Accounting for the "Tragedy" in the Prisoner's Dilemma*

JOHN J. TILLEY
INDIANA UNIVERSITY–PURDUE UNIVERSITY INDIANAPOLIS

[Final draft of a paper that appeared in *Synthese* 99(2) (1994): 251–76.]

ABSTRACT. The Prisoner's Dilemma (PD) exhibits a "tragedy" in this sense: if the players are fully informed and rational they are condemned to a jointly dispreferred outcome. In this essay I address the following question: What feature of the PD's payoff structure is necessary and sufficient to produce the tragedy? In answering it I use the notion of a "trembling-hand equilibrium". In the final section I discuss an implication of my argument, an implication that bears on the persistence of the problem posed by the PD.

§1

A well-known fact about the Prisoner's Dilemma (PD) is that it reveals something tragic or paradoxical about rational interaction, at least if we hold a standard "utility-maximizing" view of rational choice.[1] Whether it reveals a genuine *paradox* is questionable; hence I will stick with the word "tragedy" throughout this essay. The tragedy is that if the players are rational and fully informed they are sure to achieve[2] a jointly dispreferred outcome (namely *DD*).

1:

		L	
		C	D
H	C	2 2	3 0
	D	0 3	1 1

[*] I'm grateful to J. Howard Sobel for helpful remarks on an earlier draft. I also thank Subir Chakrabarti for many valuable discussions, and for my initial acquaintance with the literature on equilibrium refinement.

[1] The PD is defined by figure 1 together with some assumptions. (The latter are not always stated; they are implicit in the anecdotes used to introduce matrix 1.) The assumptions, which I'll make about all the games I discuss or allude to, are these: first, the game is non-repeated; second, the players are confined to pure strategies; third, each player's choice is independent of the choice of his co-player; fourth, communication and binding agreements are impossible; and finally, the players choose simultaneously, or at least without direct knowledge of the other's choice. In the outcome cells of each matrix, player *L*'s utility payoff is shown to the upper right of *H*'s, and all payoffs are assumed to have interval (not merely ordinal) significance. I use standard notation when referring to the outcomes; so "outcome *CD*" refers to the outcome produced if *H* chooses *C* and *L* chooses *D*.

[2] When I say anything to the effect that the players will "achieve" or "attain" a particular cell (e.g., that their choices will "produce" or "result in" that cell), or that a particular strategy set will be "the outcome" of a game, I mean that the players will *choose* the relevant strategies, which leaves open the question whether they will *succeed* in producing the relevant outcome. And when I say that an outcome is "achievable" or "attainable", I mean that it's available (allowed by the rules of the game) and at the intersection of rational choices. Again, this leaves open the question whether the players will *successfully* bring about the outcome. These qualifications are important because later on I will assume that rational players can err when trying to play their chosen strategies, and hence fail to achieve (in the ordinary sense of that word) the cell at the intersection of those strategies. In short, there is always a slight chance that the "achieved" outcome will not actually come about. For simplicity I will ignore this point until section 5. Until then, I will speak as though a *choice* of a given strategy can be equated with a *successful play* of that strategy.

But what accounts for this tragedy? That is, what property of the PD's payoff structure is responsible for the tragedy's occurrence? In what follows I will propose an answer to this question, and argue that the standard answers to it – the answers that routinely appear in the literature on the PD – are false, at least if they are taken to be the *kind* of answer I'm seeking: one that states necessary and sufficient conditions for the tragedy. I say that I will "propose" an answer because to address the question I must enter a debate that cannot be settled in the space of this paper: the debate over the type of outcome that provides solutions to non-cooperative, normal form games. But this only means that my answer to the above question must remain tentative, not that it fails to improve on the most common ones. In defending it I will use a concept that has been much discussed by game theorists, but which, as far as I know, has drawn little attention from philosophers, and has not appeared in the literature on the PD: the notion of a "trembling-hand equilibrium".

§2

A preliminary question, however, is whether the "players" I discuss should be seen as people of flesh and blood, or instead as theoretical abstractions. To put this another way, should we have in mind beings who are *human* as well as rational and fully informed, or theoretical constructs that have all and only those features entailed by our description of them as informed and rational? If we take the second option, we can assume nothing about the players that is not already implicit in our description of them. In particular, we can assume nothing about them from our observations of actual people. If, however, we take the first option, we can assume not only that the players are informed and rational, but that they have various *other* characteristics: those that everyday experience shows people to have. For instance, they feel pleasure and pain, they possess ordinary human knowledge, and so on.

In what follows I will take the first option, and regard the players as human. This is because there is no reason to explicitly assume one thing while implicitly assuming another, and I think we implicitly regard the players in the PD as people of flesh and blood, not as theoretical abstractions. This is evident from our view that there is indeed something *tragic* about that game – that we are warranted in being disturbed by its deficient solution.[3] There is nothing

[3] The word "deficient" is one of several game theoretical terms I will use. For convenience I will define most of them in this note. To begin, "deficient" means "strictly deficient" unless otherwise specified. An outcome *M* is strictly deficient if and only if the game contains at least one other outcome *N* such that every player prefers *N* to *M*. *M* is then *strictly inferior* to *N*. Strict deficiency differs from *Pareto*-deficiency. An outcome *M* is Pareto-deficient if and only if at least one other outcome *N* exists that would increase at least one player's payoff without lowering the payoff of any other player. *M* is then *Pareto inferior* to *N*. An outcome that is not Pareto-deficient is *Pareto-optimal*, or *optimal* for short.

Next, a strategy *S dominates* another strategy *T* if and only if: (i) for *any* strategies played by the other agents in the game, *S* yields a payoff (to the agent playing it) at least as great as that yielded by *T*; and (ii) for *some* strategies played by the other agents, *S* yields a payoff *greater* than that yielded by *T*. A strategy *strongly dominates* another if and only if it yields a greater payoff than the other no matter what strategies disturbing about a pair of *theoretical abstractions* obtaining a deficient result, for they are not creatures that can be harmed in any way (in fact, they are not creatures at all, nor can they genuinely *obtain* anything). The PD is troubling because we take it to show that rational informed *people* can sometimes be condemned to a deficient outcome when they interact. We do this, I think, because we tacitly regard the players not as theoretical abstractions, but as people, different from the rest of us only in being ideally rational and informed.[4] If so, we should make this assumption explicit, for it may have implications for our analysis of the game. We should also assume common knowledge that each player is human, otherwise we have no grounds for considering the players *completely* informed about their situation.

From here on, I will use an enriched definition of a "rational, fully informed player", and will use "rational player" as an abbreviation for that term. Such a player has these traits:

(a) He is an individual utility-maximizer who reasons flawlessly. Any choice he makes is one that he judges to be utility-maximizing relative to the decisions he expects from his co-players, and his judgments and expectations are based on perfect reasoning from all relevant available facts. (The phrase "based on perfect reasoning" is meant to indicate, among other things, that the player's judgments and expectations are not conjectures – they are conclusions supported by the balance of reasons.) He will choose one of his available strategies if he has a choice that meets these conditions.

(b) Anything he knows about the game or about his co-players, and any fact from which he reasons to a thought or intention, is an item of common knowledge.

(c) He is informed of every detail of the game (the rules, payoffs, etc.), including the traits that define his co-players.

are played by the other agents. A strategy is *dominant* for a player if and only if it dominates all other strategies available to that player, and *strongly dominant* for the player if and only if it strongly dominates all his other strategies.

The next term is "equilibrium". An outcome is in equilibrium if and only if no player could gain a higher payoff by unilaterally departing from it. If such departure would result in a lower payoff to any player who effected it, the outcome is in *strong* equilibrium. It is in *weak* equilibrium if at least one player has an alternative strategy that would result in an equal payoff rather than a lower one.

The final three terms can be sufficiently clarified in three sentences. First, to take a *maximin* approach to a decision problem, one finds the minimum payoff for each option, and then picks the option with the highest of those payoffs. Second, in a game of *complete information* the players have properties (b)–(d), below. Third, to say that a fact is "common knowledge" is to say that it is known by each player, and known by each player to be known by each player, and so on.

[4] Or else we regard the players as theoretical abstractions, but then assume that because *they* would achieve a deficient outcome, rational *people* would do so as well. In either case, our belief that the PD exhibits a tragedy rests on the assumption that rational informed people would attain a deficient cell. We should make that assumption explicit from the start, by regarding the players as human.

(d) Owing to (a)–(c), he can duplicate the reasoning underlying any decision made by another rational player in the game.

(e) He is *human*, meaning that he has the characteristics we can reasonably expect an ordinary person to have, except those ruled out by (a)–(d).

Properties (a)–(d) are standard in game theory, and together ensure that if every player makes a choice, each player's choice is utility-maximizing relative to the others.[5] Property (e) is not standard, although it (or something like it) is implicit in many arguments.[6] Taking it seriously will influence my answer to the question in section 1.

§3

The question was this: What property of the PD's payoff structure accounts for the tragedy in that game? The standard answers are these:[7]

(A) The players have strongly dominant strategies intersecting in a strictly deficient cell.[8]

[5] Imagine a two-person game in which the players choose their respective components of strategy set (S_H, S_L), but at least one of those choices, say S_H, is not utility-maximizing relative to the other. This implies that: (i) *H*'s choice is not one that she judges to be utility-maximizing relative to the decision she expects from *L*; (ii) either her conclusion about what counts as a utility-maximizing choice, or her prediction about *L*, results from faulty reasoning from the available facts; (iii) the available facts do not enable her (even if she is reasoning perfectly) to form accurate expectations about *L*; or (iv) the available facts do not enable her to form a correct view about what counts as a utility-maximizing choice. But (i) and (ii) are ruled out by feature (a), and (iii) is ruled out because feature (d) ensures that *H* can predict *L*'s decision. The latter prediction, when combined with *H*'s reasoning capacities and her detailed knowledge of the game, enables her to form a correct view about what choices qualify as utility-maximizing. This rules out (iv). So if *H* and *L* are rational, and each chooses a strategy, their choices will be mutually utility-maximizing.

[6] Something akin to it is at work in Lewis 1969, pp. 35ff, and in Luce and Raiffa 1957, pp. 109f.

[7] It's not hard to show that these are the standard answers. What's doubtful (in many cases plainly false) is that those who advance them mean to be giving conditions that are sufficient and *necessary* for the tragedy. Thus, my criticism of the answers is not meant as criticism of the authors who provide them. (Among the latter are Hamburger 1979, p. 78; Luce and Raiffa 1957, p. 96; Resnik 1987, p. 148; and Zagare 1984, pp. 52f.) I criticize the answers because although they do not suffice as the kind of account I'm seeking, they are likely to suggest themselves for that purpose given their prevalence in the literature.

[8] For clarification of "strongly dominant" and "strictly deficient" see note 3. Any reference, implicit or explicit, to an outcome or strategy should be understood as a reference to an *available* outcome or strategy (one allowed by the rules of the game) unless otherwise indicated. Thus, when I say that the players have strongly dominant strategies intersecting in a strictly deficient cell, I mean they have such strategies *available*, intersecting in an *available* cell that is strictly inferior to another *available* outcome. Unless this is kept in mind, some statements will be puzzling. For example, the first part of answer (H) will seem to suggest, falsely, that some normal form games have no equilibria. But if we read "equilibria" to mean "available equilibria", and remember that in this essay the players use only pure strategies, (H) carries no false suggestions.

(B) The players have dominant strategies (not necessarily *strongly* dominant strategies) intersecting in a strictly deficient cell.

(C) The game has a single equilibrium, which is strictly deficient.

Some additional answers are these:

(D) The players' maximin strategies intersect in a strictly deficient outcome.

(E) The game has a single equilibrium, which is both a *strong* equilibrium and strictly deficient.

(F) The game contains at least one strong equilibrium, but every strong equilibrium it contains is strictly deficient.

(G) The game has a non-equilibrium that is jointly preferred to any outcome that might result if both players deviated from the component strategies of that non-equilibrium. Also, any outcome that might result from this deviation is in equilibrium.[9]

(H) The game contains one or more equilibria, all of which are strictly deficient.

The PD (figure 1) has all the above properties. It has a single, strong equilibrium, *DD*, which is both at the intersection of strongly dominant (and hence maximin) strategies, and strictly inferior to *CC*, a non-equilibrium. But the above properties fail to account for the tragedy in the PD, if by an "account" of the tragedy we mean a feature of the game that is necessary and sufficient for the tragedy to occur.

Answers (A), (B) and (D) are ruled out by game 2. In this game, only one player has a dominant strategy, and the maximin strategies lead to a non-deficient cell (*CD*). Yet the game presents the same tragedy exhibited by game 1. Player *H* will see that *L* must choose *D*, and *H* can respond rationally to that choice only if he too chooses *D*. So the ultimate outcome will be *DD*, which is jointly dispreferred to *CC*.

2:		L	
		C	D
H	C	2 3	3 1
	D	0 0	1 2

[9] Perhaps the wording of this answer is puzzling. It's formulated so that it can be applied not just to 2x2 games, but to two-person games of a larger size.

This does not mean that we can find no important differences between games 1 and 2. One such difference is that in game 1, rational choices will produce *DD* even if each player knows only her *own* payoffs.[10] In game 2 this is not necessarily true. Suppose that *H* is ignorant of *L*'s payoffs, and thus cannot predict his choice. If we extend our definition of "rational players" so that it applies to such situations, we may have to conclude that in game 2 player *H* will choose *C*, since that is her maximin, her maximax, and her minimax-regret option.[11] We will not conclude this about game 1 (unless we extend our definition in a wildly implausible way), since *D* is strongly dominant for each player.

The upshot is that in game 1, but not in game 2, the cell produced by rational choices is sure to be deficient even if each player is ignorant of the other's payoffs. If we take *this* to be the tragedy in the PD, game 2 does not share the tragedy. But this is not the tragedy that makes the PD famous, the tragedy many have dubbed a "paradox". There is nothing paradoxical about two players attaining a deficient cell if at least one player is so poorly informed that he cannot predict the other's choice. We get a paradox (arguably, anyway) only if we assume that the players in game 1 are fully informed about the game. But if we assume this about the players in game 2, we must grant that game 2 displays the same paradox – the same *tragedy*, as I prefer to call it – displayed by game 1.

Thus, we have two games that are equally tragic, although they differ in important respects (for example, one game is a PD, the other is not).[12] One thing they share, however, is the presence of a single, strong equilibrium, which is strictly inferior to another cell. So perhaps (C) or (E) accounts for the tragedy.

Those answers, as well as (F), are ruled out by figure 3. This game has two equilibria, *DD* and *DE*, neither of which is in strong equilibrium. Rational players will attain one of those cells (note that *D* is strongly dominant for *H*), both of which are strictly inferior to *CC*. So the following facts – that the PD has a *single* equilibrium, and that it has a *strong* equilibrium – are not essential to the tragedy it displays. A game with multiple equilibria, all of them weak, can exhibit the same tragedy.

[10] Another is that in game 2, but not in game 1, if the players choose in sequence with *perfect* information (meaning that each is aware of all prior choices when it's his turn to choose), the outcome will depend on who chooses first. We are assuming, however, that the players choose simultaneously (see note 1).

[11] I have clarified the first of these terms in note 3; for the others see Resnik 1987, ch. 2, or Luce and Raiffa 1957, ch. 13. I say that we *may* have to conclude that *H* will play *C*, not that we *must* do so, to allow for the view that decisions under ignorance should be treated as decisions under risk by assigning a subjective probability to each "state of nature" – in this case each of *L*'s strategies – and then maximizing expected utility. If *H* assigns a probability greater than .75 to *L*'s *D*-strategy, *H* must choose *D* to maximize expected utility.

[12] As the parenthetical remark indicates, the tragedy we are discussing is not sufficient to make a game a PD. The question, "What features of a game are sufficient to make it a PD?" is different from the one I'm addressing.

3:

			L	
		C	D	E
H	C	2 3	2 1	2 1
	D	0 4	1 2	1 2

Game 3 raises some points worth highlighting. As already stated, rational players will achieve *DD* or *DE*. But we cannot predict the exact cell they will attain, because *L* will express his indifference between *D* and *E* by choosing arbitrarily between the two. Some might be troubled by this, and require rational players to express indifference by choosing a *mixed* strategy – in *L*'s case, a .50/.50 mix of *D* and *E*. But in the games we are discussing, only pure strategies are available; so this requirement would lead to the implausible view that game 3 is unsolvable. Thus, we should allow indifference to be expressed through an arbitrary choice.[13] One result is that some solvable games have multiple solutions. A solvable game is simply a game with at least one solution, and a solution is a set of choices, one *per* player, which is consistent with the assumption that each player is rational. (In short, the outcome at the intersection of those choices is achievable).

Let us now consider two properties referred to in answer (G), and shared by all the games discussed so far. First, a non-equilibrium exists (namely *CC*) that both players prefer to any cell they might attain if they deviated from their component strategies of that non-equilibrium. In game 3 this deviation would produce *DD* or *DE*. In the other games it would produce *DD*. Second, the outcomes just mentioned are equilibria. So perhaps (G) accounts for the tragedy we have been discussing.

4:

			L	
		C	D	E
H	C	2 2	1 0	3 0
	D	0 3	0 1	1 1

[13] This constrains the way we interpret "decision" and "utility-maximizing relative to the decisions he expects from his co-players", as they appear in our definition of rational players (in section 2). For the sake of space I will state those constraints without fully explaining them. First, we must interpret "decision" so that it means either a choice of a specific strategy *or* a practical conclusion of the sort *L* makes in game 3 – a conclusion of the form, "I must choose *D* or *E* (or . . .)", which is then followed by an arbitrary choice of one of those strategies. (This point applies specifically to "decision", not to "choice" or "intention".) Second, if *L*'s "decision" is a conclusion of the latter kind, then *H*'s choice is utility-maximizing relative to *L*'s decision if and only if it is utility-maximizing relative to each of *L*'s feasible choices (*D* and *E* in this case). If we keep these two points in mind, the argument in note 5 remains sound, and our definition of rational players (specifically item (a)) has no false presuppositions.

But (G) is ruled out by game 4. In this game, a joint departure from *CC* would not necessarily produce an equilibrium, for it might result in *DD*. Of course, a joint *rational* departure from *CC* would produce an equilibrium, for it would produce *DE*, which is at the intersection of strongly dominant strategies. But if we revise (G) so that it speaks of a "rational" departure from *CC*, our account of the tragedy in the PD will be unsatisfactory, or at least unsatisfying.[14] The revised version of (G) will merely restate, with a few embellishments, the tragedy it is designed to explain.[15] To say that a game "has a non-equilibrium that is jointly preferred to any outcome that might result if both players deviated, in a rational way, from the component strategies of that non-equilibrium" is not much different from saying that if both players are rational they will attain a jointly dispreferred outcome. So let us put (G) aside and consider (H).[16]

§4

Answer (H) is vulnerable to counterexamples unless we revise it as follows: (H´) The game is *solvable*, and contains one or more equilibria, all of which are strictly deficient. To see the need for the revision consider matrix 5, in which *DD* and *EE* are equilibria. Answer (H) is true of this matrix, but the game does not exhibit the tragedy we found in the PD. In game 5, rational players cannot choose strategies, from which it trivially follows that they will not choose strategies that lead to a jointly dispreferred cell. They cannot choose strategies because neither player can make a choice that he judges to be utility-maximizing relative to what he expects from his co-player. To make such a judgment, he must form a determinate expectation about the strategy his co-player will choose, but the structure of the game precludes this.[17] However, the example

[14] An equally poor way of revising (G) is to delete its second sentence. This makes game 4 ineffective as a counterexample to (G), but at the cost of making (G) true of many non-tragic games – e.g., the game produced if we change the first matrix in note 16 so that *H*'s payoff in cell *CD* is 3.

[15] Similar things can be said about other possible accounts of the tragedy – for example, one that speaks of the "natural outcomes" in games 1–4. The game theorists who coined that term (Rapoport, *et. al.*, 1976, p. 17) explicated it in terms of the choices of *rational* players.

[16] Before doing so, let us consider some further counterexamples to (A)–(G). Games (i) and (ii) each rule out (A) and (B), and game (iii) rules out (A), (B), (D), (E) and (F). Games 6–9, in section 5, are counterexamples to (C) and (E). One of them – game 9 – also rules out (G), and the other three rule out (F).

(i)		L	
		C	D
H	C	2 3	3 0
	D	0 1	1 2

(ii)		L	
		C	D
H	C	2 3	3 0
	D	0 2	1 1

(iii)		L	
		C	D
H	C	1 3	2 1
	D	0 0	0 2

[17] Some might object by saying that game 5, being a standard *coordination problem*, is solvable if one of the equilibria is perceptually salient (see Elster 1986, p. 9; and Ullmann-Margalit 1977, pp. 83, 112). To sidestep this objection we need only stipulate that neither equilibrium is salient. But we can also meet the objection head-on by borrowing an argument from Margaret Gilbert (1989). Her argument shows that even if one of the two equilibria, say *EE*, is salient, the *balance of reasons* does not dictate, for either

makes no trouble for (H´). Since rational players cannot choose strategies, the game is unsolvable.

5:

			L	
		C	D	E
H	C	2 2	3 0	3 0
	D	0 3	1 1	0 0
	E	0 3	0 0	1 1

We now have, in (H´), a tempting answer to our question about the tragedy in the PD. For one thing, the two-fold property referred to in (H´) is shared by all the games in the previous sections. More importantly, (H´) seems to isolate the precise class of games that exhibit the tragedy we are trying to explain.

To see this, note first that when rational players confront a solvable game, their choices intersect in an equilibrium. This is guaranteed by two facts, one about rational players, the other about equilibria. First, rational players choose strategies in such a game, and their choices are mutually utility-maximizing. Second, an equilibrium is a set of mutually utility-maximizing strategies. So if the players are rational, they are sure to achieve an equilibrium.

But if (H´) is true of the game we are imagining, then *all* of the game's equilibria, and thus all of the outcomes attainable by rational players, are strictly deficient. This means that for any given outcome at the intersection of rational choices, at least one outcome exists that every player prefers to the given one. Since this is the tragedy we found in games 1–4, the property (H´) refers to is clearly *sufficient* to produce the tragedy.

That property also seems *necessary* to produce the tragedy. Suppose that (H´) is false of our imagined game. Then either the game is not solvable, and hence is clearly without the tragedy, or else it is solvable and has at least one available equilibrium that is not strictly deficient. If it's a game of the latter sort, and if we assume, as seems reasonable, that *all* of its equilibria are achievable – i.e., that in a solvable game, every available equilibrium qualifies as a solution – then our players might achieve the non-deficient equilibrium just mentioned. But this implies that our imagined game has at least one outcome which, although attainable by rational players, is not unanimously dispreferred to some unachievable outcome. So the game does not exhibit the tragedy we found in the PD.

In sum, it seems that the tragedy we have been discussing occurs in all and only those games of which (H´) is true – that is, solvable games with at least

player, the conclusion that the other player will choose *E*. This fact, together with point (a) in section 2, makes game 5 unsolvable.

some equilibria, all of which are strictly deficient. So (H´) seems adequate as an account of the tragedy.

§5

But the above argument is flawed. It assumes that in a solvable game with several available equilibria, all the equilibria are achievable. If this assumption is false, there may be games that reveal the same tragedy we found in the PD, but in which *non*-strictly deficient equilibria are available. This will be the case if the equilibria just described have a feature that makes them unattainable by rational players, and if there exist other equilibria, attainable by such players, that are strictly deficient. For it will follow that rational players will achieve a unanimously dispreferred outcome (hence the tragedy), but contrary to (H´), the game will contain non-deficient equilibria. Thus, (H´) will not state a *necessary* condition for the tragedy.

Games of this sort are shown below. Each game has an achievable equilibrium, *DD*, which is strictly deficient. But each game has one or more additional equilibria, and the latter are neither strictly deficient nor achievable. (The additional ones are *DC* and *CD* in game 6, *DC* in games 7 and 8, and *CC* in game 9.) The additional equilibria are unachievable because each has at least one dominated component, and as the following argument shows, no equilibrium with a dominated component is attainable by rational players.

6:		L	
		C	D
H	C	1 1	2 0
	D	0 2	0 0

7:		L	
		C	D
H	C	1 2	2 0
	D	0 2	0 1

8:		L	
		C	D
H	C	1 2	2 0
	D	0 3	0 1

9:		L	
		C	D
H	C	2 2	2 0
	D	0 2	1 1

First of all, a rational player, being human, has all the imperfections typical of human beings – all those, that is, which are not ruled out by the other traits essential to being rational (being a utility-maximizer, etc.). And since each player knows that her co-player is human, she must assume that he has those imperfections. One such imperfection is the tendency to make 'mistakes' in carrying out one's intentions. The word is in inverted commas because I do not mean mistakes in judgment or choice, but the sort of failings or slip-ups that can cause us to execute our choices unsuccessfully, sometimes to the point of producing an alternative we chose *against*. Since each player knows that her co-player is human, she knows that he is prone to such failings. This is consistent with knowing that he reasons flawlessly, that he employs all relevant facts, etc., for to know such things about a person is not to know that he is unerring in carrying out his intentions. It is to know how he *forms* his intentions and beliefs, not that he always *succeeds* in producing the results he intends. Clumsiness or distraction, to name just two things, can cause a choice to be unsuccessfully discharged, even if the choice is fully rational.

The above points are true even if we are speaking of a very simple action, like uttering a word or pulling a lever. Nor is this always explained by something physical, e.g., a muscle spasm or a coughing fit. For instance, in the service I knew a man who was given a choice between two branches: infantry and artillery. He cringed at the thought of joining the infantry, so he chose artillery. Later he was told by friends that the word he had uttered in delivering his choice was not "artillery", but "infantry". At first he was distressed, but he persuaded himself that his friends were joking. They were not, and he soon received written orders for the infantry. The interesting point is that it would be false to say that he *chose* to join the infantry. He chose artillery, but discharged that choice unsuccessfully, without being aware of his error.[18]

How should we incorporate these observations into our assumptions about rational players? We already know that any choice made by such a player is utility-maximizing relative to his co-players' intentions. But what is it for a choice to be utility-maximizing relative to a co-player's intention to play, say, strategy D, if the execution of that intention might involve a slip of the hand, resulting in the play of some *other* strategy? A reasonable answer is this: a choice is utility-maximizing in the stated way if and only if it maximizes the relevant player's payoff, given that his co-player's intended strategy (in this case D) will "in all probability" be played, and his co-player's unintended strategies have "virtually no chance" of being played. To think that a strategy has "virtually" no chance of being played (rather than no chance at all) is to grant, without making a specific probability assignment, that there is *some* infinitesimal chance that it will be played (perhaps owing to a hand tremor). This thought, which at first seems hopelessly fuzzy, is easy to capture mathematically (see below, including note 19).

The remainder of the argument refers to game 10 (though it can be extended to apply to any normal form game). Suppose the following: (1) Outcome DC is in equilibrium, and D dominates C for player L; (2) L assumes that she will succeed in playing whichever strategy she chooses; (3) L predicts

[18] If this sentence and the preceding one seem false, this is because we often use "chose" more performatively than we use "intended" or "decided". (I thank Dennis Stampe for warning me about this.) In such cases, what the agent "chose" to do is simply what he *succeeded* in doing. When used this way the term has no place in game theory, where it's assumed a choice is the sort of thing that can be *required* by rationality. A person can be rationally required to form a particular intention, but he cannot be rationally required to *succeed* in carrying it out.

that *H* will choose *D*, which is to predict that *H*'s *D*-strategy will "in all probability" be played, and that his *C*-strategy has "virtually no chance" of being played. From (1) it follows that c = d and b > a. Given (2) and (3) (and the fact that *L* is rational), *L* will choose *C* only if $[(1 - \varepsilon)\mathrm{c} + \varepsilon\mathrm{a}] \geq [(1 - \varepsilon)\mathrm{d} + \varepsilon\mathrm{b}]$ for all sufficiently small $\varepsilon > 0$.[19] Since c = d, we can simplify this by saying that *L* will choose *C* only if $\varepsilon\mathrm{a} \geq \varepsilon\mathrm{b}$ for all sufficiently small $\varepsilon > 0$. But this, combined with the fact that b > a, implies that *L* will *not* choose *C*.

10:

		L	
		C	D
H	C	a a´	b c´
	D	c b´	d d´

The argument shows that if an equilibrium has a dominated component, at least one player will not choose her component of the equilibrium if she thinks her co-player will choose his. The conclusion that at least one player will shun her component depends crucially on the assumption that her component is dominated. To see this, assume that *D* does *not* dominate *C* for player *L*. This assumption, combined with the fact that *DC* is in equilibrium (hence that c ≥ d), implies that either c > d, or c = d and a ≥ b. But neither disjunct provides a way of showing that *L* will not choose *C*. For instance, suppose that c > d. If *L* can rationally choose *C*, it must be that $(1 - \varepsilon)(\mathrm{c} - \mathrm{d}) \geq \varepsilon(\mathrm{b} - \mathrm{a})$ for all sufficiently small $\varepsilon > 0$.[20] Since (c – d) has a positive value (given that c > d), the consequent of the preceding statement is true; hence we are furnished with no grounds for denying the antecedent: that *L* can rationally choose *C*.

At this point a suspicion might arise about the preceding sections. (Those sections tacitly assume what we are now rejecting: that a *choice* of a strategy

[19] To grasp the idea here, suppose for a moment that *H* cannot make mistakes – his intention to play *D* will be executed with no chance of failure. Given this, and given that *L* knows that *H* is choosing *D*, the second part of the footnoted sentence should read: "*L* will choose *C* only if c ≥ d". But we are supposing that *H can* make mistakes (hence that he might play *C* through a slip of the hand), which means that in the inequality just stated, c and d must be replaced, respectively, with $[(1 - \varepsilon)\mathrm{c} + \varepsilon\mathrm{a}]$ and $[(1 - \varepsilon)\mathrm{d} + \varepsilon\mathrm{b}]$, where ε is an infinitesimal probability. (Any greater probability would be unrealistic – player *H* is not a bungler.) This captures the idea that *L* evaluates her options on the understanding that *H*'s *choice* of *D* cannot be equated with a *sure chance of playing D* – there is a very slight, or "near-zero", chance that *H* actually will play *C*. The trick now is to capture the idea that *L* evaluates her options on the understanding that *H*'s choice of *D* correlates, not just with a near-zero chance of playing *C*, but with a near-zero yet otherwise *indeterminate* chance of playing *C* – i.e., that *H* plays *C* not with some *fixed* probability, but with nothing more specific than a probability of "virtually" zero. This is where the phrase "for all sufficiently small $\varepsilon > 0$" enters the picture. Suppose that *L* assigns a small probability ε to *H*'s unintended strategy (*C*), and then assigns, as indeed she must, a probability of $(1 - \varepsilon)$ to *H*'s intended strategy (*D*). Now let ε approach zero. If *C* is a rational choice for *L*, then at some point *before* ε reaches zero, and for every positive value from there on, it will be true that $[(1 - \varepsilon)\mathrm{c} + \varepsilon\mathrm{a}] \geq [(1 - \varepsilon)\mathrm{d} + \varepsilon\mathrm{b}]$. This is the idea expressed by the footnoted sentence, given the presence of the phrase "for all sufficiently small $\varepsilon > 0$".

[20] The inequality here is equivalent to the one in the preceding paragraph, third from last sentence.

can be equated with a *successful play* of that strategy.) The assumption that each player treats his co-player's unchosen strategies as having a near-zero, though positive, probability enables us to show that some equilibria are unattainable. Perhaps it also enables us to show that some non-equilibria are achievable. If so, we must retract some of the statements in sections 2–4, particularly the claim that rational players can attain no outcome that fails to be in equilibrium.

Fortunately, this suspicion is unfounded. Suppose that *DC* (in matrix 10) is not in equilibrium because, say, $c < d$. Suppose also that *L* expects *H* to choose *D*. If *DC* is achievable then *C* is a rational choice for *L*, which in turn implies, falsely, that $(1 - \varepsilon)(c - d) \geq \varepsilon(b - a)$ for all sufficiently small $\varepsilon > 0$. The latter implication is false because as ε approaches zero, the right side of the inequality approaches zero and the left side approaches $(c - d)$. But $(c - d)$ has a value *less* than zero, given that $c < d$. So *DC* is unattainable. The general point is that the assumptions introduced in this section do not undermine the claim that only equilibria are attainable.

But now an objection arises concerning the preceding five paragraphs. It says that they ignore something, namely that each player resembles her co-player in being prone to occasional mistakes. If *L* assigns an infinitesimal probability to *H*'s unchosen strategy, she must do the same to her own (no matter what it is). This will affect the way she calculates and compares the utilities of *C* and *D* in game 10. Given that *H* is choosing *D*, *L* can rationally choose *C* only if $\{(1 - \varepsilon)[(1 - \varepsilon)c + \varepsilon d] + \varepsilon[(1 - \varepsilon)a + \varepsilon b]\} \geq \{(1 - \varepsilon)[(1 - \varepsilon)d + \varepsilon c] + \varepsilon[(1 - \varepsilon)b + \varepsilon a]\}$, for all sufficiently small $\varepsilon > 0$.

This objection is reasonable, but it alters none of the results in the paragraphs to which it pertains. For example, if *DC* is in equilibrium and *D* dominates *C* for *L*, then $c = d$ and $b > a$. Given that $c = d$, we can simplify the last point in the above objection by stating that *L* will choose *C* only if $[(1 - \varepsilon)a + \varepsilon b] \geq [(1 - \varepsilon)b + \varepsilon a]$ for all sufficiently small $\varepsilon > 0$. This statement, combined with the fact that $b > a$, implies that *L* will not choose *C*. This is the same conclusion we reached earlier, when arguing that an equilibrium with a dominated component is not achievable.

In sum, we can say four things about the assumptions introduced in this section. First, they force us to conclude that any equilibrium with a dominated component is unattainable. Second, they do not force us to conclude that some equilibria without dominated components are unattainable. Third, they do not contradict the results in sections 2–4. In particular, they do not support the view that some non-equilibria are attainable. Fourth, the preceding three points are true regardless of whether we imagine the players assigning infinitesimal probabilities only to their *co-players'* unintended strategies, or instead imagine them doing so to *all* such strategies, including their own.

§6

Let us return to games 6–9. In those games, every equilibrium but *DD* has at least one dominated component, making each of those equilibria unattainable. Since *DD*, which is strictly deficient, is the only attainable cell, games 6–9 are as tragic as games 1–4. The upshot is that answer (H´) does not state a necessary condition for the tragedy.

Fortunately, we can repair (H´) by replacing the term "equilibria" with "trembling-hand equilibria". A trembling-hand equilibrium is an equilibrium from which no player would deviate, even after assigning a slight probability to each of the pure strategies of his co-players, excluding those strategies that produce the equilibrium.[21] Such outcomes have three important features: first, at least one of them exists in every normal form game; second, such equilibria have no dominated components; and third, in bimatrix games, every equilibrium without dominated components is a trembling-hand equilibrium.[22] Given the second two features, we can state our conclusion about games 6–9 as follows: they are as tragic as games 1–4 because they have a single trembling-hand equilibrium, *DD*, which is jointly dispreferred to *CC*.

The preceding conclusion is not meant to imply that the presence of a *single* trembling-hand equilibrium is essential to the tragedy. Game 3 exhibits the tragedy, but contains *two* trembling-hand equilibria: *DD* and *DE*. These cells are its only equilibria. So the important fact about games 6–9 – about game 3 as well – is that they have *some* trembling-hand equilibria, all of which are strictly deficient. If we examine games 1, 2 and 4, and those in note 16, we see that they too have that feature. They resemble games 6–9 in having a single trembling-hand equilibrium, *DD*, which is strictly inferior to *CC*. So we seem to have accounted for the tragedy in the PD. The tragedy occurs in any solvable game with at least some trembling-hand equilibria, all of which are strictly deficient.

I believe this account is correct. In fact, I think the argument for (H´) (in section 4) could be repaired by prefacing every occurrence of "equilibrium" with the term "trembling-hand", and making a few minor adjustments. The argument for (H´) rests on a false assumption, but that assumption would be true if it were about trembling-hand equilibria rather than about equilibria in general. The assumption is that in a solvable game, *any* available equilibrium, and no other type of outcome, is attainable by rational players. This assertion is false as it stands, but true if we replace "equilibrium" with "trembling-hand equilibrium". Hence we arrive at the following account of the tragedy in the PD:

[21] For a more formal and precise definition see Selten (1975), who developed the concept in question, but used the term "perfect equilibrium". (I have borrowed the term "trembling-hand equilibrium" from Binmore and Dasgupta 1986, p. 16.) See also Van Damme 1987, pp. 13f, 25–28.

[22] For proofs of these points see Van Damme's indispensable monograph (1987), pp. 26ff, 48f.

(H´´) The game is solvable, and contains one or more trembling-hand equilibria, all of which are strictly deficient.

§7

An objection to (H´´) might arise, similar to the criticism of (H´) at the start of section 5. The objection grants that rational players will attain a trembling-hand equilibrium, but challenges the claim that in a game with multiple equilibria of that sort, *all* of those equilibria are attainable. Perhaps the class of attainable outcomes is included in, but narrower than, the class picked out by the trembling-hand concept, and is properly singled out by a refinement of that concept. Suppose this is true, and suppose we can find games in which some trembling-hand equilibria are neither strictly deficient nor achievable, and in which the achievable trembling-hand equilibria – the ones picked out by our refinement of the trembling-hand concept – are strictly deficient. In these games rational players will attain a unanimously dispreferred cell, but contrary to (H´´), the games will have trembling-hand equilibria that are not deficient. Thus, (H´´) will not state a *necessary* condition for the tragedy it is meant to explain.

I see no *decisive* way to rule out this objection, particularly given the abundance and unceasing development of new equilibrium refinements. But I'm aware of no present refinement that could make the objection forceful.[23] Nor am I aware of any good reason to think that some trembling-hand equilibria (in solvable games with only pure strategies available) are unachievable. The nearest thing I can think of is a view inspired by R. B. Myerson (1978): that a rational player not only makes occasional mistakes, but makes them in a "rational" way, meaning that she makes a given mistake less frequently the more costly it is, its "costliness" being a function of the payoffs in the relevant

[23] This requires an explanation. In the first place, some of the equilibrium refinements that currently receive attention are not refinements of the *trembling-hand* concept. I have in mind the following: *subgame-perfect equilibrium* (Selten 1965); *persistent equilibrium* (Kalai and Samet 1984); and *sequential equilibrium* (Kreps and Wilson 1982). Second, two refinements that *do* mark off subgroups of trembling-hand equilibria – *essential equilibrium* (Wu Wen-Tsün and Jiang Jia-He 1962), and *strongly stable equilibrium* (Kojima, Okada and Shindoh 1985) – were developed to deal with perturbations of a game's *payoffs*, and thus are irrelevant to our present concerns, since we are trying to determine the outcomes that rational players will attain on the assumption that their payoffs remain fixed. (For proof that these two types of equilibria mark off subgroups of trembling-hand equilibria, and for an excellent discussion of all the equilibrium refinements in this note, see Van Damme 1987.) Finally, two other equilibrium refinements that single out subsets of trembling-hand equilibria – *strictly perfect equilibrium* (Okada 1981) and *regular equilibrium* (Harsanyi 1973) – designate types of equilibria that are not present in every normal form game, and thus are poor candidates for picking out the general type of outcome rational players are sure to achieve. This leaves the notion of a *proper equilibrium* (Myerson 1978) – a refinement of the trembling-hand concept, and a type of equilibrium that exists in every normal form game – as a possible candidate. I reject it for reasons stated below. (An afterthought: the remarks in this note might suggest that the above equilibrium refinements are the only ones widely discussed. They aren't. Unfortunately, I cannot mention *all* the important ones, since too many exist.)

matrix.[24] Myerson demonstrates that if the players have this characteristic, and know that every player has it, two things follow: first, some trembling-hand equilibria do not furnish solutions to the games that contain them; but second, every normal form game has at least one trembling-hand equilibrium that provides a solution. He calls such outcomes *proper equilibria*.[25]

For our purposes, however, Myerson's demonstration can go unexplained. This is because the assumption just mentioned – the assumption required to show that the *properness* concept, rather than the *trembling-hand* concept, singles out the class of outcomes achievable by rational players – is without support, given our definition of rational players. That definition neither entails, nor ensures in some other way, that a rational player is less likely to make a mistake the more costly it is.

For instance, we cannot assume that a player has the given attribute – the tendency to make mistakes in a "rational" way – from the fact that he is *human*, for that tendency is not displayed by actual human beings. (Remember, by "mistakes" we do not mean errors in judgment, but the sort of slip-ups caused by distractions, fatigue, etc.) If this seems doubtful, we must become clear on assumption in question. The assumption is that if two mistakes differ in terms of costliness, the less costly one is more probable than the other. For example, if a person has a tendency to slip and fall, and if there are two kinds of streets in his city, those made of asphalt and those made of concrete, he is less likely to slip on the second kind than on the first, since his injuries are likely to be slightly more serious. This, I believe, is counter to what we observe about people. Further examples reinforce my point. For instance, I find that I cut myself with sharp knives just as often as I do with dull ones, even though the former cuts are more "costly" than the latter. And I am just as likely to hit my

[24] See Van Damme (1987, p. 15), who uses this assumption to motivate his discussion of proper equilibria.

[25] Not to be confused with what David Lewis (1969, p. 22) means by that term. Lewis has in mind what I have been calling a *strong* equilibrium. Myerson's definition of a proper equilibrium is highly technical; I will not state it here. Informally, such outcomes can be defined as equilibria from which no player departs even when every pure strategy deviation from the equilibrium is assigned a small probability, the more costly deviations being assigned proportionally lower probabilities than the less costly ones. In game (iv) (adapted from Van Damme 1987, p. 14), outcomes *CC* and *DD* are trembling-hand equilibria, but *CC* is the only proper equilibrium. If each player intends to play *D*, but then assigns an infinitesimal though otherwise indeterminate probability to the *C* and *E*-strategies of the other player, neither player has an incentive to deviate from *D*. But if each player is confident that for the other player, mistake *C* is more likely than mistake *E* (*C* being the less costly of the two), the players will switch to *C*. So *DD* is not proper.

(iv)			L	
		C	D	E
	C	3 3	2 2	0 1
H	D	2 2	2 2	0 2
	E	1 0	2 0	0 0

thumb with a *heavy* hammer as with a light one. Similarly, when I try to put money in a coke machine, and through clumsiness lose one of the coins, I find that the lost coin is just as often a dime as it is a nickel.

A natural objection to all this is that although our definition of rational players does not imply that they are more likely to make a mistake the less costly it is, we should modify our definition so that it carries that implication. After all, a truly *rational* agent tries to avoid errors, and takes special measures to avoid costly ones. If we include this in our definition of rational players, we must grant that the likelihood of a mistake varies inversely with its costliness.

The objection rests on this assumption: A person who is truly rational takes precautions against error, and the effectiveness of the precautions varies directly with the costliness of the possible errors. But this premise is false, at least as a generalization. First of all, whether a person should take precautions against mistakes depends largely on the cost of the *precautions*. Sometimes those precautions – the only effective ones, anyway – will involve so much time, effort or sacrifice that an intelligent person will not take them. Secondly, very often the measures we take to avoid a costly error not only reduce the likelihood of *that* error, but reduce, in an equal amount, the likelihood of many less costly ones. Suppose that in a particular game, strategy *A* is the best choice for player *H*. Suppose also that *H* wants to avoid blundering into playing *B*, *C*, or *D*, but wants even more to avoid *E*, which would be a more costly error than the previous three. In this situation, perhaps the only effective way to avoid playing *E* rather than *A* is simply to *concentrate very hard* on playing *A*. But then the precautions *H* takes against error *E* will be equally effective against the other three.

In sum, there is no compelling reason to assume that rational players make mistakes in a "rational" way; hence we have seen no reason to think that some trembling-hand equilibria are unattainable by such players, nor that we should replace the trembling-hand concept with the properness concept to isolate the solutions to normal form games. The notion of a proper equilibrium may have great value in some contexts, but it does not single out the general class of outcomes rational players will achieve.

§8

Before going further I will summarize the points I have defended:

(1) Answers (A)–(H) fail to account for the tragedy in the PD.

(2) Answer (H´), although better than (A)–(H), is not satisfactory. This is because some equilibria fail to qualify as solutions. Such equilibria are not *robust*: they cease to be at the intersection of utility-maximizing choices once we interpret "utility-maximizing" in light of the fact that rational players, being human, can make mistakes in

carrying out their intentions. This enables us to find counterexamples to (H´).

(3) Answer (H´´) is the correct account of the tragedy in the PD. Unlike (H´), it states a condition that is both sufficient and *necessary* for the tragedy's occurrence. It succeeds at this because the deficient equilibria it speaks of are trembling-hand equilibria. Such equilibria are robust in the required way.

(4) There is no need to replace the notion of a trembling-hand equilibrium with that of a proper equilibrium to obtain a correct solution concept, and hence to account for the tragedy in the PD.

The last two points are advanced tentatively. The literature on equilibrium refinement is enormous, and the debates are too complicated to be settled in a short discussion. But one thing is certain: if we regard our players as *human* as well as ideally rational, some equilibria, specifically those mentioned in point (2), fail to qualify as solutions. This ensures that (H´´) is closer to the truth than any other answer we have seen. If a better answer exists, it will resemble (H´´) in referring to a robust form of equilibrium. Fortunately this is all I need for the next section. What I say there will stand up even if we substitute "robust" for "trembling-hand".

§9

I will finish by discussing an implication of my argument, particularly of my claim that in a non-cooperative game with multiple equilibria, only the trembling-hand equilibria are solutions.

A common assumption about non-cooperative games is that even if they have multiple equilibria, we can easily single out a unique solution if one of the equilibria *payoff-dominates* the others – that is, if it is better for every player than any other equilibrium.[26] The reasons for this assumption are seldom made clear, but it's a natural assumption to make, and many have made it. If it's true, the PD is not *quite* as problematic as it appears. For with only a minor change in the players' preferences, game 1 can be turned into game 9, in which *CC* payoff-dominates *DD*.[27] The change is "minor" because it does not reverse any preferences, and because it alters only one payoff *per* player, leaving the most striking features of the game unaffected: in both games, *D* is dominant for each player, and cell *DD* is strictly deficient.

[26] See, for instance, Elster 1986, p. 9; Gauthier 1986, p. 71; and Ullmann-Margalit 1977, pp. 34, 80. For a criticism of the assumption see Gilbert (1990, sec. 2), who shows that the assumption is neither an obvious truth nor an implication of the game theorist's definition of rationality.

[27] It's clear that Ullmann-Margalit (1977, p. 30) would see this as a remedy to the problem in the PD.

1:

		L	
		C	D
H	C	2 2	3 0
	D	0 3	1 1

9:

		L	
		C	D
H	C	2 2	2 0
	D	0 2	1 1

But if the argument of this paper is sound, game 9 is as tragic as game 1, meaning that the players will achieve *DD* rather than *CC*. This is because *DD* is the only trembling-hand equilibrium in the game. This is easily tested by focusing on equilibrium *CC*, and then assigning small probabilities to the *D*-strategies. It becomes obvious that neither player would choose *C*. But if the same test is applied to *DD*, its component strategies are shown to be rational. Of course, an even simpler test consists of noting that the components of *CC* are dominated, but those of *DD* are not.

Two things follow: First, it's false that a payoff-dominant equilibrium furnishes the solution to any game that contains one. In game 9, equilibrium *CC* is payoff-dominant but unachievable. Second, the preference change that transforms game 1 into game 9 is not sufficient to extricate the players from their predicament. The problem they face is serious: it cannot be overcome by a *minor* change in preference.

A general lesson is that the problem illustrated by the PD is easily underappreciated, owing a common way of stating it, a way that's natural given the specific *way* the PD illustrates it. The problem is that rational choices can produce an outcome that is worse for all participants than the outcome they might achieve if they behaved irrationally. When we try to state this in game theoretical terms, we are prone to equate it with an obvious thing shown by the PD: that a game can be structured so that every equilibrium is strictly deficient. The latter fact is sufficient to *produce* the problem we equate with it – this is why the PD successfully illustrates the problem. But to produce the problem is one thing, to state it accurately is another, and we do not state it accurately by saying that all of a game's equilibria can be strictly deficient. For this suggests that an optimal outcome (*any* non-strictly deficient outcome, for that matter) can be defective only by failing to be in equilibrium; hence that a game with an optimal equilibrium is unproblematic. Game 9 shows this to be false. Hence to state the problem accurately we must include the fact that optimal outcomes can be defective in a second way: they can be in equilibrium, but fail to be robust against minor uncertainties about whether the players will succeed in playing their chosen strategies.

The upshot is that the presence of an optimal equilibrium, even a payoff dominant one, does not ensure that rational players can avoid a unanimously dispreferred outcome. The problem for rationality associated with the PD is more persistent than some have thought.

REFERENCES

Binmore, K., and P. Dasgupta: 1986, "Game Theory: A Survey", in *Economic Organizations as Games*, K. Binmore and P. Dasgupta, eds., Basil Blackwell, Oxford.

Elster J.: 1986, Editorial Introduction to *Rational Choice*, New York University Press, New York.

Gauthier, D.: 1986, *Morals by Agreement*, Oxford University Press, Oxford.

Gilbert, M.: 1989, "Rationality and Salience", *Philosophical Studies* **57**, 61–77.

Gilbert, M.: 1990, "Rationality, Coordination, and Convention", *Synthese* **84**, 1–21.

Hamburger, H.: 1979, *Games as Models of Social Phenomena*, W. H. Freeman, New York.

Harsanyi, J.: 1973, "Oddness of the Number of Equilibrium Points: A New Proof", *International Journal of Game Theory* **2**, 235–50.

Kalai, E. and Samet, D.: 1984, "Persistent Equilibria in Strategic Games", *International Journal of Game Theory* **13**, 129–44.

Kojima, M., A. Okada and S. Shindoh: 1985, "Strongly Stable Equilibrium Points of *n*-Person Noncooperative Games", *Mathematics of Operations Research* **10**, 650–63.

Kreps, D., and R. Wilson: 1982, "Sequential Equilibria", *Econometrica* **50**, 863–94.

Lewis, D.: 1969, *Convention: A Philosophical Study*, Harvard University Press, Cambridge, Mass.

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

Myerson, R. B.: 1978, "Refinements of the Nash Equilibrium Concept", *International Journal of Game Theory* **7**, 73–80.

Okada, A.: 1981, "On Stability of Perfect Equilibrium Points", *International Journal of Game Theory* **10**, 67–73.

Rapoport, A., M. J. Guyer and D. G. Gordon: 1976, *The 2x2 Game*, University of Michigan Press, Ann Arbor.

Resnik, M.: 1987, *Choices: An Introduction to Decision Theory*, University of Minnesota Press, Minneapolis.

Selten, R.: 1965, "Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit", *Zeitschrift für die Gesamte Staatswissenschaft* **121**, 301–24, 667–89.

Selten, R.: 1975, "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games", *International Journal of Game Theory* **4**, 25–55.

Ullmann-Margalit, E: 1977, *The Emergence of Norms*, Oxford University Press, Oxford.

Van Damme, E.: 1987, *Stability and Perfection of Nash Equilibria*, Springer-Verlag, Berlin.

Wu Wen-Tsün and Jiang Jia-He: 1962, "Essential Equilibrium Points of *n*-Person Noncooperative Games", *Scientia Sinica* **11**, 1307–22.

Zagare, F.: 1984, *Game Theory: Concepts and Applications*, Sage, Beverly Hills.