BACKWARD-INDUCTION ARGUMENTS: A PARADOX REGAINED*

JORDAN HOWARD SOBEL†‡

Scarborough Campus
University of Toronto

According to a familiar argument, iterated prisoner's dilemmas of known finite lengths resolve for ideally rational and well-informed players: They would defect in the last round, anticipate this in the next to last round and so defect in it, *and so on*. But would they anticipate defections even if they had been cooperating? Not necessarily, say recent critics. These critics "lose" the backward-induction paradox by imposing indicative interpretations on rationality and information conditions. To regain it I propose subjunctive interpretations. To solve it I stress that implications for ordinary imperfect players are limited.

1. Introduction. I respond in this paper to recent criticisms of backward-induction reasoning, and especially to criticism conducted by Philip Pettit and Robert Sugden. The introduction to their paper provides points of reference for mine:

> 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.[1] 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 rec-

*Received November 1990; revised June 1992.

†I am grateful to Allan Gibbard, Philip Pettit, Wiodek Rabinowicz, Willa Freeman-Sobel, and Robert Sugden for comments, and to an anonymous referee for insightful criticism and suggestions.

‡Send reprint requests to the author, Scarborough Campus, University of Toronto, 1265 Military Trail, Scarborough, Ontario M1C 1A4, Canada.

[1] Brian Skyrms (1990, 130) presents such an argument. For similar arguments addressed to sequences of other games see Kreps and Wilson (1982, 255), and Milgrom and Roberts (1982, 283). "The paradoxical result[s] [have been said to be] due to [assumptions of] complete and perfect information" (Kreps and Wilson 1982, 276; see also Milgrom and Roberts 1982, 283).

Philosophy of Science, 60 (1993) pp. 114–133

> ommend 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. (Pettit and Sugden 1989, 169; emphasis added)

I spell out ungenerous indicative interpretations of emphasized premises in section 2, and present in section 3 an argument for permanent defection that is certainly not good when the premises are interpreted in those ways. Section 4 contains subjunctive interpretations that are adequate for such arguments, and, I contend, descriptive of perfectly rational players. Section 6 recycles the standard solution to the paradox thus regained. It reminds us that none of us are perfect, and that even if ideally rational and well-informed players would be doomed to permanent defection, rational and well-informed players can sometimes do better and cooperate at least in long sequences, and until late rounds. Section 5 makes an interlude in which I maintain that in place of common beliefs, game-theoretic models can feature common knowledge. "Of course they can", one might say, "for they almost always *do*". One might say this and no more but for recent challenges, such as Pettit and Sugden's, to common knowledge models.

Frank Jackson (1987) discusses surprise-examination paradoxes. The difference I work between indicative and subjunctive interpretations of rationality and knowledge assumptions is somewhat like a difference of importance to Jackson between belief-assumptions (BEC) and (BIC) for an "easy" paradox that he claims to solve, and certainty-assumptions (CEC) and (CIC) for a "hard" one. What is certain in the strong sense *would still be certain* whatever positively probable condition were to obtain as the examination week unfolded. The difficulty of the "hard" paradox which he does not solve lies, Jackson implies, at the heart of the surprise-examination conundrum (ibid., 122).

2. Indicative Rationality and Belief Premises. Players in sequences are to be rational. I take this to mean that each *maximizes causal expected utility* in each round. (Though I identify rational behavior with behavior that is maximizing *and* for which decisions would be stable on ideal reflection [Sobel 1990], the second condition coincides with the first in prisoner's dilemmas, and can be presently ignored.) Players are to "share the belief [subjective certainty] that they are rational" (Pettit and Sugden 1989, 169). This, we are told, is to be a *common* belief, something that each believes, that each believes that each believes, and so on without limit. It is stressed that the premise says only that these beliefs obtain "at the start of the game" (ibid., 172), but to give backward-induction reasoning a chance, and without prejudice to Pettit and Sugden's main objection, we want a stronger premise according to which common beliefs in players' rationality are "compounded forward to the end of the sequence" thus: For a sequence to come of length n, in each round k, $1 \leq k < n$, it will be a common belief that (for exposition k is taken to be less than $n - 4$)

> $(n - 1)$: It will be a common belief in round $n - 1$ that the players will be rational in round n;
>
> $(n - 2)$: It will be a common belief in round $n - 2$ that the players will be rational in rounds $n - 1$ through n, and that $(n - 1)$;
>
> $(n - 3)$: It will be a common belief in round $n - 3$ that the players will be rational in rounds $n - 2$ through n, and that $(n - 2)$ and $(n - 1)$, and so on; to
>
> $[n - (n - k) + 1] = (k + 1)$: It will be a common belief in round $k + 1$ that the players will be rational in rounds $k + 2$ through n, and that $(k + 2), \ldots, (n - 2)$, and $(n - 1)$.

This premise compounds through the sequence common beliefs that go one way, forward, and the compounding stops short of the last round. I cannot motivate, as required to give backward-induction reasoning a chance, common beliefs that go both ways or that are compounded through the last round, but without prejudice to Pettit and Sugden's objections, such beliefs can be assumed, and the belief premise to be impugned as inadequate can be:

Common Beliefs in Rationality Compounded
All Ways Through and Beyond the Sequence

> (1) It is always, before, during, and after the sequence, a common belief that players are rational in every round of the sequence.
>
> (2) It is always, before, during, and after the sequence, a common belief that (1).
>
> (3) It is always, before, during, and after the sequence, a common belief that (2). And so on without end.

If backward inductions are to have a chance, beliefs need to extend to things other than rationality. For example, premises need to insure that throughout sequences players have beliefs: (a) concerning the causal independence of actions in rounds; (b) concerning where they are in the sequence, and how many rounds are to come; (c) concerning possible "payoffs" in rounds; and (d) concerning how expected utilities for plays in sequences of rounds are related to payoffs for plays in rounds sequenced. For simplicity I suppose that these further matters are subjects not merely of common belief, but of common knowledge, compounded all ways through and beyond the sequence, and suppose that this knowledge not only obtains but would obtain throughout any possible play of a sequence. For concreteness, players can be in isolation booths with buttons they can press (defect) or not (cooperate), payoffs can be dollars,

	press	~press
press	\$1,\$1	\$3,\$0
~press	\$0,\$3	\$2,\$2

and utilities for sequences can be sums of payoffs.

3. A Backward Argument for Defection. 1. For rational players who believe they are rational it can *seem* that in each round k, $1 \leq k < n$, of a sequence of length n each player could reason in the following manner to the conclusion that his defection maximizes in that round (for exposition k is taken to be less than $n - 2$).

(i) We will defect in the last round, *and we would defect in that round whatever we had done in previous rounds.*

In the last round all that will matter to determinations of expected utilities of actions will be possible payoffs in it, and that is all that would matter no matter what we had done in previous rounds. As we will and would in any case realize in the last round, strategically the dilemma in it could as well be isolated. In it, as in every round, we will and would in any case see that defection dominates in terms of payoffs in it and that our actions are causally independent of one another, so that defection will, and would in any case, maximize expected utility. Finally, since we will, and would in any case, be rational in this round, we will, and would in any case, defect in it.

(ii) In the next to the last round we will, *and would whatever we had done in previous rounds,* believe that (i).

Why? Because we will, and would in any case, believe that defection will, and would in any case, maximize in the last round, and that we will, and would in any case, be rational in the last round.

(iii) We will, and would whatever we had done in previous rounds, defect in the next to last round.

Why? Because, by (ii), in the next to the last round we will, and would in any case, believe that our actions in the only future round will be, were going to be, causally independent not only of one another but of actions in the then *preceding* round (this current next to the last round); that is, we will, and would in any case, see in the next to the last round that strategically it could as well be last

> and itself isolated. (See, "Things are clear on the last trial . . .; hence the penultimate trial . . . is now in strategic reality the last . . ." [Luce and Raiffa 1957, 98]. "The last round might as well be isolated. . . . And this will be obvious in the . . . next to last round . . ." [Sobel 1985, 310].)

(iv) In the next to next to last round we will, and would whatever we had done in previous rounds, believe that (i) and (iii).

> Why? For the kinds of reasons detailed under (ii), but ramified so that they showed that we will and would in any case believe that we will, and would in any case in each coming round, (a) have beliefs and values that will make (would make) defection rational in it, and (b) be rational in it.

(v) We will, and would whatever we had done in previous rounds, defect in the next to next to last round.

> We will, and would in any case, see that actions in subsequent rounds are, were, going to be causally independent of actions in this next to next to last round so that strategically it could as well be the next to last round, or, for that matter, the very last round. We will, and would in any case, see that the dilemma in *it* could as well be isolated.

And so on to $(2(n - k))$. In round k we will, and would whatever we had done in previous rounds if any, believe that (i), (iii), . . ., and $(2(n - k) - 1)$ (i.e., that in each round to come we will, and would whatever we had done in previous rounds, defect).

Recasting the argument, we have, in strong mathematical-induction form (see Sorensen 1986, 342):

Basis. Ideally rational and well-informed players will, and would whatever they had done in previous rounds, defect in round n of a sequence of n prisoner's dilemmas. This is (i) above.

Inductive step. For every k such that $1 \leq k < n$, if ideally rational and well-informed players will, and would whatever they had done in previous rounds, defect in rounds $(k + 1)$ through n, then they will, and would whatever they had done in previous rounds, defect in round k.

For this we have the main lemma: *For every k such that $1 \leq k < n$, if ideally rational and well-informed players will, and would whatever they had done in previous rounds, defect in rounds $k + 1$ through n, then in round k they will, and would whatever they had done in previous rounds,* believe *that they will, and would whatever they had done in previous rounds, defect in rounds $k + 1$ through n*. Grounds for this are illustrated under (ii) and ramified under (iv).

Therefore, ideally rational and well-informed players will defect in every round.

2. It has been said that "the [backward-induction] argument for permanent defection is unsound" (Pettit and Sugden 1989, 169). I say that the argument just given certainly does overreach premises stressed when these are strictly and narrowly interpreted. According to the first stressed premise the players are rational. For a sequence of dilemmas that lies entirely in the future, this premise when narrowly interpreted says no more than that in each round each player *will* perform an action that, given his then current beliefs concerning possible consequences of his actions, maximizes expected utility. According to the second premise narrowly interpreted, the first premise *is* a matter of common belief compounded through the sequence. The problem with the argument from these *indicative* premises is that they do not *begin* to ground the several *subjunctive* moves it involves. Pettit and Sugden find in this problem "the solution to [the backward-induction] paradox" (ibid., 171), which they say parallels "the solution offered by Frank Jackson to a version of the surprise examination paradox [his 'easy' paradox]" (ibid.). To illustrate, the part emphasized in (ii) is not supported. The premises do entail that in rounds previous to the next to the last round all players *will* believe that in the next to the last round they *will* believe that we *will* defect in the last round. But these premises leave open that players do *not* believe that no matter what they had done in rounds previous to the next to the last round, they *would* believe in it that they were going to defect in the last round, let alone believe in it that they would defect in the last round no matter what they were to do in the next to the last round. The premises leave open that players believe that they will believe in the next to the last round that they will defect in the last round largely *because* they believe that they *will* believe in the next to the last round that they had defected in all rounds leading up to it, and that they were going to defect in it. Consistently with that opinion, however, they might believe that they are disposed to form expectations regarding each other by very simple induction, so that previous acts of cooperation would lead them in the next to the last round to expect actions in the last round to be further acts of cooperation. Also left open is the more important possibility that they believe that previous acts of cooperation could lead them to suspect that actions in the last round were not destined to be causally independent of actions in the next to the last round. For example, they could believe that some previous acts of cooperation would lead a player to suspect in the next to the last round that the other player had been, and would continue to be, disposed to reciprocate cooperation—that he had been, and would continue to be, following a tit-for-tat strategy. Given such suspi-

cions the penultimate trial would *not* in his view be in strategic reality the last trial.

The stressed premises, interpreted strictly and narrowly, that is, interpreted indicatively, fail to support (ii). Indeed, interpreted strictly and narrowly they do not support even (i). The first premise, indicatively construed, leaves open whether the players would be rational in a round no matter what actions they had taken in previous rounds. Left open, for example, is that some actions that were irrational, given then current beliefs concerning consequences, might, if performed, establish habits and dispositions not to maximize, or that they might affect others and make them soft and nonmaximizing in future rounds. Even for a length-2 sequence, the premise that players *will* be rational and maximize in each round does not entail that a player *would* be rational and defect in the last round no matter what players had done in the first round, and even if they had behaved irrationally and cooperated in the first round.

4. Subjunctive Rationality and Belief Premises. "There can be no doubt that the conclusion [defection in every round] follows from the stated premises, but the premises deserve scrutiny" (Skyrms 1990, 130).

1. The premises of section 2 are woefully inadequate to the argument of section 3. Matters could have been worse. These premises could have been not merely inadequate to the conclusion that players will permanently defect. They could have been obstacles to that conclusion, barriers to reasoning to permanent defection. They could even have implied that resourceful players would *not* permanently defect. But these indicative premises do not make such troubles, and to repair the argument it is possible simply to strengthen the constructions; it is possible to interpret Pettit and Sugden's words generously in ways that make them adequate to the subjunctive moves in the reasoning set out above, and thus better candidates for what intelligent proponents of resolutions by backward-induction reasoning can be supposed to have had more or less clearly in mind.

Pettit and Sugden disagree. They claim that their premises are barriers that get in the way of reasoning to permanent defection; they say that "the players are [not only] not necessarily in a position [but] indeed are necessarily not in position, to run the backward induction" (1989, 174). They suggest that players would need for such reasoning (a) "the belief that the common belief in rationality would survive even if cooperative moves were played" (ibid.), and they maintain that the premises they state imply that (b) "neither of the players *can* believe that the common belief in rationality will survive whatever moves the players make" (ibid.; emphasis added).

Now there is a sense in which, given certain assumptions that we should be willing to make, neither player can believe that a common belief in their rationality would survive no matter what; but this is not a sense in which it is necessary for backward-induction reasoning that players *should* believe that a common belief in rationality would survive no matter what. That is, while there is a sense of "common belief in rationality" in which (a) is needed, and a sense in which, given assumptions I am willing to make, (b) is true, these senses are different, and the sense in which (b) is true leaves open that players can believe that come what may they would have the common beliefs they need for backward-induction reasoning.

I concede that if, as we should, we take for granted that players would in every round remember what plays they had made in previous rounds, then a common belief in everyone's rationality in all rounds, *including rounds already completed*, could not survive the *last* round whatever moves the players had made in rounds including it. And if we take for granted also that the players would no matter what remember in every round what they had *believed* in all previous rounds, then a common belief in everyone's rationality in all rounds, including rounds already completed, could not survive even the *first* round whatever moves, holding beliefs constant, players had made in it. But this does not mean that our players cannot believe that a common belief in rationality sufficient for backward-induction reasoning would survive every round no matter what moves had been made in previous rounds, and no matter what moves were made in it, for that reasoning requires only that common beliefs, compounded in a certain manner, in players' rationality in all rounds if any still to *come* would survive no matter what moves the players had made in rounds completed.

2. Pettit and Sugden lose the backward-induction paradox by casting its premises as indicatives, and seek to bury it by insisting that for it common beliefs that go both ways would have to survive no matter what. To regain the paradox I frame, and briefly defend as appropriate, considerably stronger, largely *subjunctive*, rationality and belief premises that for the most part go only one way, forward.

According to our indicative rationality premise, each player is rational in the sense that in each round he will perform an action that, given his then current beliefs, would maximize expected utility. For a stronger premise, I let a player be, by definition, *resiliently rational in a round* if and only if he is rational in it, and, for every subsequent round, if any, he is rational in it, and *would* be rational in it no matter what players, himself included, had done in previous rounds; a resiliently rational player would, in any subsequent round, even if he (even if everyone) were irrational in the current round and every intervening round, finally come to his senses. Using this definition, I propose, as a stronger rationality premise for a sequence of known finite length of prisoner's dilemmas, *Resilient Rationality Through the Sequence*

> Each player is not merely rational, but resiliently rational, throughout the sequence, in the sense that, for each round, he will, and would no matter what players, himself included, had done in previous rounds, be resiliently rational in it.

This premise is in the idealizing spirit of classical game theory. While rationality in the thin sense of ever-maximizing can be superficial, and rationality in the less thin sense of a persistent underlying disposition to such actions can be eradicable, resilient rationality is deep-seated and includes not only a display of rational actions, but a deeply entrenched and ineradicable disposition to such actions that would assert itself no matter what insults it had suffered. To say that players are not merely, and perhaps only coincidentally, rational, but *resiliently* rational, is to say that they are very rational indeed. (A strong-willed and nonaddictive nonsmoker, not only does not smoke, but, even if he were to smoke once or twice, would not smoke anymore. If perfectly and completely resilient, then, for every n, he would not smoke anymore even if he were to smoke n times.)

3. According to the indicative belief premise, *common beliefs in rationality compounded all ways through and beyond the sequence*, the original thin rationality premise is a common belief compounded through and beyond the sequence in a certain manner. For a strengthened common-belief premise I conjoin the proposition that the new resilient rationality premise is a common belief compounded in a certain subjunctive manner forward to the end of the sequence:

Common Beliefs in Resilient Rationality,
Compounded Robustly Forward to the End of the Sequence

> For a sequence to come of length n, in each round k, $1 \leq k < n$, it will be, and would be "no matter what" (this from now on is short for "no matter what they had done in previous rounds, if any") a common belief that (for exposition k is taken to be less than $n - 4$)
>
> ($n - 1$): It will, and would no matter what, be a common belief in round $n - 1$ that the players are then resiliently rational;
>
> ($n - 2$): It will, and would no matter what, be a common belief in round $n - 2$ that the players are then resiliently rational, and that ($n - 1$);
>
> ($n - 3$): It will, and would no matter what, be a common belief in round $n - 3$ that the players are then resiliently rational, and that ($n - 2$) through ($n - 1$), and so on; finally to

> $(n - (n - k) + 1) = (k + 1)$: It will, and would no matter what, be a common belief in round $k + 1$ that the players are then resiliently rational, and that $(k + 2)$, . . ., $(n - 2)$, and $(n - 1)$.

This strengthened belief premise is suited to resiliently rational players. Such players would be rational in a round, and indeed resiliently rational in it, no matter what they had done in previous rounds. The new belief premise attributes confidence in appropriately robust immediate noninferential *appreciations* of themselves and their fellows in each round that would obtain regardless of what they had done in previous rounds. It elaborates on the idea that in any round they would not only be resiliently rational but would believe in their resilient rationality, and this notwithstanding contrary evidence provided by *patterned*, by even *constant*, past failings. The thought that underlies this premise is that ideal players would, no matter what, believe in a round that they would, no matter what, in every subsequent round see through any prima facie evidence provided by their behavior in previous rounds for their not being resiliently rational in it, and find it possible always in one way or another to discount such evidence. Ideal players, perfect players, would always, no matter what, believe that they would always, no matter what, know themselves and each other that well.

4. It has been said of extensive form games that "if the same player has to move at different points in the game, we want that player's knowledge to be the same at all of his information sets" (Bicchieri 1989, 336). I note *à propos* this that because it is not needed it is not part of my common-belief premise that a player's beliefs concerning the rationality of players and concerning their beliefs are to be the same at all information sets, from the first one on, that he actually reaches. More importantly, not only because it is not needed, but also because of threats of inconsistency were it included, it is most certainly not part of my common-belief premise that beliefs of a player should be the same at all information sets, including in particular ones that will not be reached. I want to leave open that information sets that will not be reached might be reached by irrational play in earlier rounds, as well as by rational play based on beliefs disproved in later rounds and jettisoned. If I were to require that at each possible information set my agents should commonly believe certain things about past rounds, I would have them commonly believe only what will be, or would be, *true* things about their beliefs and play in these rounds. It is not part of that premise that players, no matter what they had done or believed in previous rounds, would have common beliefs concerning their rationality and beliefs that went "in both directions" (ibid.), which common beliefs, depending on what had been done and believed in previous rounds, might be *false*. Connectedly, I

allow that acts of cooperation would provide evidence that *parts* "of the announced conditions of the game do not really hold" (Sorensen 1986, 345), and that "each of the players has [repeated] opportunit[ies] to undermine the other's knowledge of [parts] of the game situation" (ibid.). But I take care that no acts would provide evidence that undermined parts that are needed for backward reasoning to defection in every round.

I do not want or need that the beliefs of a player should be the same not only at all of the information sets he actually enters, but also at his only possible information sets. Nor do I understand why anyone would want that, or think that it was somehow unavoidable or traditional for the topic, or partly definitive of an extensive-form model worth investigating. This feature of Bicchieri's model seems to be important to her conclusion that assuming more levels of knowledge than is needed "leads to an inconsistency" (Bicchieri 1989, 336), that it leads "to a discrepancy between what is observed and what is known about the other player" (ibid. 1990, 77–78). That conclusion might be compared with one of Pettit and Sugden's according to which, "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" (1989, 174). They write of common beliefs that would go both ways, without noticing that backward reasoning to permanent defection does not *need* such extensive common beliefs.

On a similar note, Philip Reny writes that "there is . . . a large class of extensive form games, in which it is not possible for rationality to be common knowledge throughout the game" (1989, 363). It is consistent with my rationality and common-belief premises that the rationality of all players "throughout the game" should be a common belief, indeed a common *true* belief, "throughout the game", *if* "throughout the game" is taken in the sense of "at every information set that is actually reached". But this is not consistent with those premises if "throughout the game" is taken in the sense of "at absolutely every information set".

5. My resuscitation of backward-induction arguments does not assume that out-of-equilibrium acts of cooperation would invariably be viewed as having been "trembling hands", random and uncorrelated, accidental deviations that are for *this* reason not projected or taken as arguments against current rationality. It assumes that not even histories of constant past deviations and past irrationality would be projected by perfectly rational players, that not even unbroken histories of deviations that were viewed as not random and accidental, but as persistently confused, as unreflectively imitative, as corruptly principled, or as grounded sometimes in one of these ways and sometimes in another, would be projected. It assumes that all such projections would be contravened, because, even

after exceptionless runs of irrational actions had taken place, and regardless of what might have been their grounds, our agents would not only *in fact* come to their senses and take proper charge of their actions at last and evermore, but would immediately *see* that they had all come to their senses and, regardless of grounds for past deviations, would be taking proper charge of their actions at last and evermore. (What *might have been* the explanatory grounds for several acts of cooperation, whether scattered among acts of defection or presented in exceptionless runs, would be what *are* the grounds for these acts in "nearest" worlds in which they take place, the identities of which worlds would depend on details of particular stories for sequences of prisoners' dilemmas.) My players cannot build reputations and thereby "influence each other" (Sorensen 1986, 345–346). However this is not because they "play in total ignorance of each other's moves" (ibid.), but because whatever they knew that they had done, they would still at sight know each other for the resiliently rational players that, even if only at last, and for the first time, they in fact would be.

For theorists friendly to a game model that includes resilient rationality—rationality that always, no matter how long it had been dormant, would assert itself and be real in this relentless way—such robust and immediate recognition and confidence should seem a natural complement, and a necessary part of a balanced and complete model of perfection. It would be a strange model that supposed that always, no matter how long, and how near to completely, it had been dormant, rationality would assert itself firmly and forevermore, but that allowed that, though it would always *do* that, *that* it had done that would not always be appreciated or believed, at least not right away. The natural complementing assumption is that, on the contrary, not only would each perfectly rational player realize that *he* had emerged from his night of irrational abandonment, but each would see that he was not alone, and indeed that along with him all of his fellows were resiliently rational still, again, or at last.

Ken Binmore has written "that, in similar games after similar histories, players should normally be expected to choose similar . . . actions" (1987, 200). He writes, I assume, of *ordinary* players who come to know one another, and to learn what to expect of one another, largely from experience. In contrast, agents who are perfect maximizers according to the present account would always know one another immediately, on sight, and either without any experience of one another in past encounters, if they have only just met, or independently of any experience they may have had of one another, if they have a history of encounters.

Is this a wildly unrealistic hypothesis whose consequences must be completely irrelevant to ordinary players and real situations? Is there no continuity between the ideal condition here proposed and natural ones?

Consider animals in the wild, consider in particular young animals (I am thinking of the cub in the movie, *The Bear*). Do they at first not know what to make of the growls, barks, and licks, of the grimaces, smiles, and caresses of other animals, especially other animals of their own kinds? Do they come to know what to make of these only eventually and from experience? And are their readings of one another sensitive to every aberrant past performance, or are they sometimes resistant to change, sometimes even highly resistant, even in the face of counterevidence? I think that natural animals, babies in their mothers' arms, for example, are sometimes in their readings of other animals highly resistant to such changes, and disposed not to project "out of natural character" behavior. The robust and immediate confidence of my ideal agents in one another's resilient rationality can be viewed as a doxastic disposition whose natural resistance to change, which could be of survival-value, has been rendered extreme. That is appropriate as a feature of an idealization.

These new premises—the strengthened rationality premise that calls for resilient rationality, and the strengthened common-belief premise that calls for robust and immediate forward-looking self- and fellow-recognition compounded to the end of the sequence—are, I suggest, descriptive of ideally or *perfectly* rational and well-informed agents. And they evidently provide sufficient rationality and common beliefs in rationality for arguments of the kind explained in section 3 (1) to permanent defection in prisoner's dilemmas of known finite lengths.

For the record, I have elsewhere argued without dependence on backward reasoning for a more general conclusion. I have maintained, and still do, that "hyperrational maximizers" who have only forward-looking values would always "know each other too well to *teach* each other what to expect or to set for themselves effective *precedents*" (Sobel 1985, 311), and that, as a consequence, "[d]ilemmas in series, whether of definite [known finite] or *indefinite* length, will defeat such maximizers" (ibid., 314; emphasis added).

5. An Interlude in Defense of Common Knowledge. 1. It is evident that all references to beliefs in my common-belief premise are, given this premise and the new resilient rationality premise, references to what are, or would be under the conditions in which they would take place, *true* common beliefs. And so one supposes that these references could be to agents' common *knowledge* of the things of which my premise says they have common beliefs.

Pettit and Sugden think that there are special properties of and problems for common-knowledge idealizations. They suggest for one thing that given common knowledge, as distinct from mere belief, among players that they are rational, "each [*could*] run the backward induction and each will

. . . defect" (1989, 181). My difficulty with this is that an indicative common-knowledge premise analogous to the initial indicative common-belief premise would not put players in a position to run backward inductions. Perhaps Pettit and Sugden are influenced by the mistaken idea that common *knowledge* of rationality in all rounds would necessarily survive no matter what, as if knowledge were always "strong" in a manner analogous to Jackson's "strong certainty" (Jackson 1987, 122).

Pettit and Sugden also suggest (1989, 181–182) that since stipulations of a common-knowledge idealization would entail that the players will permanently defect, these stipulations cannot, on pain of incoherence, include stipulations concerning what would happen were the players to cooperate, or stipulations concerning what players are to think would happen were they to cooperate. The suggestion is that antecedents of conditionals that would say what would happen if players were to cooperate would be inconsistent, given stipulations that entailed that they would not cooperate. My difficulty here is that even if stipulations of common knowledge would be inconsistent with players cooperating, though that would make antecedents that had them cooperating *counterfactual*, it would not make them inconsistent, and so pointless and uninteresting. Regarding the pointlessness of inconsistent suppositions, I note that it is a valid principle of Lewis-Stalnaker conditional logic that, supposing an inconsistency, everything is the case: For any inconsistent or impossible p, and any proposition at all q, the conditional that $(p \mathrel{\Box\!\!\rightarrow} q)$ is true. "A more adequate [but more complicated] theory would allow truth-value gaps" (Sobel 1970, 432), and make these conditionals neither true nor false.

The general point is that an ideal case, and indeed any case or model, can be defined by subjunctive conditions along with indicative ones, and antecedents of subjunctive conditions can be incompatible with implications of conditions for the case. The only restriction on useful and interesting subjunctive conditions is that their antecedents be possible or entertainable for purposes of counterfactual supposition. Consider that we can of course think coherently about a case in which everyone knows that everyone is going to vote, and in which everyone also knows that even if someone, anyone, were not to vote, everyone else would. Similarly we can think about a case in which everyone knows that everyone is going to vote, and that if anyone were not to vote, then no one would.

To set conditions that logically imply that players *will* defect—for example, conditions that insure that they will as a consequence of backward-induction reasoning defect, or conditions that entail that they *know* (nevermind how) that they will defect (nevermind why)—is *not* to make it "a matter of logical necessity [that they] *must* defect" (Pettit and Sugden 1989, 181). If it were, then, given such conditions, to ask what would happen were a player not to defect *would* be equivalent to asking "what would happen under [the] inconsistent hypothesis" (ibid.) that he does not defect and does defect. It is a valid principle of subjunctive conditional logics, of Lewis-Stalnaker logics, that for any propositions p, q, and r:

$$\Box p \rightarrow \Box[(q \mathbin{\Box\!\!\rightarrow} r) \leftrightarrow ((p \,\&\, q) \mathbin{\Box\!\!\rightarrow} r)].$$

2. Perhaps Pettit and Sugden should be read as taking an "imported view" of common-belief conditions and proposing that it is uniquely appropriate to take an imported view of common-knowledge conditions in game models. This would, without making defection logically necessary, have some of the effects of doing that; it would have the effect of making suppositions of cooperation inconsistent.

Don Hubin and Glenn Ross contrast "imported views" of conditions with "exported views". To take an exported view of conditions of a practical puzzle or model is to hold "that [they] simply . . . [do], but need not, obtain" (Hubin and Ross 1985, 441) in the puzzle or model. "On the imported view . . . we consider as the outcome of an action what would happen were you to perform the action . . . when the puzzle conditions obtain" (ibid., 443). To take an imported view of a condition of a puzzle is to hold that this condition is necessary, not absolutely and logically, but conditionally on all actions in the puzzle. It is to suppose it holds "no matter what act is performed" (ibid., 445). For any action a, condition c, and state s, an imported view of c has the effect of importing c into the antecedents of conditionals that say what would be if a:

$$(a \mathbin{\Box\!\!\rightarrow} s) \leftrightarrow [(a \,\&\, c) \mathbin{\Box\!\!\rightarrow} s];$$

an imported view of c makes c independent of a:

$$c \,\&\, [\sim a \rightarrow (a \mathbin{\Box\!\!\rightarrow} c)].$$

It can be seen that when the distinction matters, when it makes a difference which view one takes of a condition, taking an imported view of a condition of a puzzle or model of someone else's making can, *pace* Hubin and Ross, be a plain mistake. (They hold that the imported view is right without exception for all puzzle conditions, "In solving a hypothetical practical problem, the stipulated puzzle conditions must be assumed to hold no matter what act is performed" [ibid.].) For example, while it is right to take an imported view of the independence conditions of standard Newcomb Problems (these conditions say that predictions made are causally independent of actions, and that contents of the boxes are causally independent of actions), taking an imported view of correctness conditions for predictions made would be a mistake. It would mistake the intents of makers of standard Newcomb Problems to suppose that the prediction that has been made of my action not only is correct, but would be correct no matter which of the things I can do I were to do. For it is part of all standard Newcomb Problems that whatever has been predicted both actions are possible, and that if I were to act in an unpredicted way the prediction made would of course not be correct. (It is part of "limit" problems that though the prediction is correct, and the predictor is infallible in the sense of never erring, the prediction is not necessarily correct, and he is not infallible in the mysterious sense of being incapable of error. See Sobel 1988.) Similarly, while imported views can be taken of common-belief conditions for standard game models, taking imported views of common-*true*-belief conditions and common-*knowledge* conditions, as perhaps Pettit and Sugden can be read as doing, would mistake intents of makers of standard game models. When it is supposed, for example, that it is common knowledge that players will behave rationally, it is taken for granted that they do not *have* to behave rationally, and that if they were not to behave rationally it would not, *per impossibile*, be known that they had behaved rationally.

3. Binmore strikes what seems a discordant note similar to Pettit and Sugden's, "Conventional arguments . . . require counterfactuals of the form 'If a rational player made the following sequence of *irrational* moves, then . . .' " (1987, 198; see also, it "seems inevitable to me [that] out-of-equilibrium behavior is to be treated in terms of mistakes" [ibid., 183]). He also suggests (without saying this straight out) that if players are perfectly rational in the ways required by these conventional arguments for equilibrium behavior (for example, permanent defection in sequences of prisoner's dilemmas), then *suppositions* made in these conventional arguments that would have players carrying out sequences of out-of-equilibrium irrational acts are not merely counterfactual but instances of "absurd speech" (ibid., 179). In fact, however, the supposition that a player who is rational makes an irrational move, while counterfactual, is not absurd or contradictory, and this easy point is one that Binmore himself seems at times to acknowledge. Rational players *do not* make irrational moves, and that is necessary, but this is not to say that rational players *cannot* make irrational moves, or that irrational moves are impossible for rational players.

4. There are well-known ambiguities to do with scopes of modal operators that may be relevant here. Thus,

"A rational player cannot do an irrational thing"

can have the true sense of, "it is logically necessary that if a player p is rational, then he does no act a that is irrational", which, roughly symbolized is,

$$\Box[(Rp \mathbin{\&} Ia) \rightarrow {\sim}Dpa)].$$

And it can (perversely) have the false sense, indeed the preposterous (and problematic, in that it would feature quantification into a modal context) sense of "If a player is rational, then it is logically necessary that he does nothing that is irrational", which, roughly symbolized is,

$$[(Rp \mathbin{\&} Ia) \rightarrow \Box{\sim}Dpa)].$$

If a player is rational, then it would be merely out of character for him to do an irrational thing, and "impossible" only in this very weak sense that is consistent with his doing six impossible things before breakfast. Compare the sense in which it is impossible for someone really honest to tell a lie.

Further to the possible relevance to our texts of well-known modal ambiguities, consider, "[If] rationality is a matter of common knowledge [for players,] it is not something on which they could be mistaken" (Pettit and Sugden 1989, 180). This could have the false and ridiculous sense of, "if players *p* know that they are rational *r*, then it is not logically possible for them to be mistaken about *r*", which, roughly symbolized, is

$$[Kpr \rightarrow {\sim}\Diamond Mpr]$$

or equivalently,

$$[Kpr \rightarrow \Box{\sim}Mpr].$$

And it could have the true, albeit completely unremarkable, sense of "it is logically necessary that if players *p* know that they are rational *r*, then they are not mistaken about their rationality *r*", which, roughly symbolized, is

$$\Box[Kpr \rightarrow {\sim}Mpr].$$

Possible confusions aside, we should say, though it will rarely need saying, that if a rational player were to make an irrational move, he would in making it not be rational. What else we should say would obtain if a rational player made certain irrational moves will depend on our "background theory" (Binmore 1987, 189). In the idealization I have framed, the background theory features resilient rationality, and spells out several other things we should say.

6. The "Paradox" and Its "Solution". "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" (Pettit and Sugden 1989, 171). The standard and I think correct solution to this "paradox" argues that while this predicament is logically inescapable for perfectly rational players, it is of only limited predictive and prescriptive relevance for real players, including very well-informed and reasonable ones. Perfect players would be trapped in permanent defections to their mutual disadvantage. The conclusion that they defect in every round is logically inescapable. But imperfect players, even very intelligent and well-informed maximizing players, can be well-advised on maximizing grounds to cooperate at least in early rounds. After all, real and thus imperfect maximizers can sometimes, even when they have well-based probabilities informed by appreciations of the subtleties of game-theoretic analyses, interact without error in nonequilibria even in "static" one-off games. (See Binmore 1987, 211, and Bernheim 1984, *passim.*)

What would be a player's maximizing play in a round of a given sequence of dilemmas depends on his probabilities and preferences; it depends especially on his probabilities for possible effects of his play on his opponent's plays in future rounds. These probabilities of real players will often have little to do with what they would expect, and realize they would expect, were they possessed of robust common beliefs compounded through the sequences in their resilient rationality. Real players, including very thoughtful ones, may with good reason think that they and their opponents are not resiliently rational, and that robust common beliefs that say they are are certainly not compounded through the sequence—it will often be obvious to real players that they are far from the perfect rationality and common-belief conditions of game-theoretic idealizations.

This familiar solution to the backward-induction paradox has advantages over those that would pick sides and either say that permanent defection is the rational strategy for all rational and well-informed players, or say that not even *ideally* rational and well-informed players would necessarily defect in every round. The first would-be solution is certainly wrong; it ignores the dependence of players' rational strategies on their subjective probabilities. And the second would-be solution at least runs into difficulty given how widely and implausibly it would cast logical aspersions. To be interesting it would need to say that there is "nothing in" the continuing tradition of idealizations that purport to support backward-induction resolutions, not even when these idealizations are interpreted generously.

Rather than taking sides I have tried to give each its due by explaining conditions of rationality and common beliefs that would put players that satisfied them in positions to reason backward to defection, and by stressing that this result does not have direct and unmoderated implications, predictive or prescriptive, for real players. However, though the main result of this paper, namely, that perfect players would defect in sequences of dilemmas, lacks direct and unmoderated implications for real players, it is not I think completely irrelevant to predictions and prescriptions for them. Coupled with the plausible idea that real players in *short* sequences are apt to be nearly ideal in the sense of my conditions, this result predicts that even when cooperation is established in a sequence it is apt to be unstable as the sequence winds down (see Luce and Raiffa 1957, 98). If, as I think, real players are sometimes nearly ideal maximizers in short sequences, the main result of this paper, together with details of its grounds, are things that real players might usefully take into account in their deliberations when beginning short sequences and when in sequences that have run to points at which only short segments remain.[2]

Finally, and on a very different note, details of my recovery project, in particular, of my subjunctively strengthened rationality and belief premises, may be suggestive of exact points of departure for studies of species of bounded, less than perfect, rationality in finitely repeated prisoner's dilemmas.[3] I think that in the case of the backward-induction paradox, as in the case of Surprise Examination Paradox (see Montague and Kaplan 1974, 271), interests of theory are best served not by finding ways to deny it, but by articulating conditions that make it genuine.

REFERENCES

Bernheim, B. D. (1984), "Rationalizable Strategic Behavior", *Econometrica 52*: 1007–1028.

Bicchieri, C. (1989), "Backward Induction without Common Knowledge", *PSA 1988*, vol. 2. East Lansing: Philosophy of Science Association, pp. 329–343.

———. (1990), "Paradoxes of Rationality", *Midwest Studies in Philosophy*. Vol. 15, *The Philosophy of the Human Sciences*. Notre Dame: University of Notre Dame Press, pp. 65–79.

Binmore, K. (1987), "Modeling Rational Players: Part 1", *Economics and Philosophy 3*: 179–214.

Hubin, D. and Ross, G. (1985), "Newcomb's Perfect Predictor", *Noûs 19*: 439–446.

Jackson, F. (1987), *Conditionals*. Oxford: Blackwell.

[2] Challenging the backward-induction argument, which they observe "holds [for what it is worth] however large *n* may be", Pettit and Sugden (1989, 169) observe that "if *n* is a large number, it appears that I might do better to follow a strategy such as tit-for-tat . . ." (ibid.). These words suggest that they may think that defection can be reasonable when *n* is small. How, if they believe in it, could they explain that size sensitivity? They could not say that players in short sequences often approach conditions of perfect rationality and information that support backward inductions to permanent defection, for they think that conditions of perfect rationality and information do not support backward inductions, and their strictures have nothing to do with the size of *n*.

[3] "The issue . . . is whether this puzzle [of defection round after round in finitely repeated prisoner's dilemmas] can be resolved in the context of rational, self-interested behavior. The approach we adopt is to admit a 'small amount' of the 'right kind' of incomplete information" (Kreps et al. 1982, 246). Two illustrations are given. In one, players begin with small probabilities for their opponents' not simply maximizing (which is what they in fact will do) but playing tit-for-tat (ibid., 247). In the other they have small probabilities for their opponents' having cooperative preferences (which they in fact do not have), for their enjoying "cooperation when it is met by cooperation" (ibid., 251). These small initial probabilities are shown to be "right kinds" of "incomplete information" (read as "misinformation"). Details of my hyperrational model could suggest other hypotheses.

Kreps, D. M.; Milgrom, P.; Roberts, J.; and Wilson, R. (1982), "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma", *Journal of Economic Theory 27*: 245–252.

Kreps, D. M. and Wilson, R. (1982), "Reputation and Imperfect Information", *Journal of Economic Theory 27*: 253–279.

Luce, R. D. and Raiffa, H. (1957), *Games and Decisions: Introduction and Critical Survey*. New York: Wiley.

Milgrom, P. and Roberts, J. (1982), "Predation, Reputation, and Entry Deterrence", *Journal of Economic Theory 27*: 280–312.

Montague, R. and Kaplan, D. (1974), "A Paradox Regained". In R. H. Thomason (ed.), *Formal Philosophy: Selected Papers of Richard Montague*. New Haven: Yale University Press, pp. 271–285.

Pettit, P. and Sugden, R. (1989), "The Backward Induction Paradox", *The Journal of Philosophy 86*: 169–182.

Reny, P. J. (1989), "Common Knowledge and Games with Perfect Information", *PSA 1988*, vol. 2. East Lansing: Philosophy of Science Association, pp. 363–369.

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

Sobel, J. H. (1970), "Utilitarianism: Simple and General", *Inquiry 13*: 394–449.

———. (1985), "Utility Maximizers in Iterated Prisoner's Dilemmas". Reprinted with corrections in R. Campbell and L. Sowden (eds.), *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*. Vancouver: The University of British Columbia Press, pp. 306–319.

———. (1988), "Infallible Predictors", *Philosophical Review 97*: 3–24.

———. (1990), "Maximization, Stability of Decision, and Rational Actions", *Philosophy of Science 57*: 60–77.

Sorensen, R. (1986), "Blindspotting and Choice Variations of the Prediction Paradox", *American Philosophical Quarterly 23*: 337–352.