RMM Vol. 4, 2013, 191–204
Special Topic: Can the Social Contract Be Signed by an Invisible Hand?
http://www.rmm-journal.de/

David Gauthier

Achieving Pareto-Optimality: Invisible Hands, Social Contracts, and Rational Deliberation

Abstract:

I begin with two simple, similar interactions. In one, maximizing agents will reach a Pareto-optimal equilibrium, in the other, they won't. The first shows the working of the Invisible Hand; the second, its limitations. Using other simple interactions in which equilibrium and P-optimality are incompatible, I argue that the rational outcome of interaction answers to optimality rather than maximization, and requires agents to cooperate in realizing an agreed outcome, rather than to seek their best reply to their fellows. The terms of cooperation are set by a social contract, which coordinates choices to achieve a Pareto-optimum when the Invisible Hand is absent.

Keywords: best-reply, equilibrium, Invisible Hand, maximization, Pareto-optimality, rational choice.

1. Introduction

Consider two very simple, seemingly similar, interaction matrices. The two agents, A and B, are farmers who are each considering erecting a dam to improve irrigation for their crops. The payoffs are represented in *0–1* normalization.

		B	Builds dam	Doesn't
A	Builds dam		*¾, ¾*	*1,0*
	Doesn't		*0,1*	*¼, ¼*

		B	Builds dam	Doesn't
A	Builds dam		*¾, ¾*	*0,1*
	Doesn't		*1,0*	*¼, ¼*

The second matrix is, of course, a form of the Prisoner's Dilemma.

The two interactions access the same payoffs. If we allow the full panoply of randomized strategies, including joint randomization, they access the same area of expected-utility space—the area bounded by *(1,0), (¾, ¾), (0,1)* and *(¼, ¼)*. We suppose that the two farmers are expected-utility maximizers. In my view this is a substantive assumption; it is not entailed by what we mean by 'utility' and 'rational'. Evaluation and choice are logically distinct. Many economists (and

some philosophers) think otherwise. They accept revealed preference theory, which I consider a conceptual straitjacket. But I shall not embark on this here, which clearly would require at least an essay of its own.

In both interactions, it is better for each that both construct a dam rather than that neither does. The matrices represent two possibilities. In the first, each farmer would like to be a monopolist, controlling the water supply. Building a dam would achieve that end if the other did not reply in kind. If both build a dam, then they will establish a system assuring an adequate supply of water for irrigation. Neither will attain a monopoly, but each, in seeking his own benefit, will have contributed to a mutually advantageous arrangement. Hence the preferences shown in the first matrix: each would like best to be the sole dam builder, will prefer two dams to none, and will prefer none to only one built by the other farmer.

In the second matrix, once again each prefers two dams to none. But neither has an interest in constructing a dam himself. Each benefits from the other farmer's dam. Each in effect determines whether the other has an ample supply of water for irrigation, or a minimum supply. But neither is directly concerned with the water available to the other. Any concern with the other farmer's water is indirect, arising from the possible effect of the other's water supply on one's own. How might this arise?

In the first interaction A does better to build a dam, whatever B does. And since the interactions are symmetrical, B also does better to build a dam, whatever A does. So both will undertake dam building, and although each would have liked the other to refrain, if both were to do that the cost to each would be substantial—half the difference between one's best possible and worst possible payoffs from the interaction. Joint dam building is Pareto-optimal (henceforth P-optimal), which is to say that neither can do better than the payoff from both building dams, without the other doing worse.

In the second interaction A does better to refrain from dam building whatever B does. B also does best to refrain. So no dam is built, although they would each be better off were both to build one. The outcome they reach falls far short of P-optimality, so that each could receive a greater payoff while the other's payoff also increased. The tension between actions that are utility-maximizing, and payoffs that are P-optimal is my theme. But can there possibly be anything new to say about it?

The first situation plausibly illustrates the working of the Invisible Hand. For each seeks to benefit himself by assuring an adequate supply of water for irrigation, and the outcome achieves this for both, though neither intended joint benefit. What does the second illustrate? Again, each seeks only to increase his water supply, but he achieves nothing. Do we detect the presence of a malign Invisible Hand, who denies each the payoff that he could have had without overall cost to the other – indeed, with positive benefit to the other? Or is the Invisible Hand always benevolent—when he is around? Does he just not show up for the second interaction?

Whatever we say here, what is evident is that not all interaction naturally comes to a happy ending. It depends on the structure of the situation—in this case, the relation of outcomes that are possible but that no theory of rational choice could ever recommend. Any consideration a person might propose in support of one of the outcomes that she would find best and the other worst would be countered by an equal and opposite consideration supporting one of the outcomes that he would find best and she worst. A theory of rational choice must make the same recommendation for each since the two persons occupy strictly symmetrical positions.

Suppose our first interaction were characteristic, in that we could normally expect that if each person concerned himself with and only with his own benefit, the resulting outcome would be P-optimal. We should then be living in the anarchist's paradise, and most of our existing social institutions and practices would be rejected as distorting the natural harmony among human beings. The workings of the Invisible Hand would lead us to a mutual benefit that was no part of anyone's intention, but that would obviate the need for social controls.

I do not believe we live in such happy circumstances. If our second interaction were characteristic, then P-optimality might seem unattainable. That would be an unwarranted counsel of despair. Social institutions and practices exist to alter either the net payoffs of our interactions, or the way in which we relate our preferences to our actions, so that we may approach P-optimality.[1] Indeed, we should want social institutions and practices to yield outcomes that each person would expect to be both P-optimal and, when possible, P-superior to the outcome that she could expect in their absence. But how successful these alterations could be, cannot be determined at the present abstract level of our discussion.

I also do not believe that *most* of our interactions are Prisoner's Dilemmas. But many of them are, and it seems to me that the only way to approach P-optimality requires more extensive social institutions and practices than anarchists, or libertarians, or followers of Nozick, will readily acknowledge. But I shall not try to argue this here.[2]

2.

We have moved very rapidly from considering particular, structurally simple interactions to discussing some of the most general and abstract features of the

[1] This is a normative claim disguised as a factual one. Our social institutions and practices exist for all kinds of reasons and alleged-reasons. What I mean, of course, is that *justifiable* institutions and practices shape our interactions in P-optimizing ways. My claim is on a par with Rawls; "Society is a cooperative venture for mutual advantage." Tell it to the Egyptian laborers who built the Pyramids.

[2] The argument depends on the extent of negative externalities in our societies, and the best means of eradicating them.

circumstances in which interactions occur. Let us return to the beginning, and examine the reasoning of maximizers as they face the problems that arise in the very simple interactions that were our starting point. In these interactions, there is a unique outcome that must be attained by players each of whom maximizes his expectation of utility. Any other outcome can be reached only if at least one of the players does not perform his maximizing action. In these interactions, maximizing is very easy because each has an action that maximizes his utility whatever the other does. In less simple interactions, each player's maximizing action may depend on the actions chosen by the other players. We may suppose that each forms an expectation of what the other will choose, and tailors his own choice to maximize his utility given this expectation. He seeks his *best reply* to what he expects the other (or others, if there are three or more persons interacting) to choose.

It may not be possible for every person to perform his actual best reply to the other or others. There may be no n-tuple of pure strategies, one for each person such that each is the agent's best reply to the others. But if we allow randomized strategies, then we have Nash's proof that it is always possible for everyone to perform an expected best-reply—a randomized strategy with an expected utility at least as great as the expected utility of any other strategy, pure or randomized. Expected best-reply outcomes are Nash-equilibria. If (as economists and rational choice theorists suppose), a rational player seeks to maximize his utility given the actions of the others, then if each has correct expectations about the others, the outcome of interaction among rational players must be a Nash-equilibrium.

For economists and rational choice theorists, an interaction can be fully rational only if the outcome is in equilibrium. Equilibrium is a necessary, and not a sufficient, condition of rationality, for there may be multiple equilibria each of which will be a possible outcome of an interaction. But my concern is not with how rational players would coordinate on one outcome, or how equilibria may be found in situations much more complex than those we have been examining. It is sufficient for my purposes to conclude that *if* a theory of rational interaction prescribes a strategy for each person, then on the received account of practical rationality, each strategy must be a best reply for the player to all the other prescribed strategies, and so the result is a Nash-equilibrium.

A Nash-equilibrium need not be P-optimal. And in some interactions, the PD being the simplest example, no outcome is both a Nash-equilibrium and P-optimal. The received account of practical rationality, being wedded to equilibrium, must reject P-optimality as a condition on rational interaction. The received account must allow that an interaction may be fully rational, yet its outcome may leave on the table benefits that could have been enjoyed by some persons at no cost to the others. Advocates of the received account will suppose that if all of the equilibria in an interaction were sub-optimal, then it would be desirable were the payoffs of some of the outcomes to be altered by sanctions or rewards, or were the preferences of some of the players to be altered endogenously, or were the outcome to affect the reputation of the players and so their

future payoffs, or were the interaction to be repeated indefinitely with players' future choices dependent on expectations arising from others' present choices.

So let us suppose that in our PD the farmers make an agreement that each will build a dam—an agreement which, if it were kept, would yield each his second-best outcome. But building a dam is neither farmer's best reply to his fellow constructing one. On the received view, any freestanding agreement must fail. But if sanctions could be imposed for not building a dam, or if failing to fulfill one's agreement to build a dam affected one's reputation adversely and increased the likelihood that one would be barred from future agreements, or if the farmers expected to encounter other choices similar in kind and adopted a tit-for-tat strategy, keeping one's agreement if the other had kept his on the most recent similar occasion, or if the farmers simply developed a preference for dam building—if any of these considerations could be introduced, then the structure of the interaction would be changed and dam- building, which is P-optimal, could be a mutual best reply.[3]

But all of these are *ad hoc* devices for getting persons from mutually undesirable to mutually desirable outcomes when best-reply reasoning in itself would fail to do so. Something is added to the situation—a sanction, a reputation, a repetition—which is no part of the interaction itself. An endogenous preference change is somewhat different, in remaining with the original interaction. But the structure of the interaction is supposed to be transformed by the recognition that by changing one's preferences, one can better satisfy them. If one's preferences have really changed, then why should one be concerned about the satisfaction of one's original preferences? I do not deny the reality of adaptive preference formation, and that it typically occurs when one is failing to get what one wants, but it also typically leads the person to want what she can get. It is hardly a resolution to the Prisoner's Dilemma to say that it dissolves if the players come to prefer the equilibrium outcome to the outcome originally P-superior to it. Or that it dissolves if the players interchange their initially most and least preferred outcomes, so that the PD is converted into the first of the interactions we examined.

3.

I want to free our minds from the dogma that individual actions are rational only if maximizing, while keeping hold of the deeper idea that rational agents seek to bring about what they most value. Then we shall take the achievement of a P-optimal outcome as a mark of rationality, since a P-optimal outcome affords every person as much as possible of what he values given what it affords the

[3] But devices to ensure compliance may be costly. And these costs may well fall on the parties involved, leaving them less well off than had they simply complied voluntarily, The outcome may be P-superior to the outcome if no dam is built, but not P-optimal because of the costs of compliance. The economist dismisses talk as cheap; I praise it because it is free.

other persons. Not every P-optimal outcome will be rational; in particular, if a P-optimal outcome gives some person less than he could expect to achieve were everyone to choose a best reply, accepting it would not be rational for him. We might suppose that a P-optimal outcome would be rational, only if it was P-superior to any Nash-equilibrium. But in some situations, although no Nash-equilibrium is optimal, no P-optimal outcome is P-superior to all of the Nash-equilibria. For every non-optimal Nash-equilibrium, there must be a P-superior P-optimal outcome; but there may be no P-optimal outcome that is P-superior to every Nash-equilibrium.

We can however say that *if* a best-reply theory of interaction prescribes a strategy for each person, then a P-optimizing theory of rational interaction prescribes a strategy for each person that, taken together yield an outcome P-superior to the outcome of the prescribed best-reply strategies. I propose this as a necessary condition on practical rationality. Let us call it Condition P-O.

Persons *cooperate* if, in a situation in which no equilibrium outcome is P-optimal, they agree to a set of strategies satisfying Condition P-O, and each executes his agreed strategy provided he may expect similar execution from the others. We may draw two immediate and fundamental implications from this definition:

1. Cooperators may expect to bring about more of what they most value than those who interact on a best-reply basis.
2. Cooperation would be irrational if (a) only maximizing actions were rational, or if (b) only best-reply reasoning were rational.

Since the cooperative outcome differs from the outcome of best-reply in just those situations in which it is P-superior to it, the first implication may appear trivial. But in addition to situations in which the outcomes differ, cooperators may expect to find themselves enjoying opportunities from which maximizers are excluded, because the benefits depend on establishing between persons, relations of trust that cannot be accommodated in best-reply reasoning.

The upshot should be evident. The orthodox theory of practical rationality, embraced by economists and theorists of rational choice, must treat cooperation as in itself irrational. But everyone may expect to benefit from cooperative interaction. The orthodox theory is therefore mistaken. It should be superceded by a theory based on Pareto-optimality and cooperation.

But the orthodox view may not simply be dismissed. There are contexts in which cooperation would be misguided—contexts in which Adam Smith's Invisible Hand does lead individuals who seek to maximize their own utility to a P-optimal outcome. Attempting to introduce cooperation in such situations imposes unnecessary costs, and so undermines the pursuit of P-optimality. It is unfortunate that the theory of rational choice has been so closely linked with economics, and that the theory of competitive markets has led us to think of human interaction as generally conforming to a maximizing standard. But certainly some important areas of human interaction have been illuminated by the economist's approach.

Furthermore, there are contexts in which cooperation could provide benefits to all, but ignorance, disagreement, ideological blindness, make it unreasonable to expect many or most persons to behave cooperatively. If a person cannot expect others to follow cooperative guidelines that would lead to an outcome P-superior to the outcome of expected best reply reasoning, then if she is rational she will not let herself be used by others, and will look to her own interest. Best-reply reasoning offers a fallback position when genuine cooperation, on terms that are mutually beneficial, is not to be had.

And we must recognize that while cooperation does not require any common good, or sense of common purpose, beyond the demand that each cooperator benefit in his own terms, yet some possible values cannot be promoted together. Some situations are zero-sum, so that a gain for one person must be a loss for some other person. When I took a course in anthropology as a college student, I was introduced to a South Pacific people called the Dobu, who, it was claimed, saw the possession of yams as the greatest good, but who thought that yams were in fixed supply, so the more in my garden, the fewer in yours. Strategic alliances might be possible, so that you and I could join forces to deprive some third party of his yams, but we should then fall out over the division of our spoils. Genuine cooperation would be impossible, since there is no way of achieving overall mutual benefit. In such a world, reasoning in a cooperative manner would lead nowhere.

That is not our world. And in our world it is reasonable for persons to think of themselves as cooperators, seeking opportunities to interact on the basis of mutual benefit. But lacking the Invisible Hand, cooperation in all but the simplest interactions requires the solution of a coordination problem—each person's action must be directed towards the same Pareto-optimal outcome, and as I have noted, there may be many such outcomes, differing in their payoffs to the interacting persons. If we think, with Rawls, of society as a whole as properly being "a cooperative venture for mutual advantage" (Rawls 1971, 4)—I should prefer 'fulfillment' to 'advantage'—then we should think of institutions, practices, roles, expectations—the normative world into which each of us is born and raised, as addressing two problems. First they must address the *coordination* problem—choosing one from the set of P-optimal outcomes. And then they confront the *cooperation* problem—acting so that the outcome chosen is realized, whether or not it satisfies the conditions of best-reply reasoning.

4.

And now we have an explanation of the social contract. The contract is a hypothetical agreement, an agreement that persons would make were they in a position to choose, with their fellows, the terms of their interaction. The question, 'what would you agree to if . . . ?' is, I think, readily intelligible and does not reduce to any question in which the idea of agreement is eliminated. But its

practical role is limited. In particular, we cannot transfer the binding force of actual agreement to hypothetical agreement. That I would have agreed yesterday to sell you this painting I found in my attic for $50 does not have any bearing on the situation today, when I have learned that it is in fact a sketch by Lawren Harris.[4]

The social contract occupies a unique position in assessing norms of interaction. We are obviously not able to choose most of the terms on which we interact with our fellows. We are born into a particular society and we internalize its norms as part of normal development from child to adult. But we may come to reflect on the claim of these norms to provide us with genuine reasons for acting. And our reflection may suggest that the binding power of these norms depends on whether they would be agreed to by persons were they able to choose, with their fellows, in an initial position of equality, the terms of their interaction. Substantive social values would result from their agreement, and not be presupposed by it.

Actual human societies are at best imperfect approximations to what we may call the contract society. Applying the contractarian test to actual social institutions, roles, and practices, would no doubt result in major social changes. And of course many societies would reject the test, as incompatible with their view of social values as imposed by (divine) authority, or some alleged 'objective' standard. But these are fictions that mask arbitrary power. The social contract justifies the exercise of power in society only as instrumental in enabling individuals to reap the benefits of cooperation. It does not subordinate any person's concerns to a fictitious 'collective good', but coordinates individual concerns the better to advance each.

The social contract works through the normative demands and expectations society addresses to its individual members. This is not a matter to be pursued at length here. We ordinarily treat the demands and expectations as considerations that we accommodate in our deliberations as reasons for acting. Examples abound, serious and trivial. Skimming through a ladies' fashion magazine as I waited for a medical appointment, I found my eye caught by the word 'expected' in an advice column. The question concerned the propriety of bare legs and sandals for women in the work place, and the answer was favourable, but with a cautionary note, "You will be expected to paint your toenails". The impersonal phrasing indicates a social expectation, rather than an expectation of particular individuals. As such, it seems a poor candidate to pass the contractarian test, which accepts only practices that are necessary to terms of cooperation that all persons might accept. If individual autonomy is a fundamental concern of a cooperative society, personal adornment will be an area of freedom, not of imposed expectations. The example should remind us of how pervasive social expectations are, and how great will be the task of weeding out those that impose straitjackets from those that advance individual liberty and prosperity.

[4] A member of Canada's most esteemed school of painting, the Group of Seven.

The social contract answers to Pareto-optimality but not to best-reply or equilibrium. It underlies what we can call social reasons for acting—practical considerations that follow from the practices and expectations that the contract validates. These considerations bear little resemblance to those based directly on a person's utility function. But our discussion should have made it evident that these considerations are not unrelated to the agent's utilities, since acting on them, when he may expect others to do so as well, affords him greater utility than were each person to treat utility-maximizing and best-reply considerations as providing their sole reasons for acting.

Reasons for acting cannot be formally idiosyncratic. In similar circumstances, and with similar capacities and values, different persons must have similar reasons for acting. That a particular action would contribute to attaining an agreed Pareto-optimal outcome is reason to perform it. Such an action is of course not always available. But genuine cooperation among persons is possible only if they endorse the rationality of Condition P-O.

5.

I have claimed that in the context of a cooperative agreement, it may be rational for persons to perform non-maximizing actions. I have noted that the orthodox view, held by economists and rational choice theorists, denies this. I have used this denial as ground for rejecting the orthodox view. I want now to show that the orthodox view must deny the rationality of achieving mutual benefit in temporally extended interactions in which the last person receives his share of the benefit before he acts.

Let us begin with a familiar example. Two farmers, call them Fred and Ed, each with a crop to harvest. Fred's crop is ready now, Ed's will be ready next week. They can go it alone, or they can work together. The work will be easier for each if they work together. So, as they are having a pint in the pub in town, they agree to work together. "Let's drink to that", says Fred. "Another round", Ed calls to the barmaid.

The economist, who has been nursing his sparkling water in the corner, comes over to their table. "Planning to harvest together?", he asks, rhetorically, since he has overheard their conversation. "That's the idea", says Fred. "More hands, lighter work." The economist smiles his when-will-they-learn smile. "Aren't you fellows forgetting something?" "What?" says Ed, "what could we be forgetting?" The economist sighs. "This week you (addressing Ed) are going to help him (looking at Fred)—" Ed breaks in "And next week he's going to help me." "But why ever would he do that?" the economist asks. "If you help him this week, he's got what he wants. Helping you next week would be a sheer cost to him."

Fred, who had been paying little attention to the direction of the conversation, is suddenly all ears. "What's this?" he asks. "I get helped without having to

make any return?" "I'm afraid not", the economist replies. "Since you have nothing to gain from helping your friend here with his harvest, it would be irrational of you to help. So you won't. But your friend can reason this out too, and won't then expect your help. And so helping you would be a sheer cost to him, since it won't gain him your help. So it would be irrational of him to help you now. So he won't. You're surely both rational men? Rational men harvest alone." Fred says "But—", and realizes he has nothing to say. Or at least nothing relevant (as we shall see). Ed just sits there, staring into his beer. The economist smiles his always-willing-to-assist smile, and makes his departure. The farmers stare glumly at each other.

An immediate reaction to this story is "But next year!" Another is "What about Fred's reputation?" A third is "But aren't they friends?" And a fourth—"Fred could post a bond." But these replies miss the point, as Fred presumably realized when he found himself with nothing to say. For surely the mutual advantageousness of the agreement to harvest together is all that is needed to make it rational. Whether or not there is a next year, or Fred has a reputation to keep, or he would not want to benefit at a friend's expense, or he would run afoul of the law, the agreement is in itself advantageous and that should be enough for it to be rational for Fred and Ed to make and comply with it.

The orthodox theory cannot accommodate this. An action is rational only if performing it would maximize the agent's expected utility. That it is part of an advantageous set of actions is in itself irrelevant to its rationality. When a person who is rational by the standards of the orthodox theory decides what to do, he considers only the consequences of performing or not performing each of his possible actions. Of course, a person may value honoring his commitments, so that for him, that an action would honor a commitment would be a consequence entering positively into his utility-function. But this is a contingent matter.

When we make an agreement, or undertake a commitment, we normally suppose that we give ourselves a reason for acting that previously we did not have. A person benefits from being able to do this. But on the orthodox view, the commitment in itself does not create a reason for him to act unless it affects his preferences. And its strength as a reason is determined by the extent of its effect on these preferences. So when Fred considers whether he should assist Ed with his harvest, he looks, not to the mutual advantage gained if and when they assist each other, but solely to the benefit he would then receive, were he to help Ed. Any benefit he has received from Ed would be relevant to his deliberation only insofar as it affects his current preferences.

Suppose Fred and Ed could each make a single decision that would fix their behaviour in harvesting.[5] It is evident that Fred would decide to assist Ed provided Ed had assisted him, and Ed would decide to assist Fred provided Fred had decided to assist him provided he had assisted Fred. And they would harvest together. On the orthodox view, one cannot simply make a single upfront decision fixing one's subsequent actions, but must take every decision as and

[5] Whether they might be able to do this, and how they would then do it, I ignore.

when it comes. The upshot is that the time at which actions occur becomes relevant to the outcome of interaction. An outcome that could be achieved rationally by a single decision may not be achievable rationally by multiple decisions, tied to the times of performance. This, I would argue, is to give time a role in determining rational behaviour that is unwarranted. If persons can achieve benefits by increasing the scope of their decisions, so that a single decision could cover a sequence of actions, or a plan, or a policy, then they must rationally do so. A full theory of rational deliberation would address the appropriate scope of decisions. But for our purposes, we need only note that treating each decision separately, and viewing the reasons for taking it only in terms of its consequences, is to treat rationality as a hindrance to attaining mutual benefit in many temporally extended interactions.

6.

On the orthodox view, at every choice point in an interaction, a rational person must choose the action that maximizes his expected utility. Choosing this way, Fred and Ed harvest alone. Something is amiss here. So let us make matters worse. Consider a version of the well-known Centipede puzzle. Two persons, call them Odd and Even, can win up to $2000 between them. They are to play a game with 100 possible moves, taking turns, so each has up to 50. Odd moves first (he has the odd-numbered moves). A move consists of either terminating the game, or giving the other player the next move. If the game terminates at move n, the player who ends the game receives \$10(n + 1), the other receives \$10(n – 1). If the game is terminated on the first move, Odd receives $20 and Even $0, and on the hundredth (and last) move, Odd receives $990 and Even $1010. If the players do not terminate the game within the hundred moves, each receives $1000.

What should the players do? They may not communicate with each other (apart from the communication of their moves), offer or accept side payments, bind themselves legally to one or more choices, or do anything else that would affect their behaviour or payoffs. And they may not commit themselves by a single, upfront decision. Suppose that they reach the hundredth move, which belongs to Even. Termination gives her $1010, otherwise she gets $1000. On the orthodox theory of rationality, she must terminate. This is common knowledge. So at the ninety-ninth move, should they reach it, Odd knows that if he does not terminate, Even will on her move. Termination gives him $1000, letting Even move and terminate gives him $990. He must terminate.

The argument repeats itself all the way back to the beginning. Should they reach the n^{th} move, the player with the move knows that if he or she does not terminate, the other will on the next move—and it will cost the player with the move $10. So he or she must terminate now, whenever 'now' is. They must

reach the first move, and following the orthodox path, Odd must terminate. He pockets $20. Even's pocket remains empty. $1980 remains on the table.

But it is surely a *reductio ad absurdum* to suppose that in the Centipede puzzle, rational players would abort the game immediately. A theory of rational choice that has this as a result is plainly unsound. What is more, as Brian Skyrms has shown for a relevantly similar game (Skyrms 1990, 133; see also Pettit and Sugden 1989, 182), it is within the power of the players to defeat the presumption of common knowledge of rationality, and to do so in a way that yields them larger payoffs than the supposedly rational outcome. Believing Even will follow suit, Odd does not terminate the game immediately. Aware of Odd's choice, Even infers that he expects her to follow suit, since otherwise his choice would be irrational. And she will follow suit, if she believes, as he does, that he will again refrain from terminating the game. In regaining the move, Odd knows that the worst possible outcome for him would yield him $30, a gain of $10 on terminating initially. And continuing, Odd again refrains from terminating, and Even knows that the worst outcome that can now befall her would give her $40, $10 more than she would have received had she terminated on her first turn. So the 'rational' outcome is not utility-maximizing. The orthodox view seems to lead to a contradiction.

This contradiction is avoided if we reconceptualize the Centipede game in cooperative terms. Odd gives Even the move. Even makes sense of Odd's decision by supposing Odd to be proposing mutually advantageous cooperation. If he and Even both refrain from terminating the game, then both benefit. If Odd were to terminate, he would get $20 but Even would have nothing. If Odd gives Even the move and she terminates, Odd would get only $10 and so would have done better not to have given Even the move. But if Even returns the move to Odd, then since she must get at least $20, each will have done better than if Odd had ended the game—even if Odd does not return the move to Even. Each pair of moves constitutes an interaction between Odd and Even in which both can benefit provided each gives the move to the other.

Note that each pair of moves can be treated as self-contained, on this cooperative analysis. The rationality of not ending the game does not depend on the expectation that the other person subsequently will not end it. It depends on the benefit each receives from the other not ending it as his or her present move. And so Odd continues to turn the move over to Even who returns it to him, and each pair of interactions adds $20 to each person's payoff.

What happens at the end? If the game goes to the hundredth move, Even can end it and receive $1010 or accept the imposed ending and receive $1000. If she ends it, Odd gets $990; the automatic ending would give him $1000. By ending the game on the 99th move, Odd could walk away with $1000; he would not have to depend on Even not ending the game. But Even would get only $980. Recall Principle P-O. Cooperation aims at P-optimality. If Odd terminates the game on the 99th move, $20 will be left on the table. So he will not terminate, but will turn the move over to Even. And she will accept the imposed ending, forgoing the additional $10 she could receive by terminating, but leaving Odd as well off

as he would have been had he terminated the game at the 99th move. Each then pockets $1000.

The cooperative analysis of the game raises none of the problems faced by the orthodox maximizing analysis. Common knowledge of rationality coheres with rational cooperative choice. The rational outcome is not the absurd ending of the game at the first move, and there is no parallel to the contradiction between the strategy called for by backwards induction, and the strategy that actually maximizes the players' expected utility. In the Centipede puzzle, the orthodox understanding of practical rationality fails; the cooperative understanding carries the day.

7.

I began with two seemingly similar interaction matrices. A sound theory of rational deliberation will preserve that similarity. In the type of situation exemplified by the first matrix, the Invisible Hand does the work necessary if we are to treat each persons as deliberating in interaction as he would deliberate when only he is involved. But this work will not yield a general theory of deliberation, since outside the privileged enclave of the competitive market, the Invisible Hand loses its power. In the type of deliberation exemplified by the second matrix, the persons engaged in interaction recognize the inadequacy of deliberating as solitary persons. In our very simple example, Pareto-optimality and symmetry are sufficient to pick out a single outcome, which persons can realize by cooperating with each other. But cooperation typically involves setting a P-optimal goal, reflecting the concerns of the several parties to the interaction, and the prescription for each of an action which, when combined with the actions prescribed for the others, reaches this goal. The action prescribed for a person need not be his or her best reply to the actions prescribed for the others, but rather what, taken with the actions prescribed for the others, will result in a P-optimal outcome. We have seen this at work with the farmers, Fred and Ed, and the Centipede players, Odd and Even. And we have seen the disaster in attempting to treat these interactions from the perspective of the agent alone.

The social contract embodies the intent manifest in cooperation. It thus emerges as, not a rival to, but rather a counterpart of, the Invisible Hand. When cooperation is required to reach a P-optimal outcome, the contract sets out what persons would agree to were they determining *ex ante* their terms of interaction. These terms must afford a place to cooperation at the forefront of any adequate theory of rational deliberation.

References

Pettit, P. and R. Sugden (1989), "The Backward Induction Paradox", *The Journal of Philosophy* 86, 169–82.

Rawls, J. (1971), *A Theory of Justice*, Cambridge/MA: Harvard University Press.

Skyrms, B. (1990), *The Dynamics of Rational Deliberation*, Cambridge/MA, Harvard University Press.