Ratio (new series) XX 1 March 2007 0034–0006

PRISONER'S MISTRUST

Erin I. Kelly and Lionel K. McPherson

Abstract

The standard, non-repeated prisoner's dilemma poses no true dilemma about rationality, we argue. What the prisoners ought rationally to do, unless they are selfless, depends on the relationship of trust that they have or lack with one another. This helps to diffuse the apparent conflict between individual and collective rationality. If the prisoners have reason to trust one another, pursuing a joint strategy would be rational for them. In the absence of trust, pursuing an individual strategy would be rational. The solution that is supposed to be puzzling – because each prisoner confessing is worse for both than the alternative on which both remain silent – is simply the rational solution for persons who have no reason to trust one another. Philosophers have been misled by the apparent availability of the better alternative. Collective rationality is not relevant for persons who do not stand in a trusting relationship. When they do, there is no unresolved conflict between their individual and collective points of view.[1]

I.

We challenge the notion that the prisoner's dilemma, despite the attention it has received from philosophers, involves a deep puzzle about rationality. Our aim is to unpack some of the issues that generate the impression there is a puzzle. The question about what the prisoners ought rationally to do, we will argue, cannot be answered without examining what they value. More specifically, the rational course of action for the prisoners – if neither would accept fully sacrificing his own interests – depends on the relationship of trust that they have or lack with one another.

The relevance of this relationship helps to diffuse the apparent conflict between individual and collective rationality. We will focus on whether the prisoners have reason to trust one another.

[1] We thank Virginia Held and Michael W. Klein for helpful discussion and the Edmond J. Safra Foundation Center for Ethics, Harvard University for generous fellowship support.

Trust, as we understand it, exists when persons have strong mutual concern and believe that none would advance his own interests or goals at the other's expense. If the prisoners have reason to trust one another, it would be rational for them to pursue a joint strategy. If they do not have reason to trust one another, each ought rationally to pursue an individual strategy. Hence no special 'solution' to the prisoner's dilemma is needed because there is no real dilemma, a conclusion that runs contrary to Elizabeth Anderson's recent proposal.[2]

Acting on trust does involve risk – there is no guarantee that the other person will cooperate. Yet trust is not unconditional. It is built on mutual concern. A relationship in which persons have not yet made themselves vulnerable to one another can provide a rational basis for venturing trust under conditions in which vulnerability arises: strong mutual concern supports the confidence that trust requires.

Here is a characterization of the standard dilemma. Slim and Shady have been arrested for a crime. They are held in separate jail cells and each is offered the following deal by the prosecutor. If you confess and your partner remains silent, I will use your confession against him, he'll get ten years, and I'll set you free. Likewise, if your partner confesses and you remain silent, I will use his confession against you, you'll get ten years, and I'll set him free. If you both confess, I will convict you both but, in exchange for your confessions, each of you will get a reduced sentence of five years. If you both remain silent, I can convict you both only on lesser charges, which means you'll each get two years.[3]

Slim reasons in this way. If Shady confesses and I don't, I'll get ten years. But if I confess too, I'll only get five years. So, if he confesses, I should confess. If he does not confess and I do, I go free. Therefore, no matter what he does, I should confess. Shady reasons in the same way. So both prisoners confess, and each receives harsher sentences (five years) than if they both had remained silent (two years).

[2] See Elizabeth Anderson, 'Unstrapping the Straitjacket of "Preference": A Comment on Amartya Sen's Contributions to Philosophy and Economics', *Economics and Philosophy*, 17 (2001), pp. 21–38, and 'The Prisoner's Dilemma: Solved', Harvard University, Cambridge, MA, 24 October 2002.

[3] It is important to specify and not merely rank the payoffs: if the worst outcome is to be avoided at all costs (e.g., death) or if the differences between the options are not significant enough, this could factor into judgments about the reasonableness of individual risk taking.

As many philosophers have interpreted the prisoner's dilemma, it illustrates a conflict between individual and collective rationality. Individuals, each rationally pursuing his self-interest, may all end up worse off than if they had identified as a group and acted contrary to individual self-interest. Derek Parfit puts the point like this: the individually rational course of action, the aim of which is to make the outcome better for oneself, is collectively self-defeating. If all persons follow this course of action, the outcome will be worse for everyone. The aims of each person will be worse achieved. This may seem to imply that 'if we were choosing a collective code, something we will all follow, then a theory that advises us each to do what would be better for ourselves would tell us to reject itself.'[4] However, Parfit claims, a defender of the individually self-interested point of view can argue that the group perspective is irrelevant to the prisoners. Since they do not operate with a collective code, 'to be collectively self-defeating is, in the case of prudence, *not* to be self-defeating.'[5] It is not irrational for the prudent prisoner to seek the best outcome for himself.

Game theorists agree that in the standard, non-repeated prisoner's dilemma, collective rationality does not make sense. The concept of Nash equilibrium is helpful here. A Nash equilibrium is a profile of strategies according to which each player's strategy is an individually maximizing response to the other players' strategies.[6] The unique Nash equilibrium in the prisoner's dilemma is the outcome where each prisoner confesses – that is, the second worst of the four possible outcomes for each of them, where the second best outcome for each would be better for both. This result is not understood as a failure of the Nash model for consistently predicting rational behavior. Rather, the features of the non-repeated prisoner's dilemma are such that rational individuals can expect to do relatively badly in the game.[7]

This unhappy outcome has led game theorists to alter the game. The purpose is to introduce conditions whereby it would be rational for each of the players to make choices that would be better for them both, rendering the outcome Pareto efficient. A

[4] Derek Parfit, 'Is Common Sense Morality Self-Defeating?'. *Consequentialism and its Critics*, ed. Samuel Scheffler (Oxford: Oxford University Press, 1988), p. 185.

[5] Ibid.

[6] Drew Fudenberg and Jean Tirole, *Game Theory* (Cambridge, MA: MIT Press, 1991), p. 11.

[7] Roger B. Myerson, *Game Theory: Analysis of Conflict* (Cambridge, MA: Harvard University Press, 1991), pp. 97–8.

brute modification allows the players to make binding agreements.[8] If this move is allowed, it is obvious that the players will cooperate (assuming they believe the agreement really is binding). A more subtle modification transforms the game into a repeated prisoner's dilemma: the players believe there is a high probability that they will play again, and they do not know in advance when the games will come to an end.[9] The players therefore would have reason to adopt a long-run strategy that would reward cooperation and penalize defection from what in effect becomes a joint strategy. Drew Fudenberg and Jean Tirole observe that 'repeated games may be a good approximation of some long-term relationships in economics and political science – particularly those where "trust" and "social pressure" are important. . . .'[10] Of course, cooperation in repeated games is generated by individual rationality, with no influence from moral or social values.

The real question for the prisoners, we will argue, concerns whether or not they can trust one another. We have in mind genuine trust, as compared to 'trust' that would have to be compelled by a guarantee of enforcement of binding agreements or by a demonstration of rational foresight across an indefinite horizon of repeated games. That is, we are concerned about the ramifications of trust for a non-repeated prisoner's dilemma.

II.

Anderson rejects the type of analysis Parfit gives and proposes a solution to the prisoner's dilemma. She argues that if the prisoners could shift to a 'we' perspective from which each would abide by a collective code, for each to remain silent would clearly be rational. The 'we' perspective involves a joint strategy. In particular, each person must be guided by concern for everyone: 'the parties are committed to acting only on reasons that are univer-

[8] See, e.g., Fudenberg and Tirole, *Game Theory*, pp. 74–5; and Myerson, *Game Theory*, pp. 244–45.

[9] Myerson, *Game Theory*, pp. 398–99.

[10] Fudenberg and Tirole, *Game Theory*, p. 145. For an influential discussion of the evolution of cooperation in iterated prisoner's dilemmas, see Robert Axelrod, *The Evolution of Cooperation* (New York: Basic Books, 1984).

salizable to their membership.'[11] The question is whether adopting a joint strategy – in which the parties act only on the basis of shareable reasons – could be rational from each prisoner's perspective, apart from the likelihood that this strategy would advance his own interests. Anderson believes she can show that it could be and that this dissolves the dilemma.

Anderson's view is not straightforwardly about the requirements of morality versus the rational pursuit of individual self-interest. She does not argue that morality could require the prisoners to act collectively, on the basis of loyalty and mutual concern, even though a joint strategy would conflict with individual rationality. In fact, she believes that morality might require the prisoners to confess, assuming they are guilty, in the interests of society. So her proposed solution to the prisoner's dilemma does not rest on the claim that morality trumps individual rationality.[12]

Nor does Anderson hold that a moral course of action is rationally required. While she is drawn to a Kantian analysis of the content of moral claims, she does not endorse the Kantian position that rationality demands the moral point of view, assuming for the moment that this would require the prisoners to adopt a joint strategy. She writes, 'I have no argument that would show that identification as a member of a universal community of humanity – a kingdom of ends – is rationally required.'[13] Her focus is on how it could be rational for each of the prisoners to make an effort to cooperate with his partner, not why it must be.

According to Anderson, an account of rational action must recognize a distinction between the rational agent who does what would be best for himself and the rational agent who acts on a social norm of responsibility.[14] She claims that each of these orientations could be rational under appropriate circumstances. Her discussion of the latter orientation elaborates what Amartya Sen refers to as 'commitments.' Commitment is not to be confused with sympathy. The case of sympathy, Sen writes, is one 'in which the concern for others directly affects one's own welfare. If the

[11] Anderson, 'Unstrapping the Straitjacket of "Preference" ', p. 29.

[12] Moral considerations may be relevant, however. We suppose that the prisoners need to be guilty, since there are strong moral reasons not to accuse someone falsely, whether or not doing so would better one's own situation.

[13] Anderson, 'Unstrapping the Straitjacket of "Preference" ', p. 37.

[14] Ibid., p. 22.

knowledge of torture of others makes you sick, it is a case of sympathy; if it does not make you feel personally worse off, but you think it is wrong and you are ready to do something to stop it, it is a case of commitment.'[15] Sen's point is that sympathy with others may change the conditions that would advance an agent's welfare. Agents who sympathize with one another may in effect represent their own interests by refusing to improve their situation at any of the others' expense. Commitment, by contrast, opens up a rational course of action that may well conflict with self-interest; an agent's commitments express her values and may be rational even when they do not serve to make her better off.

Sen thus challenges a model of rationality whereby a chosen alternative must be better than (or at least as good as) the others for the person choosing it. He suggests that persons guided by that model may turn out to be 'rational fools' who fail to appreciate possibilities that could be brought about by acting on commitment.[16] Nevertheless, Sen is careful to claim that 'in a situation where the outcome depends on other people's actions in addition to one's own . . . the choice of rational action depends then on the actions of others, and ultimately on the preferences of others.'[17] This point applies to a wide range of personal commitments as well as to self-interested concerns. Sen is not claiming that persons are fools if they qualify their commitments by thinking strategically about whether they can count on the cooperation of others. Instead, he urges us to think about how social pressures toward other-regarding behavior might evolve and serve to make morality and rational behavior more consistent.[18]

Anderson wants to apply the insight of Sen's rational fools challenge directly to the problem of choice framed by the standard prisoner's dilemma. Her question is how it could be rational for an agent to choose a strategy that would serve collective interests, were others likewise to follow the strategy, when doing so conflicts with a maximizing strategy focused on advancing his own interests. Anderson's answer is that a person can act rationally on

[15] Amartya Sen, 'Rational Fools: A Critique of the Behavioral Foundations of Economic Theory', *Philosophy and Public Affairs*, 6 (1977), p. 326.

[16] Ibid., p. 341.

[17] Sen, 'Choice, Orderings and Morality', *Choice, Welfare and Measurement* (Cambridge, MA: MIT Press, 1982), p. 75.

[18] See also Byeong-Uk Yi's discussion of the rationality of forming dispositions versus the rationality of acting on them; 'Rationality and the Prisoner's Dilemma in David Gauthier's *Morals by Agreement*', *The Journal of Philosophy*, 89 (1992), p. 486.

a commitment to collective interests when he identifies as a member of a social group. The opportunities for identification are various: persons can identify with friends, family, workplace, community, nation, etc. This would suggest that identification is a matter of psychological affinity, perhaps based on perceived commonalities or special concern. But Anderson also claims that one may opt to identify with others when doing so would solve a deliberative problem, even when one has no prior or independent reason to identify.[19]

The prisoner's dilemma, Anderson believes, presents such a problem. She does not argue that rationality requires persons to identify with one another. Yet when they do, they thereby constitute a collective agent who will properly regard the object of choice to be a single, joint strategy. This means that persons who identify with one another will universalize their reasons across membership in the group: 'only the reasons that *we* can share are reasons on which people who identify as "we" can accept as a ground for their action.'[20] More strongly, Anderson argues that collective identification is fundamentally at odds with the pursuit of individual utility. She writes, 'The universalization principle *rules out* the principle of maximizing expected utility (individual preference satisfaction) as an acceptable principle of rational choice for members of a collective agency who constitute the parties to a prisoner's dilemma.'[21] Hence her solution to the prisoner's dilemma.

But the solution Anderson proposes faces a serious problem. A person cannot rationally expect the cooperation of others simply given the fact or possibility of identifying with a social group. Without this expectation, it is hard to see how rationality could require a member to act on reasons that universalize over membership in the group. To repeat, while Anderson does not claim that rationality requires collective identification, she appears to argue that once a person does identify with a larger group, perhaps for moral or psychological reasons or to avoid an apparent paradox of rationality, that person is then bound by the universalization principle. This is unconvincing. If you identify with your fellow prisoner on grounds of your shared ethnicity, for example, this seems insufficient rationally to ground a cooperative

[19] Anderson, 'Unstrapping the Straitjacket of "Preference" ', p. 31.
[20] Ibid., p. 29.
[21] Ibid.

strategy. Your reason for identification must be good enough to justify the collective strategy. Specifically, your reason must be good enough to support the broader goal – promoting the interests of the group or, at least, of your fellow prisoner – apart from any assessment of the risks this may pose to your own interests.[22] Mere ethnic identification does not seem rationally or morally weighty enough to accomplish this.

Anderson's alternative suggestion is that perhaps the relevant sense of group identification is provided simply by the recognition that if you both were to act on a joint strategy, you both would do better than if you each were to attempt to maximize your individual utilities. But this suggestion is not helpful. Identification then would not be the basis for solving a deliberative problem. It would simply be the solution, with no apparent basis. The emptiness of this solution as an answer to the question of what could ground the claims of collective rationality suggests that collective rationality has preconditions and that the relevance of a 'we' perspective for the prisoners depends upon whether the preconditions have been met.

In fact, Anderson admits that further conditions may need to be satisfied in order rationally to defend the commitment involved in identification with a particular group. She claims that 'particular collective agencies may fail to survive rational scrutiny, in which case uncritical identification with them would be irrational.'[23] Now this suggests a standard for collective rationality that stands apart from the matter of whether an individual identifies with a group. She does not pursue this thought.

As Anderson understands the prisoner's dilemma, a choice between the perspectives of individual and collective rationality can be decided by a decision to identify with the group. But her characterization of the presuppositions of collective rationality as a matter of whether individuals identify with a group is inadequate. She fails to analyze the conditions under which group membership rationally supports the adoption by members of a joint strategy for choice. Some account is needed of what the basis of collective identification could be such that it could support a joint strategy. Moreover, we want to know whether the rationality

[22] On the normative bases of African American solidarity, for example, see Lionel K. McPherson and Tommie Shelby, 'Blackness and Blood: Interpreting African American Identity', *Philosophy and Public Affairs* 32 (2004), pp. 171–92.

[23] Ibid., p. 32.

of a joint strategy is contingent on its prospects for success. Advancing the interests of your fellow prisoner by leaving your cooperative strategy open to exploitation by him does not advance the interests of the group.

III.

Some commentators have thought that the problem for the prisoners is one of communication.[24] If only they could talk, they could avoid an outcome that would be worse for them both. This would be to view the prisoner's dilemma as a coordination problem. Coordination problems can arise even when persons are not self-interested.

Suppose you want to do what would serve your group's anti-war cause. The best outcome would have most members demonstrate against the war, while a few remain in the office to send email alerts of the group's next action. The problem is you are unsure of what the other members will do. If most go to the demonstration, for you to go to the office would be better. Most, though, might go to the office, each worrying that the general preference to demonstrate will result in the office being unattended. For you to go to the office as well would then be better; your group would appear silly with only a few members at the demonstration. So, either way, you should go to the office. But if all members reason in this way, you all end up in the office, which would be a worse outcome for your cause than if you were all to end up at the demonstration.[25] A dilemma such as this is easily solved through communication: some in the group could plan to go to the demonstration and others to the office.

The prisoner's dilemma is not a predicament of this sort. The main issue for the prisoners is not about coordinating their actions, if only they could, by agreeing to remain silent. Unless they trust each other to remain bound by a joint strategy, an opportunity to make one is irrelevant. The temptation to defect remains – even after discussion produces 'the agreement' – and perhaps even becomes stronger. The prisoners may well be

[24] See, e.g., Myerson, *Game Theory*, p. 244.
[25] We thank T. M. Scanlon for this example.

inclined to exploit what they take to be the other's good word.[26] This is because they ultimately endorse competing goals: viz., each is trying to do what would be best for himself.

With trust, however, there is no dilemma. When trust is in place, no communication or agreement is needed, and there is no problem of enforcement. As we have described it, trust characterizes a relationship in which persons come to identify the other's interests closely with their own and are aware that this concern is mutual. We are supposing that concern for the other does not directly affect each person's own welfare via sympathy. Each person's concern for the other depends on there being a reciprocal commitment to maintaining this relationship – a commitment that is conditional on the other party's sharing it.

Imagine that the prisoners' decision-making is guided by a principle expressing their mutual commitment: Do not inform on your partner, since he will not inform on you. Each accepts the principle conditionally, on the trust that his partner also accepts it. The principle is a reasonable one. It expresses loyalty and reciprocity, and it supports an outcome that would be better for both prisoners than the alternatives that are supported by reasoning from self-interest. Commitment to a principle like this would not be unreasonable, even if it means receiving a punishment that one could otherwise avoid by defecting. Only a very strong self-interest theory would require, as a matter of rationality, that a person abandon his principles whenever this would most benefit him.[27] For persons who have mutual concern and trust, the temptation to defect is gone. Each constrains his self-regarding behavior by their relationship of trust – a relationship that both find worth maintaining, given their reciprocal investment – despite the vulnerability that each incurs to having this trust exploited by the other. There is no troubling conflict between the interests of each and the collectively best outcome.

This shows that Anderson is right to some extent. The prisoner's dilemma might be dissolved, yet not for the reasons she offers. Prisoners who trust each other face no dilemma. A joint strategy is rational for them. Within a relationship of trust, the question

[26] This suggests, contra Anderson, that voting dilemmas are not prisoner's dilemmas; see 'Unstrapping the Straitjacket of "Preference" ', p. 26. What characterizes the prisoner's dilemma is a temptation to defect, even when a collective solution could be worked out.

[27] David Lewis seems to assume a strong self-interest theory of rationality in 'Prisoner's Dilemma as a Newcomb Problem', *Philosophy and Public Affairs*, 8 (1979), pp. 235–40.

'What should we do?' supersedes the question 'What should I do?' But we cannot make sense of the idea that it could be rational – in the absence of mutual concern and trust – for a person to play his part in what would be a collectively rational scheme. It is not irrational for each prisoner to avoid the worst outcome for himself while at the same time seeking the best outcome for himself. Contrary to what Anderson argues, rationality here actually seems to require the individual perspective. Without trust, there is no dilemma.

We have been supposing – perhaps in light of the emphasis on free markets and individual rationality prevalent in the United States – that background social conditions are such that persons generally have reason to mistrust each other when the stakes for them as individuals are high. Such a context seems to support adopting what Virginia Held refers to as 'the competitive policy'; we would agree with her that 'if one party has solid grounds for predicting that the other will adopt the competitive policy, then not to do so as well amounts to altruistic self-sacrifice. . . .'[28] The competitive policy is at odds with building trust.

Still, we are not claiming that the prisoners must be self-interested.[29] Prisoners with strong other-regarding commitments and no expectation of reciprocity could act rationally on those commitments. But this would not amount to a strategy of collective 'as if' reasoning. If others do not cooperate, collective goals are not advanced. No joint strategy is needed for an individual to accept fully subordinating his interests. Anderson argues that a person may rationally identify with a group and attempt to promote collective aims without the assurance that others will cooperate. Where common knowledge of everyone's (rational) conditional willingness to join together is absent, she claims, one may still regard oneself as part of an imagined common agency 'in the hope that others will join and make it real by cooperating.'[30] Our view is that this would represent a kind of wishful thinking, which lacks the support of reason.

[28] See Virginia Held, 'Rationality and Reasonable Cooperation', *Social Research*, 44 (1977), p. 728.

[29] It has been acknowledged that a prisoner's dilemma-type situation could arise between altruists as well as egoists. On prisoner's dilemmas between altruists see, e.g., John J. Tilley, 'Altruism and the Prisoner's Dilemma', *The Australasian Journal of Philosophy*, 69 (1991), pp. 264–87, and 'Prisoner's Dilemma from a Moral Point of View', *Theory-and-Decision*, 41 (1996), pp. 187–93. See also Sen, 'Choice, Orderings and Morality', pp. 81–2.

[30] Anderson, 'Unstrapping the Straitjacket of "Preference" ', p. 32.

Consulting a 'we' perspective will not solve the prisoner's dilemma without a mutual commitment to the pursuit of collective goals. Trust cannot be presupposed in the name of collective rationality: trust must be in place if the directives of collective rationality are to be relevant. But then a collective strategy is not the solution to a dilemma. When trust is in place, the collective strategy is clearly the rational response to the prisoners' situation.

IV.

Philosophers other than Anderson have been drawn to the hope of a collective solution. Held contemplates the possibility that the prisoners bear a moral obligation to undertake a joint strategy.[31] If they were morally obligated, they would have reason to cooperate. But we doubt that a reasonable morality would obligate persons to cooperate under conditions in which there is no expectation of cooperation by others. Such a morality may well be thought to be overly demanding, if not plainly irrational. There is a gap between what morally would be better – both persons adopt a joint strategy and hence the morally relevant interests of both are best served – and what could be morally obligatory for each person.

Certain scenarios, such as those in which leaders of nuclear powers cooperate in refraining from striking first, may appear to provide a counterexample.[32] Cooperation has ensued among parties that have reason to mistrust one another. Yet it is important to be clear about whether the scenario in question retains the structure of a prisoner's dilemma; each party, in refraining from cooperation, must appear to do better from the perspective of individual rationality as compared to the cooperative alternative. This is not the case in nuclear deterrence scenarios. Victory in a nuclear war may be nothing more than a Pyrrhic victory: winning the war may not be worth the resulting devastation to the victorious side, to say nothing of the devastation overall. Moreover, deterrence scenarios extend over time in such a way as to constitute a repeated prisoner's dilemma. As we earlier acknowledged, cooperation as a strategy may evolve over time in response to 'tit for tat' reinforcements. Indeed, this process may serve to solidify

[31] Virginia Held, 'On the Meaning of Trust', *Ethics* 78 (1968), pp. 156–59.
[32] Held, 'Rationality and Reasonable Cooperation', p. 728.

expectations, at least regarding the other side's rational pursuit of its self-interest. Such scenarios represent a different game than the game we have considered. It could be, however, that over time a relationship of mutual concern and trust emerges.

Our examination of the non-repeated prisoner's dilemma raises the prospect that the rational claims of morality presuppose commitment to cooperation.[33] In other words, morality may presuppose what we have been referring to as trust. Morality would require not just that persons have some other-regarding aims but also a mutual commitment to those aims. Acting on trust would not involve a leap of faith. Rather, it would preserve a relationship that the participants value. Without trust, rationality forces persons to think more strategically about how to advance their aims. Persons may have other-regarding aims but lack the moral community enabled by basic, mutual concern. The trust relevant to 'solving' the prisoner's dilemma and underwriting moral practices has to be earned–with its elements of shared goals, reciprocity, and some common understanding of the conditions of another person's reliability–and it may hold to a greater or lesser degree.[34] The depth of mutual commitment and the dynamics of vulnerability thus serve to define the possibilities and limits of moral relationships.

We have argued that the standard, non-repeated prisoner's dilemma poses no true dilemma. The solution that is supposed to be puzzling – because each prisoner confessing is worse for both of them than the alternative on which they both remain silent – is the rational solution for persons who care about their own well-being and have no reason to trust one another. It is rational for them to be determined to avoid the individually worst outcome, especially when that outcome is quite bad. Philosophers have been misled by the apparent availability of the better alternative. From a rational perspective, that option is not really available, even for a prisoner who is not thoroughly self-interested. It is not rational to gamble on what would be mutually best when the individual stakes at the losing end are high, the odds are not

[33] We find support for this claim in Yi's criticism of David Gauthier. Yi argues that Gauthier's contractualism relies on a moral constraint on the direct pursuit of individual interest that Gauthier claims is absent in the concept of rationality alone. See Yi, 'Rationality and the Prisoner's Dilemma'.

[34] For an interesting discussion of different forms of trust, some precarious and morally compromised, see Annette Baier, 'Trust and Antitrust', *Moral Prejudices: Essays on Ethics* (Cambridge, MA: Harvard University Press, 1994), especially pp. 123–29.

favorable, and the sure-bet alternative of avoiding the worst outcome may lead to the individually best outcome. It does not seem rational to take seriously the possibility that if each person is willing to risk unfavorable odds, together they could do better than the second worst outcome. Collective rationality is not relevant for persons who cannot be confident that they stand in a trusting relationship. And when they do have reason to be confident, there is no unresolved conflict between their individual and collective points of view.

Department of Philosophy
Tufts University
Medford, MA 02155
Erin.Kelly@tufts.edu
Lionel.McPherson@tufts.edu