on Figure 1. However, in the ultimatum game, it is no longer the case that “always reject offers below  $x\%$ ” is a best response to “always offer  $x\%$ ,” since this will result in some unprofitable rejections. Indeed, “never reject” dominates “always reject offers below  $x\%$ ” in the presence of trembling hand errors, or, indeed, any other deviations from equilibrium play with any positive probability, however small. In other words, only the Nash equilibrium of the original game is a perfect strong equilibrium of the supergame. This occurs because the information structure of the ultimatum game does not permit the proposer to condition his strategy on that chosen by the responder. That is, when played by a coalition, the game is one of imperfect recall. (Kuhn 1954)

The significance of perfect equilibrium *in the supergame* is that, to interpret experimental results, we must admit the possibility of some errors. Perfect equilibria are robust against at least one category of small errors; imperfect equilibria are not. Therefore, cooperative game theory in this sense cannot improve on noncooperative game theory so far as correspondence to the evidence is concerned.

The discussion in the text differs in that, by focusing on the equilibrium *in the supergame*, Aumann tacitly excludes utilitarian generalization. The utilitarian generalization is not an individual best response, but rather a best response constrained by the condition that all responders choose the same rule. As such, it allows the responder to anticipate, and avoid, the shift to a noncooperative equilibrium that occurs when other responders choose their dominant “never reject” rule.

## References

- Akerlof, G.A. and Yellen, J.L (1986), *Efficiency Wage Models of Labor Market* (Cambridge University Press, Cambridge).
- Akerlof, George A. (2001), "Behavioral Macroeconomics and Macroeconomic Behavior," (Nobel Foundation) available at: [http://nobelprize.org/nobel\\_prizes/economics/laureates/2001/akerlof-lecture.html](http://nobelprize.org/nobel_prizes/economics/laureates/2001/akerlof-lecture.html), as of August 9, 2008.
- Andreoni, James and Emily Blanchard (2006), "Testing Subgame Perfection Apart from Fairness in Ultimatum Games ," *Experimental Economics* v. 9, no. 4 pp. 307-321.
- Aumann, R. J. (1959), "Acceptable Points in General Cooperative n-Person Games," *Contributions to the Theory of Games, Volume IV (Annals of Mathematics Studies, Number 40)* v. 4, pp. 287-324.
- Aumann, R. J. (2005), "War and Peace (Nobel Prize Lecture)," (The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2005) available at: [http://nobelprize.org/nobel\\_prizes/economics/laureates/2005/aumann-lecture.html](http://nobelprize.org/nobel_prizes/economics/laureates/2005/aumann-lecture.html), as of June 9, 2007.
- Camerer, Colin (2003), *Behavioral Game Theory: Experiments in Strategic Interaction* (Princeton:Princeton University Press).
- Cooper, Russell W. and Douglas V. DeJong, Robert Forsythe, and Thomas W. Ross (1990), "Selection Criteria in Coordination Games: Some Experimental Results," *American Economic Review* v. 80, no. 1 (Mar) pp. 218-233.
- Elster, Jon (1977), "Ulysses and the Sirens: A Theory of Imperfect Rationality," *Social Science Information* v. 16, (Oct) pp. 469-526.
- Guth, W and R. Schmittberger and B. Schwartz (1982), "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization* v. 3, pp. 376-388.
- Henrich, Joseph and Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, Richard McElreath, Michael Alvard, Abigail Barr, Jean Ensminger, Natalie Smith Henrich, Kim Hill, Francisco Gil-White, Michael Gurven, Frank W. Marlowe, John Q. Patton and David Tracer (2005), "'Economic man' in cross-cultural perspective: Behavioral experiments in 15 small-scale societies," *Behavioral and Brain Sciences* v. 28, no. 6 (Dec) pp. 795-815.
- Hoffman, Elizabeth and Kevin McCabe and Vernon L. Smith (1998), "Behavioral Foundations of Reciprocity: Experimental Economics and Evolutionary Psychology," *Economic Inquiry* v. 36, no. 3 (July) pp. 335-352.

- Kuhn, H. W. (1953), "Extensive Games and the Problem of Information," *Contributions to the Theory of Games, Volume II (Annals of Mathematics Studies, Number 28)* edited by H.W. Kuhn and A. W. Tucker (Princeton: Princeton University Press) pp. 193-216.
- McCain, Roger A. (2009), *Game Theory and Public Policy* (Forthcoming, Elgar).
- Mill, John Stuart (1863), *Utilitarianism* (Kitchener, Ontario, Canada: Batoche Books).
- Nash, John (1953), "Two-Person Cooperative Games," *Econometrica* v. 21, (Jan) pp. 128-140.
- Oosterbeek, Hessel and Randolph Sloof and Gijs van de Kuilen (2004), "Cultural Differences in Ultimatum Game Experiments: Evidence from a Meta-analysis," *Experimental Economics* v. 7, no. 2 (June) pp. 171-188.
- Roth, Alvin and Malouf, Michael W. (1979), "Game-theoretic models and the role of information in bargaining," *Psychological Review* v. 86, no. 6 (Nov) pp. 574-594.
- Roth, Alvin E. and Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir (1991), "Bargaining and Market Behavior in Jerusalem, Ljubliana, Pittsburgh, and Tokyo: An Experimental Study," *American Economic Review* v. 91, no. 5 (Dec.) pp. 1068-1095.
- Roth, A. (1995), "Bargaining Experiments," *Handbook of Experimental Economics*, edited by Kagel, J., Roth, A. (Princeton: Princeton Univ. Press) pp. 253-348.
- Schelling, T.C. (1980), "The Intimate Contest for Self-Command," *The Public Interest*, no. 60 (Summer) pp. 94-118.
- Selten, Reinhard (1975), "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory* v. 4, pp. 25-55.
- Sen, Amartya K. (1979), "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory," *Philosophy and Economic Theory* (London: Oxford University Press).
- Sen, Amartya and Bernard Williams (1982), "Introduction," *Utilitarianism and Beyond* edited by Amartya Sen and Bernard Williams (New York: Cambridge University Press) pp. 1-22.
- Simon, H. A. (1955), "A Behavioral Model of Rational Choice," *Quarterly Journal of Economics* v. 69, no. 1 (Feb) pp. 99-118.
- Simon, H. A. (1995), "Artificial Intelligence: An Empirical Science," *Artificial Intelligence* v. 77, pp. 95-127.
- Stahl, Dale O. and Paul W. Wilson (1995), "On Players' Models of Other Players: Theory and Experimental Evidence," *Games and Economic Behavior* v. 10, pp. 218-254.

Stanley, T. D and Ume Tran (1998), "Economics Students Need Not Be Greedy: Fairness and the Ultimatum Game," *Journal of Socio-Economics* v. 27, no. 6 pp. 657-663.

von Neumann, John and Oskar Morgenstern (2004), *Theory of Games and Economic Behavior* (60th anniversary edition. Princeton: Princeton University Press).

Henry R. West (2006), *The Blackwell Guide to Mill's Utilitarianism* (Oxford: Blackwell).