The Backward Induction Paradox

Philip Pettit; Robert Sugden

The Journal of Philosophy, Vol. 86, No. 4. (Apr., 1989), pp. 169-182.

Stable URL:
http://links.jstor.org/sici?sici=0022-362X%28198904%2986%3A4%3C169%3ATBIP%3E2.0.CO%3B2-8

The Journal of Philosophy is currently published by Journal of Philosophy, Inc..

THE JOURNAL OF PHILOSOPHY

VOLUME LXXXVI, NO. 4, APRIL 1989

THE BACKWARD INDUCTION PARADOX*

SUPPOSE that you and I face and know that we face a sequence of prisoner's dilemmas of known finite length: say n dilemmas. There is a well-known argument—the backward induction argument—to the effect that, in such a sequence, agents who are rational and who share the belief that they are rational will defect in every round. This argument holds however large n may be. And yet, if n is a large number, it appears that I might do better to follow a strategy such as tit-for-tat, which signals to you that I am willing to cooperate provided you reciprocate. This is the backward induction paradox.

Although game theorists have been convinced that permanent defection is the rational strategy in such a situation, they have recognized its intuitive implausibility and have often been reluctant to recommend it as a practical course of action. We believe that their hesitation is well-founded, for we hold that the argument for permanent defection is unsound and that the backward induction paradox is soluble.

I. THE PARADOX

The argument involved in the generation of the paradox involves a familiar sort of backward induction. Suppose that two players A and B face and know they face a finite sequence of n prisoner's dilemmas. Suppose also that they are both rational and that their rationality is a matter of common belief: each believes each is rational, each believes each believes this, and so on. Under those assumptions, it seems that either is in a position to run the following induction:

> My partner, being rational, will defect in the nth round of the sequence, since defecting at that stage will not have any undesirable effects in further rounds—there are none—and since it will dominate coopera-

* This paper was written while Sugden was a Visiting Fellow at the Research School of Social Science, Australian National Univ. We are grateful for a helpful discussion when it was presented at a seminar in the Department of Philosophy.

0022-362X/89/8604/169–182

> tion, just as in the one-shot prisoner's dilemma. Since he believes that I am rational, my partner will also expect me to do likewise in the nth round and so, being rational, will himself defect in round $n - 1$: there will be no undesirable effects in further rounds—the nth result is already fixed—and defecting will dominate cooperation. But, believing that I believe that he is rational, my partner will also expect me to do likewise in round $n - 1$ and so, being rational, will himself defect in round $n - 2, \ldots$.

The backward induction argument is that such an induction will lead each to the conclusion that his partner will defect in every round and that he therefore ought to do the same. If the argument is sound, then in this situation *a rational player will never cooperate.*

But can this be right? Two players who always defect will do worse in every round than if they had cooperated; so it is in both their interests to arrive at some cooperative understanding. Of course, this is also true of the one-shot prisoner's dilemma. But in the one-shot game the players have no means of communicating before they make their moves; and even if they *were* able to reach an agreement before the game they would have no means of enforcing it, for once a player has made a move it is too late for his partner to respond in any way. In the repeated game, in contrast, the players can communicate during the course of play by means of their choice of moves, and, in every round except the last, each player has to take account of the possibility that his choice of move may influence the later moves of his partner. Intuition surely suggests that, at least if the number of rounds is sufficiently long, a player would do better by following a conditional strategy such as tit-for-tat (that is, the strategy whereby a player cooperates in the first round and then does whatever his partner did in the previous round) than by permanently defecting. By playing such a strategy, a player signals his willingness to cooperate provided his partner will do the same, and if the signal is read correctly, it will be in the partner's interest to cooperate—in the case of tit-for-tat, in every round except the last. And even if the signal is misunderstood, little is lost.[1]

Game theorists often concede that reasonable (if not perhaps rational) people would be better advised to play conditional strategies than permanently to defect. Reinhard Selten,[2] for example, writes that "the [backward] induction theory is the game theoretically correct one but the co-operation theory [i.e. the recommendation to play conditional strategies] seems to be the better guide to practical

[1] The robustness of tit-for-tat—that is, its ability to do well against most of the strategies that a partner might plausibly play—is explained by Robert Axelrod, *The Evolution of Cooperation* (New York: Basic Books, 1984). See also Philip Pettit, "Free Riding and Foul Dealing," this JOURNAL, LXXXIII, 7 (July 1986):361–379.

[2] "The Chain Store Paradox," *Theory and Decision,* IX, 2 (1978):127–159.

behaviour" (*ibid.*, p. 138). Similarly, Duncan Luce and Howard Raiffa,[3] while endorsing the theoretical validity of the argument for permanent defection, remark that they do not think it "reasonable" to single out this solution to the game. "By reasonable," they write, "we mean that we predict that intelligent players will play accordingly and, furthermore, that they will still do so after a full airing of the 'theory'."

The conclusion that rational players must defect seems logically inescapable, but at the same time it is intuitively implausible. That is the backward induction paradox.

II. THE SOLUTION

The solution to this paradox is surprisingly straightforward, paralleling the solution offered by Frank Jackson[4] to a version of the surprise examination paradox. The solution consists in recognizing that neither player is in a position to run the backward induction required.[5]

On careful exposition, the induction depends on the capacity of each of the players, before making his first move, to run the following sequence of arguments:

Argument 1. In round n, my partner will act rationally; therefore, he will defect.

Argument 2. In round $n - 1$, my partner will act rationally. In addition, he will then believe that I will act rationally in round n; therefore, he will believe that I will defect in round n. Given this belief, he will defect in round $n - 1$.

Argument 3. In round $n - 2$, my partner will act rationally. In addition, he will then believe, not just what he believes at round $n - 1$, but also that in round $n - 1$: (i) I will act rationally and (ii) I will believe that he will act rationally in round n. Therefore, in round $n - 2$, my partner will believe that I will defect in round $n - 1$. Given these beliefs, he will defect in round $n - 2$.

And so on, up to argument n. Taken together, these arguments yield the conclusion that the partner will defect in every round; thus, the only rational response is to defect in every round too.

But it is mistaken to think that a player is in a position to run these arguments before making his first move. At this point, he believes

[3] *Games and Decisions* (New York: Wiley, 1957), pp. 97–102.

[4] See his *Conditionals* (New York: Blackwell, 1987), ch. 7. See also Crispin Wright and Adrian Sudbury, "The Paradox of the Unexpected Examination," *Australasian Journal of Philosophy*, LV, 1 (1977):41–58.

[5] After completing this paper we discovered a rather similar critique of backward induction, but in the context of a different game, in Ken Binmore, "Modeling Rational Players," *Economics and Philosophy*, III, 2 (1987):179–214.

that his opponent is rational, he believes that his opponent believes he is rational, and so on; this we have already assumed. But that does not entitle him to believe that in subsequent rounds his partner will still believe he is rational (etc.), irrespective of how he, the first player, has acted in the interim. Thus, for example, argument 2 is unavailable to him, because from the premise that my partner *now* believes that I am rational, I cannot derive the premise required, that in round $n - 1$ my partner will believe I am rational. Similarly, argument 3 is unavailable, because from the premise that my partner *now* believes that I *now* believe he is rational, I cannot derive the extra premise required there, that in round $n - 2$ my partner will believe that in round $n - 1$ I will believe that he is rational; and so on.

To put this point more generally, let p be any proposition that, if true at time t_1, is also true at the later time t_2. Let E be an event which may occur between t_1 and t_2. Suppose that at t_1 I believe that p. Then at t_1 must I also believe that, regardless of whether E occurs or not, I will still believe that p at t_2? Not if E is an event that I believe will not occur. If at t_1 I believe that E will not occur, then at t_1 I may believe that, were E to occur in the meantime, I would not believe p at t_2. On 31 December I may believe that it will not snow in Queensland in February. But on 31 December I may also believe that, were there to be heavy snowfalls in Queensland in late January, on 31 January I would not believe that it would not snow in Queensland in February. There is no contradiction here, provided that on 31 December I do believe that there will not be heavy snowfalls in late January.

The backward induction is supposed to show each player, at the start of the game, that rationality requires him to defect in every round—and in particular, in the first. It would indeed be irrational for me to cooperate in the first round if, regardless of whether I cooperate in that round, my partner will defect in rounds $2, \ldots, n$. But does the backward induction show this? Only if arguments $1, \ldots, n - 1$ all hold even on the supposition that I cooperate in round 1.

Argument 2 assumes that in round $n - 1$ my partner will believe I will act rationally in round n. But the only premise we are allowed is that at the start of the game my partner believes I am rational. Suppose that at the start of the game my partner also believes that I will not cooperate in round 1. Then we cannot argue that, even were I to cooperate in round 1, still my partner will believe in round $n - 1$ that I am rational. And similarly for arguments $3, \ldots, n$.

We have shown then that, just given the common belief in their rationality, neither of the players will necessarily believe that the common belief will endure. Thus, neither will necessarily be in a

position to run the backward induction. But the considerations rehearsed in showing this also enable us, with some supplement, to show something stronger: not just that neither will necessarily believe in the survival, come what may, of the common belief, but necessarily that neither will believe in this.

Consider the two players at the start of the game, holding a common belief in their rationality. We know that they cannot run the backward induction unless they believe that that common belief will survive, regardless of what either does.[6] But we can readily show that neither can believe that the common belief will survive, come what may; each must recognize that the belief will break down in the event of one of them cooperating.

We can show this, if we can show that any act of cooperation would cause the common belief in rationality to break down. For if *we* can show this, arguing just from the belief that the players are rational and have a common belief in their rationality, then each of the players will be in a position to do so too. Each of them will be in a position to see that, were either of them to cooperate, then the common belief in their rationality with which they start would break down.

We can show that any act of cooperation would cause the common belief in rationality to break down by starting with the nth round and seeing that cooperation would cause a breakdown there; then going back to round $n - 1$ and seeing that cooperation would have a similar effect in that round; and so on, making our way back to round 1.

Suppose one of the players, say A, cooperates in the final round. This is an irrational act, since defection dominates cooperation there, and so would cause B, if rational, to believe that A is irrational. In that case, the common belief in rationality breaks down at the first level.

Suppose instead that A cooperates in round $n - 1$. B may again be led to believe that A is irrational, in which case we also get a first-level breakdown of the common belief. But may B hold that A is rational? Perhaps, but only if B believes that A believes that B can be induced by A's cooperation in round $n - 1$ to cooperate in round n, which would be an irrational act. So in that case the common belief breaks down too, though now at the second level up: B does not believe that A is irrational but he believes that A believes that he, B, is irrational.

Now suppose that A cooperates in round $n - 2$. B may be led to believe that A is irrational. Alternatively, B may believe that A be-

[6] We do not mean to suggest that it would be sufficient to allow the players to run the backward induction if they began with the common belief and with the belief that it would survive; we think other beliefs would also be necessary, such as the common belief that it would survive, the common belief that the common belief it would survive would survive, and so on into nightmarish possibilities.

lieves that by cooperating in round $n - 2$ he can induce B to cooperate in at least one of rounds $n - 1$ or n. But cooperating in round n is irrational, so this belief would involve a breakdown in the common belief at level 2. And cooperating in round $n - 1$ is either irrational —in which case we also get a breakdown at level 2—or it springs from the belief that such cooperation would induce A irrationally to cooperate in round n, in which case we get a breakdown of the common belief at level 3.

The generalization should be obvious. For any act of cooperation by one player in round $n - j$, where $0 \leq j \leq n - 1$, the partner, if rational, must respond with a belief that causes the common belief in rationality to break down at level $j + 1$ or at some lower level. If we can see this, so can the players (who, by assumption, are rational). So, necessarily, neither of the players can believe that the common belief in rationality will survive whatever moves the players make.

We have now solved the backward induction paradox, for we have shown that the players are not necessarily in a position, and indeed are necessarily not in a position, to run the backward induction. One of the beliefs required for them to run the induction—the belief that the common belief in rationality would survive even if cooperative moves were played—is a belief that the players could not have.

III. THE SOLUTION SUPPORTED

The backward induction argument has been supposed to show that permanent defection is the uniquely rational strategy for each player. We have shown that argument to be unsound, and so we have resolved the paradox arising from the conflict between the conclusion of this argument and a more intuitive conclusion: that it could be rational for a player to cooperate in the first round, in the expectation of establishing some kind of cooperative understanding with his partner.

We have not yet shown, however, that this intuitive conclusion is correct, and we turn now to that task. We can show that it is correct, if we can show that there are beliefs which a player might hold consistently with the common belief in rationality—as we shall say, beliefs which he might rationally hold—which would make it rational for him to cooperate in the first round.

Intuitively, we might expect a player to do quite well by playing tit-for-tat, and in particular that tit-for-tat would be more successful than permanent defection. This does not mean, however, that it can be rational to play tit-for-tat. On the contrary, tit-for-tat cannot be a rational choice, since it involves cooperating without possible benefit in the last round: and this, indeed, in the expectation that the other, being rational (if this is still believed at that stage), will defect in that round.

Still, tit-for-tat points us in the right direction. From A's point of view, the intuitive argument for playing tit-for-tat is that by doing so he can induce B to believe that this *is* what he is doing. Suppose, to keep things simple, that B forms this belief as soon as he observes A's first cooperative move. Then, provided the sequence of rounds is sufficiently long, B will do best to respond in a way that induces A to cooperate from round 3 on. (If, as we suppose, B defected in round 1, there is no way he can prevent A from defecting in round 2.) One possible response is "delayed tit-for-tat": to cooperate in rounds 2 and 3 and from then on to repeat A's previous move. But this cannot be a rational choice for B, since he can gain nothing by cooperating in the last round. A better strategy is "delayed tit-for-tat minus 1," which is exactly like delayed tit-for-tat except that it always defects in the last round. If, after observing A's opening move, B forms the belief that A is playing tit-for-tat, delayed tit-for-tat minus 1 is a rational strategy: in terms of his beliefs, B can do no better than this.

Now suppose that A believes that these *are* the beliefs which B will form, and that this *is* the strategy which B will play. Then it is not rational for A to play tit-for-tat, but it *is* rational for him to play "tit-for-tat minus 2"—the strategy that is exactly like tit-for-tat except that it always defects in the last two rounds. By playing this, he makes the best possible reply to the behavior that he expects of his partner.

So, we have shown that there are beliefs which A might hold which would make it rational for him to cooperate in the first round. But are these beliefs ones which he can rationally hold? At first sight, it is tempting to think not, because A is attributing to B the belief that A is playing an irrational strategy, namely, tit-for-tat. But this is quite consistent with an initial common belief in rationality. According to A's belief, it is only after A has cooperated that B forms the belief that A is irrational. In other words, A believes that his act of cooperation will cause the common belief in rationality to break down. And this in itself cannot be used as an argument against the rationality of A's beliefs. Quite the contrary: as we showed in section II, a rational player *must* believe that any act of cooperation will cause the common belief in rationality to break down.

This is enough to show that there are beliefs which a player might rationally hold which would lead him to cooperate initially. But it is worth remarking that there are also many other beliefs which he might rationally hold which would lead him to do the same thing. Thus, just to stick with the tit-for-tat family of responses, there are different beliefs, all compatible with the initial common belief in rationality, which would rationalize different members of that family.

There is a pattern in the cases involved and we can catch it nicely in

tabular form. The table has the advantage that it gives an idea of how the progression goes.[7]

A rationally chooses to play:	because *A* believes that in response to an initial cooperation *B* would reason in this way:
1. Tit-for-tat minus 2	*A* is irrational and will play tit-for-tat to the end, so I should play delayed tit-for-tat minus 1.
2. Tit-for-tat minus 4	*A* is rational and believes I am rational. But he believes that I believe he is irrational and that I will play delayed tit-for-tat minus 1. Hence, he will play tit-for-tat minus 2. And so I should play delayed tit-for-tat minus 3.
3. Tit-for-tat minus 6	*A* is rational and believes I am rational. Furthermore, *A* believes that I believe he is rational. But he believes that I believe that he believes that I believe that he is irrational and that I will play delayed tit-for-tat minus 3. Hence, he will play tit-for-tat minus 4. And so I should play delayed tit-for-tat minus 5.
And so on.	And so on.

Any one of the strategies listed in the left-hand column may rationally be chosen by *A*: each is a rational response to the corresponding beliefs held by *A* about *B*. In each case, the choice *A* makes involves an initial cooperation; and the belief about *B* that makes this a rational choice entails that the common belief in rationality breaks down. In the case in which *A* plays tit-for-tat minus 2, the belief breaks down at the second level: after the first round, *A* believes that

[7] The progression cannot go on indefinitely because the number of rounds is finite. The rationale for *A*'s playing tit-for-tat minus j (where j is any positive even number no greater than $n - 2$) requires that this is a better response than permanent defection to a player who will respond to an initial cooperative move by playing delayed tit-for-tat minus $j - 1$. It also requires the stronger condition that delayed tit-for-tat minus $j - 1$ must be a better response than permanent defection to tit-for-tat minus $j - 2$. Let the payoffs in utility units for each player be c if they both cooperate, d if they both defect, 1 to a player who defects while his partner cooperates, and 0 to one who cooperates while his partner defects. For the classic repeated prisoner's dilemma game we require that $1 > c > d > 0$ and $c > \frac{1}{2}$. In order for tit-for-tat minus j to be rationalized, we need that $n > j + (c + d - 1)/(c - d)$. Thus, the higher the value of j, the longer the game must be in order for tit-for-tat minus j to be rationalized.

B believes that *A* is irrational. In the case in which *A* plays tit-for-tat minus 4, it breaks down at the fourth level: *A* believes that *B* believes that *A* believes that *B* believes that *A* is irrational. In general, the beliefs that make it rational for *A* to play tit-for-tat minus *j* involve a breakdown in the common belief in rationality at level *j*.

We have shown how it is that permanently defecting is not a uniquely rational strategy and, in particular, how it is that defecting in the first round is not uniquely rational. But that leaves the question of whether it could ever be rational to defect in round 1. Could it be that initial cooperation is uniquely rational?

It had better not be, for this reason. If initial cooperation were uniquely rational, then each would believe that the other would cooperate in round 1. But this ought to have the same effect as the later belief that he cooperates or cooperated in that round. It ought to mean that the common belief in rationality does not obtain. And that is inconsistent with the assumption that rationality is a matter of common belief at the beginning.

But, in any case, it ought to be clear that it is not uniquely rational to cooperate. Consistently with the common belief in rationality a player might believe that were he to cooperate initially then the other would believe that he was irrational and beyond the strategic influence of anything the other might do. In that case, he would expect the other to defect in response to an initial cooperation and so it would be rational for him to defect in round 1.

The import of this section then is twofold. It is not uniquely rational to defect permanently, in particular, to defect in the first round. And neither is it uniquely rational to cooperate in the first round. Both options are left open by the different belief sets that rational players might hold.

IV. THE PARADOX, WITH EVERY BELIEF COMMON

We have solved the paradox presented at the beginning of this paper and we have provided some useful support for the solution. But there are related paradoxes that may be thought to create equal problems. We consider a first here and a second in the next section.

One possible objection to the argument of section III is that we have allowed *A* to believe that he can outguess *B*. In the simplest case, *A* plays tit-for-tat minus 2, believing that *B* will play delayed tit-for-tat minus 1. But if *B* believed that this was what *A* was doing, *B* would not play delayed tit-for-tat minus 1. (He would do better to play delayed tit-for-tat minus 3.) If *A* believes *B* to be just as rational as he is, how can he expect to outguess him?

The idea that rational players cannot expect to outguess one another is often presented in game theory as the following requirement: any belief that one player holds, or at least any belief that he

holds about the other, must be a common belief.[8] To see the point of this, suppose that A holds the beliefs Z that make a particular strategy S rational for him. Then, if all beliefs are common, A must believe that B believes that A believes Z. Thus, if A's beliefs are such that S is rational for him, he must believe that B believes this to be the case too.

The idea that all beliefs must be common is thought to be a consequence of there being a common belief in rationality. Suppose A is rational and believes Z. Suppose also that B is rational and believes A to be rational. Then any reasoning that A can follow through can be replicated by B. In forming his beliefs about B, what has A to go on except his knowledge of the structure of the game and the common belief in rationality? And these are exactly the same data as B must use in forming his beliefs about A's beliefs about B. How, it is asked, can rational people, reasoning from exactly the same data, arrive at different conclusions? And hence the conclusion that B must believe that A believes Z; and so on: that A believes Z must be a common belief.

We do not think this requirement is plausible, because it seems to us that what one player should believe about another may be underdetermined by the data at his disposal. Suppose the data available to A are compatible with more than one system of beliefs, and that A happens to light on one of these systems. There seems to be no reason why B should guess that that is what he comes to believe, let alone that the belief should be common between them.

But let us put these misgivings aside and assume not just that the players have a common belief in their rationality but that any belief they have about one another is also common. May the backward induction argument be invoked under those assumptions? Game theorists have thought so. But the conclusion of the argument, that rational players ought never to cooperate, offends against intuition in this case too. And so we have another version of the backward induction paradox.

Does our solution of section II serve also to solve this paradox? The answer, happily, is that it does. Even under the strengthened assumptions the player who runs the backward induction has to believe that, given at the beginning a common belief in the rationality of the players, each will continue to maintain that belief, regardless of what the other does. But he cannot believe this, for the considerations invoked in section II apply here also. He cannot believe that the common belief in rationality will survive regardless of what hap-

[8] See, e.g., R. J. Aumann, "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica*, LV, 1 (1987):1–18; or Ken Binmore and Partha Dasgupta, *Economic Organizations as Games* (New York: Blackwell, 1985), pp. 1–6.

pens, because any act of cooperation would cause that belief to break down at some level. There is no paradox here because, as before, the argument that generates it is unsound.

It remains to provide support for that solution of the paradox by showing that the intuitive conclusion, that it can be rational to cooperate in the first round, is correct. To show this we must show that there are beliefs which a player might rationally hold which would make it rational for him to cooperate initially; and that these beliefs could be common in the sense required by the claim that a common belief in rationality entails the commonality of all beliefs.

This requires us to look more carefully at that claim. Suppose that A believes Z. The claim that therefore B must believe that A believes Z depends on the additional premise that B is rational, for if B were irrational there would be no grounds for expecting him to be capable of replicating whatever reasoning led A to his belief. The further claim that therefore A must believe that B believes that A believes Z depends on the further premises that A is rational and that A believes B to be rational; and so on: higher orders of belief in Z require higher orders of belief in the rationality of the players. The upshot of this is that, if the common belief in rationality breaks down at any level, so too does the argument for the commonality of all beliefs.

At the beginning of the sequence, when the common belief in rationality is in place, the argument for the commonality of all beliefs goes through, at least given our earlier concession. But since any act of cooperation will undermine the common belief in rationality at some level, we know that equally it must undermine the case for the commonality of whatever beliefs are held by the players.

With this point in mind, let us now ask whether it could be rational under the strengthened assumptions for either of the players to cooperate in the first round as part of one of the strategies discussed in the last section: tit-for-tat minus 2, tit-for-tat minus 4, and so on. If it could be rational to do this, then the support provided for the solution in the case of the original paradox serves also to support the solution here.

Consider the simplest of the patterns of beliefs we presented in section III—the pattern of beliefs which leads A to play tit-for-tat minus 2. A believes that B has the initial belief that A will defect in every round. A also believes, however, that, were he to cooperate in round 1, B would be led to believe that he was irrationally playing tit-for-tat and would therefore respond by playing delayed tit-for-tat minus 1. In the light of all these beliefs, A plays tit-for-tat minus 2. To this it might be objected that B, as a rational player, ought to be capable of replicating A's reasoning and hence of working out what A's beliefs are; and that, if he did this, he would realize that A was

playing tit-for-tat minus 2 and not tit-for-tat. Furthermore (the objection would go), *A* ought to be able to see that *B* would realize this. Is there therefore not an inconsistency in *A*'s beliefs? But by now it should be clear that the answer has to be "no." *A* believes that *B* believes he is irrational, so in *A*'s belief, *B* has no reason to replicate the reasoning through which he, *A*, has in fact gone. Thus, *A* may rationally hold the beliefs that would lead him to play tit-for-tat minus 2 even under the assumption that each player's initial beliefs about his partner must be matters of common belief.

It should be clear that this defense of the possible rationality of playing tit-for-tat minus 2 can be replicated in defense of the possible rationality of tit-for-tat minus 4, and so on. We may conclude then that, even if a common belief in rationality entails the commonality of all beliefs, still it may be rational for a player to cooperate initially.

This means that, contrary to the conclusion of the backward induction argument, the strategy of permanent defection, and in particular any strategy involving initial defection, is not uniquely rational. But, by the same argument as that presented in the last section, we can see that neither is initial cooperation uniquely rational: a player would be rational to defect initially if he believed, as he still might, that an initial cooperation would cause the other player to believe that he was irrational and beyond the reach of strategic considerations.

We must conclude, then, that the solution provided for our original paradox, and the support given to that solution, are both still relevant in the case of the strengthened paradox. The assumption that the common belief in rationality entails the commonality of all beliefs, even if it is conceded to be reasonable, does not significantly complicate things.[9]

V. THE PARADOX, WITH COMMON KNOWLEDGE OF RATIONALITY

Some game theorists postulate, not just that the players have a common belief in their rationality, but that this rationality is a matter of common knowledge: it is not something on which they could be mistaken, could think they were mistaken, could think they thought they were mistaken, and so on. What becomes of our paradox and our solution under this way of strengthening the assumptions?

It should be clear that, whatever is true about the paradox, the solution fails. Under the assumption of common knowledge of their rationality, each of the players is able rationally to believe that the common belief in rationality will endure; indeed, each knows that common knowledge will endure: otherwise, he could not know it to be knowledge. Thus, each is in a position where he can rationally

[9] Under this assumption, many game theorists will wish to argue that permanent defection is a sequential equilibrium. A critique of the significance of that argument is developed by Binmore in "Modeling Rational Players."

endorse arguments 1 . . . n of section II, or versions of these arguments with 'know' and 'knowledge' substituted for 'believe' and 'belief'. Each can run the backward induction and each will therefore defect.

What we wish to argue, however, is that, if the solution to our paradox fails under common knowledge of rationality, so does the paradox itself. The paradox arises from the conflict between the backward induction argument for permanent defection and the more intuitive thought that a player might do better to play a conditional strategy involving an initial act of cooperation. But this intuitive thought depends on a particular way of thinking about games which, although entirely natural, is not allowed by the assumption of common knowledge. The intuitive idea is that each player is free to choose any one of the strategies that he is allowed by the rules of the game. To decide which strategy to choose, the player asks himself what would happen in the event that he chose each strategy. If he is rational, he must then choose the strategy that leads to the best consequences; but it is possible for him to discover that this is the best strategy only because the hypothesis that he chooses a different (and, as it turns out, irrational) strategy is intelligible. Confronted with the backward induction argument for defection, the intuitive response is to think that, despite the apparent logic of that argument, the consequences following from a conditional strategy might be preferable to those following from permanent defection.

But if the players have common knowledge of their rationality, this intuitive response misses the point, for in a sense the backward induction argument does not show that a player does better by defecting: it shows that, as a matter of logical necessity, both players *must* defect and presumably therefore that they know they must defect. The reason is that, when a player A makes the intuitive response and asks himself how B is likely to respond in the event of A's initially cooperating, he is asking after what would happen under an inconsistent hypothesis. For it is inconsistent to postulate both that there is common knowledge of rationality, in which case there is common knowledge that both will defect in every round, and that one of the players cooperates in round 1.

If this analysis seems too quick, then a parallel may help to make the point vivid. I may believe that of two options before me, p and q, p promises the better outcome, and still ask what would happen, the world being as it is, were I to choose q instead; indeed, I am very likely to do this in order to check my belief that p does actually hold out the better prospect. I am able to ask this question, because there is nothing inconsistent in the hypothesis that, the world being as it is, I might believe that q was better and choose it rather than p. But suppose it is required now that knowledge states—indeed, states

involving knowing one knows, knowing one knows one knows, and so on—replace belief states. In that case, I cannot ask after what would happen, the world being as it is, were I to know that q is better; the world being as it is, i.e., p being actually better, it is not possible that I might know that q was better.

The situation where the players are ascribed common knowledge of their rationality ought strictly to have no interest for game theory. Under the assumption of common knowledge, neither player is allowed to think strategically. On pain of inconsistency, neither can ask after what effect he might have on the other through doing something besides what that assumption entails he should do: viz., permanently defect. We believe that the assumption of common knowledge of rationality, as it is invoked in game theory, should not be taken completely literally. It should probably be taken to mean something like what we have meant by 'common belief'.[10]

CONCLUSION

The backward induction paradox, as presented in sections I and IV, is solved by the fact that rational players are necessarily not in a position to run the backward induction argument. That is our principal claim. But two subsidary theses are also worth noting. The first is that, in the situations that give rise to the paradox, neither an initial cooperation nor an initial defection is uniquely rational: either may be rational in terms of beliefs that a player can rationally hold. The second is that there is no backward induction paradox under the one assumption we have considered which would license the use of the backward induction argument: the assumption of common knowledge of rationality. Understood literally, this assumption is inconsistent with what we take to be the whole point of game theory —the analysis of strategic thinking.[11]

PHILIP PETTIT

Australian National University

ROBERT SUGDEN

University of East Anglia

[10] Recent work in game theory has implicitly recognized this: see R. Selten, "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory,* IV, 1 (1975):25–55; and D. M. Kreps and R. Wilson, "Sequential Equilibrium," *Econometrica,* I., 4 (1982):863–894. Selten's concept of "perfect equilibrium" and Kreps and Wilson's concept of "sequential equilibrium" are both based on the idea that a player must initially assume that other players are rational, but that each player's strategy must include contingency plans to be brought into action if this initial assumption is disproved in the course of the game. The usual way of describing this approach is to say that it is common knowledge that the players are almost certain to be rational; this has much the same effect as making their rationality a matter of common belief.

[11] After this paper was in proof, an article by Cristina Bicchieri, in which she introduces an argument similar to that presented here, was brought to our attention [cf. "Self-refuting Theories of Strategic Interaction: A Paradox of Common Knowledge," *Erkenntnis* (forthcoming)].