TRUST OUT OF DISTRUST*

Ever since people started reading Thomas Hobbes[1] with game theory-informed eyes, it became commonplace to see his state of nature as a prisoner's dilemma (PD-) structured situation. Starting out with his assumption of thoroughgoing egoism and given his description of the state of nature, Hobbes's conclusion is considered compelling: that human beings are destined to be locked into a state of mutual suspicion and distrust, of war of all against all. This state is a stable equilibrium, where distrust dominates trustful cooperation.

The interpretation of Hobbes's state of nature as a PD-structured situation has given rise to a puzzle: How can cooperation, which is based on trust, arise out of the distrust in which human beings seem to be irredeemably locked? After all, social life exists and depends on cooperation, and cooperation requires trust. To be sure, trust and cooperation are not identical, and neither of them is a necessary or a sufficient condition for the other. Yet there is significant overlap between the two. Much of what we regard as trustful behavior also falls under cooperative behavior. This is especially so when the relationships involved are thin rather than thick: when there are no special ties of kinship, friendship, a shared past, and the like among the people who are facing the choice of whether or not to cooperate trustfully with one another.[2] In a Hobbes-like state of nature, where relationships are thin and the options crude ("cooperate with others" or "fend for yourself"), cooperation more or less converges on trust. 'I trust you' reduces to my belief that you will do your part in the cooperative venture that serves the interests of us both, where the success of this venture depends on each of us doing our part.

But how do people mired in distrust and suspicion succeed in bootstrapping themselves onto promise keeping and trust? Hobbes's

* I wish to thank Robert Aumann, Russell Hardin, Dimitri Landa, Ariel Rubinstein, Frederick Schick, Cass Sunstein, and also the participants in the New York University Colloquium on Law, Philosophy, and Social Theory (fall 2001) and its conveners Ronald Dworkin, Thomas Nagel, and Jürgen Habermas. I am especially indebted to Piero La Mura and Gil Kalai for their generous help with the mathematical and game-theoretical aspects of this article, and to Avishai Margalit for his role in its shaping and sharpening throughout.

[1] *Leviathan*, C. B. MacPherson, ed. (New York: Penguin, 1982).

[2] For an analysis of trust in thick relationships, see, for example, Russell Hardin, "Distrust," *Boston University Law Review*, LXXXI, 3 (2001): 495-522; and my "Trust, Distrust and in between," in Hardin, ed., *Distrust* (New York: Sage, 2002, forthcoming).

0022-362X/02/9910/532–48

own solution, in the form and figure of the sovereign, is often regarded as somewhat of a deus ex machina and hence as not altogether satisfactory. The sovereign is external, and no satisfactory mechanism is or can be provided for his installment within the Hobbesian framework. And so the puzzle persists.

It is my aim here to establish the possibility of trust from within the Hobbesian framework. I shall show that distrust can be structured in two different ways: the first is commonly associated with Hobbes, and the second is sometimes associated with that other patron saint of social-contract theory, Jean-Jacques Rousseau. The two structures are here referred to as *hard* and *soft*, respectively, and they are both in essence compatible with Hobbes's stark assumptions about human nature. Hard distrust is PD-structured: the distrust strategy is dominant in it. Soft distrust is not PD-structured: the choice of trust in it is an equilibrium choice. In order to establish the possibility of trust, I do not need to claim or show that the state of nature is soft rather than hard, nor even that it is likelier to be so. All that is needed for trust to get off the ground is ambiguity, or uncertainty, as to which of the two basic situations of distrust obtains.

I. TWO SIMPLIFIED GAMES OF DISTRUST

What explains some people's tendency to consider distrust safer than trust? Why does some of the literature on trust take it almost for granted that "fairly generalized distrust might make sense in a way that generalized trust does not," and that suspicion and distrust are "inherently well grounded"?[3] There are, in principle, two alternative presumptions for dealing with cases of doubt about trust, one in favor of trust and one in favor of distrust. If one of them is to prevail as the default presumption, why is it often instinctively felt that the presumption of distrust should be the one?

Consider the following rough calculation of best and worst scenarios. First, the case of trust: acting on trust when trust is reciprocated can lead to successful cooperation and hence to mutual benefit and potentially to significant gain. Acting on trust when trust is not reciprocated and when one's partner is untrustworthy inevitably involves disappointment. It often involves worse: being betrayed or exploited. This may lead to serious damage. Consider next the case of distrust: acting on distrust when distrust is reciprocated leads roughly to whatever gains or benefits one is able to achieve on one's own. But what does acting on distrust lead

[3] Hardin, p. 500; while these are quotes from Hardin's article, they do not express Hardin's view.

to when distrust is not reciprocated (that is, when one's partner is trustworthy and trusting)?

Here we may want to look at two different possibilities. One possibility involves a structure in which unilateral distrust, while entailing foregone opportunities, basically yields to the distrusting partner the same benefits as yielded by reciprocated distrust. On a natural interpretation of this type of case, the distruster is an individualist who "goes it alone," indifferent as to whether his partner chooses to trust or to distrust him. In the other possibility, the structure is such that unilateral distrust does benefit the distruster, at the expense of her trusting partner. In this type of case, the truster is being exploited by her partner. It is not two psychological types that are differentiated here, but rather two different structures of an interactive situation, as will now be shown.

II.A. Soft distrust. Here is a simple two-person payoff matrix to represent the structure of the first type of situation:

	T	DT
T	*4*;4	*0*;2
DT	*2*;0	*2*;2

Figure 1: Soft distrust (lack of trust)

The two strategies open to each of the players are T for trust (or cooperation) and DT for distrust (or defection). If I as row-chooser (payoffs in italics) choose to trust you, I derive a payoff of 4 when you trust me, too, but as low as 0 when I am a lone truster. If I choose to distrust you, I derive a payoff of 2 regardless of whether you trust me or not. And the same goes for you, column-chooser.

This is a version of the stag-hunt game, which is famous in game theory as well as in philosophy[4]: "Separately we can catch rabbits and eat badly. Together we can catch stags and eat well. But if even one of us deserts the stag hunt to catch a rabbit, the stag will get away; so the other stag hunters will not eat unless they desert too" (*ibid.*, p. 7; cf. 95n.).[5] The credit for the original version goes to Rousseau,[6] who intended the story as a prototype of the social contract.

[4] See David Lewis, *Convention* (Cambridge: Harvard, 1969).

[5] For further discussion, see my *The Emergence of Norms* (New York: Oxford, 1977), pp. 121-24.

[6] *A Discourse on Inequality*, Maurice Cranston, trans. and intro. (New York: Penguin, 1984).

So the trust strategy (T) can lift you high or it can make you fall. It can be disappointing. Its alternative strategy (DT) is less risky: (DT) guarantees a payoff of 2, as compared with the risk of getting 0 for choosing (T). Distrust is here basically constant. It leaves you on some plateau that is insensitive to changes in your environment. At the same time, it is insensitive to the disappointments you may cause to others. There are no possibilities for you to exploit the other, nor are you in danger of exposing yourself to being exploited by the other. To the extent that 'safe' means hedging your bets, minimizing your potential losses, being risk-averse—distrust here seems indeed to be safer.[7] Both (T; T) and (DT; DT) are (Nash) equilibria. While the first equilibrium point is efficient in that it promises higher payoffs for both players, the second is the minimax choice.

But let us now generalize this situation and think of it not as a one-round encounter but rather as a repeated game with pure strategies. If the players start by playing it safe and adopting distrust, they may remain stuck with the suboptimal equilibrium in the future repetitions of the game. Suppose, however, that they succeed in coordinating on the better equilibrium, whether through communicating with each other or by responding to some commonly observed cues, or by any other way. Then they will both reap the higher fruit of trustful cooperation and neither of them will be tempted to deviate to distrust in the future repetitions of the game.

As was already noted, in this version of the two-person trust-distrust game, the truster is not exploited by the distruster. While the unilateral truster, whose trust is not reciprocated, suffers a loss (0), the distruster's payoff does not increase at her partner's expense (it remains 2 regardless of what the other player does). Strictly speaking, then, this game captures soft distrust; namely, it is a situation where one's unilateral employment of the DT strategy reflects what may be interpreted as mere lack of trust. Distrust here involves foregoing opportunities: while causing modest harm, it mostly protects against it.[8]

Let us generalize the interactive situation further: not just from a one-round to a repeated situation but also from two participants to a community of n players. If we imagine this community to be repeat-

[7] Not only risk aversion but also loss aversion may play a role here. If you do not trust, you fail to get certain gains, as compared to your starting point; if you distrust, you protect yourself from losses, as compared to your starting point. For most people, losses are much "more bad" than gains are "good." So there might be a connection between a presumption of distrust and loss aversion.

[8] Compare Hardin, pp. 495-96

edly involved in situations of the soft-distrust kind, we may speculate that over time they will tend to develop "soft distrusting" dispositions. We may imagine them as a Wild West community of rugged individualists, honest folks who rely on no one and exploit no one. At the same time, however, there is no need to think of them as blind or averse to the possibility of joint ventures requiring trustful cooperation (like constructing a dam, say) and to their mutual benefits.

Note, moreover, that the payoff matrix of figure 1 and the story of the generalized soft distrust which goes with it may be given two different interpretations: a threshold one and an all-or-nothing one. (These are, in fact, well-known variants of the *n*-player stag hunt.[9]) In the first, in order to enjoy the higher payoff it is *not* necessary that all *n* participants choose the T-strategy: it is enough that the number, or proportion, of participants who choose T goes above a certain threshold. Think of the case, say, of daylight saving time, where those who do not switch their clocks are those who suffer the inconvenience, once a sufficient number have switched. In the second interpretation, in order to enjoy the higher payoff it *is* necessary that all *n* participants choose the T-strategy. Here the interpretation of the matrix is as in the original story of the stag hunt mentioned earlier.

II.B. Hard distrust. Let us now consider the second way to conceptualize the two-person trust/distrust situation. This version captures distrust of the hard variety, where one's unilateral trust turns out to benefit the distruster at the expense of the truster.

	T	DT
T	*4*;4	−*2*;6
DT	*6*;−2	*2*;2

Figure 2: Hard distrust (betrayal of trust)

If I as row-chooser (payoffs in italics) choose to trust you (T), I derive a payoff of 4 when you trust me, too, and our cooperation succeeds. But I get as low as −2 when my trust is not reciprocated: my trust here may be seen as betrayed and exploited. If I alone choose to distrust you (DT), I in effect exploit you. My payoff is as high as 6, to

[9] See, for example, Hans Carlsson and Erik van Damme, "Equilibrium Selection in Stag Hunt Games," in Ken Binmore, Alan Kirman, and Piero Tani, eds., *Frontiers of Game Theory* (Cambridge: MIT, 1993), pp. 237-53.

your –2. When we both distrust each other, each of us must settle for 2, which is what we can achieve on our own. Once again, the game is symmetrical between row-chooser and column-chooser.

It should come as no surprise that this is a PD-matrix. After all, lone defection in the canonical PD-situation is, for either prisoner, precisely what amounts to betrayal of the other prisoner's trust. Mutual trust is not an equilibrium point here: it is unstable. If no credible way can be devised for trust to be thrust upon the players, this jointly desired point remains all but inaccessible to them (in the one-round version of this game).

Except for the case where this game is indefinitely repeated, the distrust strategy (DT) is dominant.[10] It is hence, a fortiori, the safe strategy: trust has no chance here. It should be noted, however, that this statement holds so long as we focus on strict theory, and on agents whose behavior is restricted by the requirements of instrumental rationality. In practical PD-type situations, boundedly rational agents will find routes to cooperate and to justify cooperation.[11] Many experimental results, as well as ordinary experience, attest to this discrepancy between the predictions of (game) theory and actual behavior.[12]

In fact, more than risk aversion and loss aversion may be at work here. People are often also highly "betrayal averse" in the sense that they will tend to punish those who betray their trust. Indeed, the punishment for injury that comes from betrayal of trust will often be

[10] More precisely: the DT strategy is the only strategy that survives the process of iterated elimination of dominated strategies. See D. Fudenberg and J. Tirole, "Perfect Bayesian Equilibrium and Sequential Equilibrium," *Journal of Economic Theory*, LIII (1991): 236-60.

[11] See, for example, Reinhard Selten, "The Chain Store Paradox," *Theory and Decision*, IX, 2 (1978): 127-59, where Selten promotes the distinction between *rational* (theory-driven) behavior and *reasonable* (practically oriented) behavior. This echoes the point made early on, by Duncan Luce and Howard Raiffa, in *Games and Decisions* (New York: Wiley, 1957), pp. 97-102.

[12] This statement should be qualified, as nowadays the theory attempts to make somewhat different predictions between strictly dominated strategies, where trust has really no chance, weakly dominated ones, where it depends, and iteratively dominated ones (like cooperation in the finitely repeated PD), where it also depends. See J. Y. Halpern, "Substantive Rationality and Backward Induction" (2000), in ewp-game/0004008 (consult http://econwpa.wustl.edu/eprints/game/papers/0004/0004008.abs) where both Robert J. Aumann's and Robert Stalnaker's approaches to the epistemic conditions behind backward induction are compared and discussed.

more severe than for an equivalent injury that involves no betrayal of trust.[13]

If a distrust situation of the soft kind is pictured as conducive to a Wild West community of individualists, we may picture the distrust situation of the hard kind as conducive to a community of classical Western gunmen. A John Wayne type is forever suspicious, forever on the alert to be the first to draw his gun and to take advantage of any moment of weakness on the part of his mates, as well as to engage in self-defense against them.

III. THE STATE OF NATURE

Two games or models of archetypal situations of distrust have been presented: soft and hard, or if you will, mild and harsh. Not every situation where the question of trust versus distrust comes up is PD-structured, and, as we saw, in those cases that fall under the soft category, mutual distrust is not a dominant strategy. In fact, in this class of cases, mutual trust not only leads to a jointly beneficial outcome but it is a strict equilibrium and, as such, it is accessible to the participants. Still, when people think of paradigmatic cases of trust versus distrust, it is often PD-structured situations that they have in mind, namely, cases of the hard variety. To many people, it would seem almost axiomatic that a general distrust strategy dominates a general trust strategy inasmuch as being a truster is often considered much worse than simply taking a chance: it is actually taken to *mean* being a sucker by exposing oneself to exploitation.

The general outlook of hard distrust may well derive from the powerful hold that Hobbes's grim picture has over us, the picture of the state of nature as a state of suspicion of all in all and of a war of all against all. A Hobbesian will tend to interpret many social interactions, whether micro or macro, as one-round PD-structured games. And it is this general outlook that motivates a general presumption in favor of distrust.

But even Hobbes himself, in presenting what he calls the "precept, or general rule of Reason," distinguishes between two situations. He says, first: "Every man ought to endeavor Peace, as farre as he has hope of obtaining it"; and then he continues: "and when he cannot obtain it, that he may seek, and use, all helps, and advantages of Warre." The first part of this precept contains what is for Hobbes the Fundamental Law of Nature: *to seek peace and follow it.* The second part

[13] For more on this, see Jonathan J. Koehler and Andrew D. Gershoff, "Betrayal and Aversion" (unpublished manuscript, 2000). I owe this point to Cass Sunstein, whose example is our attitude to injury caused by our own car's airbag.

sums up what he refers to as the Right of Nature: *by all means we can, to defend our selves* (*op. cit.*, chapter XIV). There is nothing far-fetched or strained, as far as I can see, in interpreting the first part of the precept as applying to situations of soft or mild distrust, and the second as applying to situations of hard or harsh distrust.

True, Hobbes did not believe that one's endeavoring peace in the hope of obtaining it would get one very far. A close reading of the relevant passages reveals how deeply convinced he was that we are doomed constantly to seek and use the advantages of war to defend ourselves. But still, the important point is that he did seem to recognize the possibility that the state of nature be construed in terms of soft distrust as well as hard distrust. And, as I shall now proceed to show, this possibility is all we need in order to establish the possibility of trust.[14]

In principle, if we are able to diagnose our situation as belonging to either the hard or the soft category, we are in familiar territory. But what if we cannot tell? More intriguingly, what if we do not know to which category the state of nature belongs—if we do not know, that is, which of the two games nature has put us in? The possibility of such an ambiguity is my leading idea in establishing the possibility of trust.

IV. THE GAME OF NATURE: SOFT OR HARD?

Suppose that we are playing a distrust game. We know that the game we play is either hard or soft distrust, but we do not know which of the two it is. (A reminder: going clockwise from top left, the payoff matrix of soft is (4;4), (0;2), (2,2) (2;0) and of hard it is (4;4) (−2;6) (2;2)

[14] There is a question whether the adequate model for the Hobbesian state of nature is that of *compulsory* interactions or that of *optional* ones. As introduced by Philip Kitcher—"The Evolution of Human Altruism," this JOURNAL, XC, 10 (October 1993): 497-516—an optional game is a situation in which a player is allowed two degrees of freedom: the possibility of signaling which partners she is willing to play with, and the possibility of opting out of the game altogether. Now, it may well be that the evolution of altruism (or of trust) works better, along the lines suggested by Kitcher, in the context of repeated optional PD-games than of repeated compulsory ones, where players have neither degree of freedom. The account I shall offer here is not evolutionary, and it is meant to apply to the pessimistic framework of a Hobbesian state of nature with nonoptional interactions. A fortiori, then, it will apply to the optional framework, where discriminating partners—prepared to trust anyone who has not played DT in a past interaction with them—would be expected to do better. (Note, however, that on my interpretation of the two distrust games, the choice of DT amounts to "going it alone," and hence to opting out of the game altogether. This signals the existence of the intermediate category of nonoptional yet *partially compulsory* games, with just one degree of freedom: you have no choice about your partner/opponent, but you can opt out. This category seems relevant in both evolutionary and nonevolutionary accounts of the emergence of trust.)

(6;−2); hard is a PD, soft is not; the strategies are T for trust and DT for distrust; the strategy (T;T) is an equilibrium in soft but not in hard.) How are we to think about our situation? How should we play? And, crucially, might trust prevail?

If both of us play T, each of us gets 4 regardless of which game we are playing, and if both of us play DT, we get 2 each; in either case with a symmetrical choice of strategy, we shall still not know which game we are playing. Only if we play asymmetrically, that is, if one of us plays T and the other plays DT, will we know for sure which game we are playing. Let us look at a story that is meant to make vivid the possibility that two people might find themselves not knowing whether the situation they are in is, in fact, one of hard or soft distrust.

Two survivors of an air crash find themselves on the desert side of an island. Separately, they can survive by farming, each having their own well of water. Before they crashed, however, they were able to see that across the desert there is a river, the other side of which looks opulent and green. But crossing the river requires a bridge, and it takes two to build a bridge. Let us assume that the payoff for each islander from staying put and farming is 2, and from crossing to the other side of the river, where the living is (or promises to be) easy, is 4. Now suppose they agreed that, if they were to survive the crash, they would meet at the river and join forces to cross it. Al sets out on his journey across the desert, trusting that Bill will do the same. But what if Bill does not reciprocate?

We are to imagine two situations. In the first, Al realizes after a while that he has been stood up and he returns to farm his land. He pays a price for having crossed the desert in vain and for having lost farming time, so his payoff is 0, while Bill's remains 2. (Of course, the situation is symmetrical between them: had Bill been the lone truster, his payoff would be 0 to Al's 2.) This corresponds to the soft-distrust game. In the second situation, it turns out that there is a subterraneous connection between Al's and Bill's wells. This means that when Al sets on his journey and stops farming, Bill has double the amount of water at his disposal. Moreover, while Bill's crops become better and more bountiful, Al's farm dries up. Here, then, Bill not only benefits from Al's lone trust, but his benefit is at Al's expense: his payoff goes up to 6, while Al's decreases to –2. This situation corresponds to the hard-distrust game.

Note that the different way distrust plays out in the two situations derives from the objective circumstances in which the islanders find themselves, and not from their being differently motivated. When Bill stands Al up, he certainly proves himself untrustworthy. But it is not

as if he can be described as "exploitative" when he does it in the second case as compared with his being more "benign" in the first. Rather, it so *happens* that in the second case Bill gains at Al's expense, whereas in the first case he does not. And, of course, as long as Al's and Bill's choice of strategy is symmetrical—that is, they either both stay put and farm, or they both set out for the river—they will not know whether they are in a soft or hard situation.

We now go back to the general case of a distrust game, where it is known that the game to be played is either hard or soft, but it is not known which of the two it is. Suppose that the probability p of the occurrence of the game of soft distrust is known. When this is the case, the two games of hard and soft can be superimposed on one another to yield one compound game: we can calculate the payoffs accordingly, check for equilibrium points, and decide on the best strategy.

	T	DT
T	4;4	$p \times 0 + (1 - p) \times (-2)$; $p \times 2 + (1 - p) \times 6$
DT	$p \times 2 + (1 - p) \times 6$; $p \times 0 + (1 - p) \times (-2)$	2;2

Figure 3: The compound game of hard and soft distrust

With these particular payoffs, if $p \geq 0.5$, that is, if chances are half (or more) that the probability that the game we are facing is soft, then the trust strategy T remains an equilibrium strategy. (The minimal value of p that sustains T as an equilibrium strategy will vary with varying payoffs.)

IV.A. Infinitely repeated game. Suppose now that this game situation recurs: we are to play this game repeatedly. That is, we go from a one-round game to a *super game*, and we are further to suppose that the repetitions we are talking about are infinite (or at least with no fixed horizon in view). If the game we play is soft, then T is an equilibrium strategy for each round, regardless of whether the game is played once or repeatedly. Repetition does make a difference, however, if the game we play is hard, which is a PD-type game. Here, as is well known, even though in the one-round game DT is the dominant strategy, in the repeated game the trustful strategy T may

be an equilibrium outcome.[15] Both players come to realize that there is less to gain by optimizing (that is, choosing DT) in the short run than to lose in the long run. Consider the strategy, known as *tit for tat*, that says: "I will play T in the first round and then do whatever my partner/opponent did in the previous round." This strategy, if adopted by both of us, is (Nash) equilibrium in the infinitely repeated game.

So it turns out that in the infinitely repeated case, T is an equilibrium outcome regardless of whether the game we play is soft or hard. And this means that *trust* will remain an equilibrium outcome in our compound super game, where we do not know which of the two games we are playing. Note that so long as we adopt the above strategy and stick to it without mistakes or glitches, we shall not know which game it is that we are playing. Nor will we have any incentive to want to know.

IV.B. Finitely repeated game. Things are different, however, when there is a finite and definite horizon. At least, things are different with regard to the hard game: with regard to soft, trust is an equilibrium strategy regardless of whether the game is played once, or infinitely many times, or finitely many times. But with hard PD-structured games, as soon as it is common knowledge between us that we shall play the game no more than, say, 1000 or 100 times, the entire cooperative scheme might "unravel from the back."

Consider: if we get to round 100 and it is common knowledge between us that it is the last, then "there is no longer-term loss to weigh against the short-term gain," and both of us will "defect" to DT. But then when we get to round 99, we both know that we shall play noncooperatively in round 100, and the same consideration applies

[15] This, roughly, is what is known as the *folk theorem*. It holds in this formulation as long as both players are patient enough, that is, when their "discount factor" is close enough to 1. For more details and technical niceties, see, for example, David M. Kreps, *A Course in Microeconomic Theory* (New York: Harvester Wheatsheaf, 1990), pp. 503-15; and Martin J. Osborne and Ariel Rubinstein, *A Course in Game Theory* (Cambridge: MIT, 1994), pp. 143-46. Note: it turns out that T is also an equilibrium outcome in yet another version of the infinite PD-game. Instead of imagining one pair playing indefinitely, we imagine a sequence of PD-games being played by pairs of people in a society with an infinite number of members. Everyone gets to play against everyone else, and each pair plays only once. The trustful cooperative solution applies here as it does in the case of PD being repeatedly played between the same two partners. The equilibrium strategy is: I play T unless somebody plays DT with me, in which case I switch to DT against my next partner and stick to it. The threat of "punishment" here is indirect, as I shall never play against you again, but you will play many times against people who will have played with me, so you will eventually be harmed. See, for example, Michihiro Kandori, "Social Norms and Community Enforcement," *Review of Economic Studies*, LIX (1992): 63-80.

to the current round as well; and so on, by so-called backward induction, to the very first round. "To get cooperation, there must always be a future substantial enough to outweigh immediate considerations."[16] The conclusion seems to be that in the finitely repeated hard game, there is no escaping distrust.[17] As was noted earlier, this conclusion is not necessarily borne out in practice: people do not betray each other as often and as harshly as the game-theoretic analysis predicts or prescribes. Cooperation in iterated PD, when the number of iterations is large, may even be the norm. And in any case, "unless the whole fixed-number iteration is extremely pristine, one cannot get the deduction going from the *n*th to the first."[18]

But now we come to consider the case where the game that is finitely repeated is our *compound* game. If we do not know which of the two distrust games we are playing, and we are to play the game finitely many times, what strategic recommendations apply? Specifically, might trust be an equilibrium choice here?

Trust may well prevail in the simple situation where there is common belief between us that chances are that we are playing soft (say, that $p > 0.5$). But now suppose that this is not the case. Specifically, we are interested in the case where we are both "pessimistic," in the sense that, in Hobbes's spirit, we both think chances are that the game we are facing is hard rather than soft. Various situations might be considered; I shall offer two.

(i) Chivalry. Let us consider first the following setup: one of our two games (or states of nature) is picked by nature, and played N (say, 1000) times. The probability that nature picked soft is a small epsilon (ϵ), the probability that the game to be played is hard is $1 - \epsilon$. The outcomes and payoffs in each round are commonly observed by both players, he and she. Consider the following strategy: both players play T up until k steps before the end of play. Now k steps before the end,

[16] This quote, as well as the previous quotes in this paragraph, are from Kreps, p. 514.

[17] More precisely, the conclusion is that trustful cooperation cannot be a subgame-perfect equilibrium. But, in fact, the stronger result holds as well, that trustful cooperation cannot even be a Nash equilibrium, as can be shown by a different argument; see Osborne and Rubinstein. Note, also, that in the case of a society where every pair plays only once, if the number of players is finite the backward induction will once again push them back to DT.

[18] Hardin, *Collective Action* (Baltimore: Johns Hopkins, 1982), p. 149. For his dissenting view of the backward induction argument, see Hardin's discussion there, at pp. 145-50. See also Philip Pettit and Robert Sugden, "The Backward Induction Paradox," this JOURNAL, LXXXVI, 4 (April 1989): 169-82; and Cristina Bicchieri, "Self-Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge," *Erkenntnis*, XXX (1989): 69-85.

that is, in round $N - k$, he continues to play T but she switches to DT. Since this is an asymmetrical strategy and the payoffs are commonly observed, both players will now know which game it is that they are playing. If it turns out that they are playing hard, the strategy calls for them both to play DT in the remaining $k - 1$ steps. If it turns out they are playing soft, the strategy is that they both play T in the remaining $k - 1$ steps. (And if either player deviates at any point from the prescribed behavior, then the two players switch immediately to DT and stick to it until the end of the game.[19])

This strategy is (Nash) equilibrium.[20] The optimal k is calculable: it depends only on the probability ϵ (that the game is soft) and on the given payoffs, but not on the horizon of play.[21] Thus, for the particular payoffs we have been considering, and with a "pessimistic" probability ϵ of 0.1 that the game is in fact soft, it turns out that it is optimal for both players to play T up until something like ten rounds before the end. In other words, even if the players are "pessimistic" and believe with high probability that they are, in fact, facing hard, then when N is sufficiently large the mere possibility of soft may produce the desired bootstrapping which will block the backward induction and "pull" toward trust, for most of the duration of play. What makes a better equilibrium than DT possible here is the fact that knowledge is possible: at any stage it is possible for the players to find out what game they are playing (and also, finding out at some stage promises the players better payoffs than playing DT throughout). At the same time, interestingly, the Nash equilibrium that yields the best payoffs for both players is the one that postpones the moment of truth as far as possible.

Note that, although the setup we consider is symmetric between the two players, the possibility of knowledge, and hence the Nash equilibria, are essentially predicated on an asymmetric step: he continues with T, she deviates to DT. (There is, to be sure, at least one symmetric equilibrium point, namely, the least efficient one where they both play DT throughout.) Of course, the roles of the players can be reversed, or assigned to a commonly observed flipped coin, but this is still an asymmetric solution. This circumstance may be

[19] This requires the tiny caveat that, if the deviating partner is she, in that she neglects to switch to DT when she is supposed to, then she is to switch to DT at the next step.

[20] This is, in fact, not one strategy but a family of strategies—for each choice of k. These strategies are Nash equilibria, provided k is not too small (that is, if it is not too close to the end of play). The smallest k for which this is a Nash equilibrium yields the highest payoffs overall for the players.

[21] See Osborne and Rubinstein, p. 156.

taken to make a case for the importance of certain social conventions, like chivalry (he sacrifices; she deviates) or deference to old age, and so on which may signal the direction of the asymmetry and thus facilitate a solution.

(ii) "Perturbation": one-sided information. Let us consider now a setup where knowledge is not possible: the payoffs in each round are not accessible to the players.

Once again, one of our two games (or states of nature) is picked by nature, and played N times; soft is played with a small probability epsilon (ϵ). There is a certain probability p (or rather $(p - \tau)$, as explained below) that both players are informed of the game they are playing, and a certain probability $(1 - p)$ that both players are not informed. There is, however, also a small probability theta (τ) that only one of them is informed. (We are to suppose that the uninformed remain uninformed until the end.)

Suppose first that he is uninformed. Now, if he observes that she chooses T in the first round, he might infer that she probably knows which game is being played and that she believes he knows as well. He might further infer that she must know that they are playing soft, and that this is why she is offering cooperation by choosing trust. But then he should also choose T from that point on (up until close to the end of play). Suppose next that he is informed. If his information is that the game they are playing is soft and he observes that she chooses trust, then the previous reasoning applies and he should choose trust as well. Consider, finally, the case when he is informed, and the information is that the game they are playing is hard. In this setup, if he observes that she chooses T in the first round, he will infer that she is probably uninformed and that, if he were to reciprocate with T, she might infer that he knows that they are playing soft and she will therefore continue to cooperate. Now, the entire chain of reasoning in this passage can, of course, be replicated for her by starting out supposing that she is uninformed and that she observes that he chooses trust in the first round.

So it is, in fact, in the interest of the uninformed to mimic the informed—or to mimic anyway on the belief that the other might be informed. And it is also in the interest of the informed to choose trust even when he or she is informed that they are playing hard, given that it is common knowledge between them that it is in the interest of the other to mimic. This will eventually "pull" toward coordination on mutual trust (almost to the end) even among the uninformed and

even when the game picked by nature is in fact hard.[22] Note that, if it is common knowledge between them that they are both uninformed, they will just look at the expected payoffs. Given our assumption that ϵ is small, these look pretty much like the payoffs for hard, and so the trustful equilibrium will be destroyed and both players will distrust (DT). The point is that the "pull" toward coordination on trust will occur even when the players are in fact uninformed, provided that there is some uncertainty in the mind of each, however small, as to whether the opponent is informed or not.[23]

There is a theoretical result in the game-theoretical literature[24] that amounts, roughly, to the following. In the finitely repeated PD, a one-in-a-thousand chance that one's opponent is irrationally benevolent (in the sense that she is determined to stick to the strategy of tit-for-tat) completely changes the theoretical prediction that there will be an unstoppable unraveling from the back toward noncooperation from round 1. That is, when a tiny "perturbation" (or a "trembling hand") is introduced in the form of incomplete information about the rationality of player 1, then player 2 can know without inconsistency what player 1 will do because, with small probability, player 1—being irrationally benevolent—will permit this to be known. And "from this small wedge, we are able to get them both to play [the cooperative strategy] happily until near the end of the game."[25] Moreover, even if player 2 is not "irrational" in this fashion,

[22] Note: there is always a Nash equilibrium in which neither the informed nor the uninformed ever trust the opponent. But the point here is that there cannot be an equilibrium in which the informed cooperate by choosing trust whenever possible and the uninformed never choose trust. If the informed cooperate, then it is also in the interest of the uninformed to mimic the informed, and this breaks the proposed equilibrium. Therefore, any other equilibrium besides the one of full distrust must involve mimicry, which "pulls" both informed and uninformed toward trust.

[23] The two examples of finitely repeated games here discussed are special cases of a general result of Benoit-Krishna: J.-P. Benoit and V. Krishna, "Nash Equilibria of Finitely Repeated Games," *International Journal of Game Theory*, XVI (1987): 197-204. See also the related discussion in Osborne and Rubinstein, pp. 155-60.

[24] Kreps, P. Milgrom, J. Roberts, and R. Wilson, "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma," *Journal of Economic Theory*, XXVII (1982): 245-52.

[25] Kreps, p. 541. Indeed, as it turns out, even minimal violations of the hypothesis of common knowledge are sufficient to justify the emergence of cooperation in the repeated PD. See Abraham Neyman, "Cooperation in Repeated Games When the Number of Stages Is Not Commonly Known," *Econometrica*, LXVI, 1 (1999): 45-65, and also Aumann, "Irrationality in Game Theory," in P. Dasgupta, D. Gale, O. Hart, and E. Maskin, eds., *Economic Analysis of Markets and Games: Essays in Honor of Frank Hahn* (Cambridge: MIT, 1992), pp. 214-27 (reprinted in Aumann, *Collected Papers*, Volume 1 (Cambridge: MIT, 2000), pp. 621-34).

the "rational" thing for player 1 to do is to mimic this sort of behavior, to keep cooperation alive.

The distrust case just discussed, where there is a tiny element of uncertainty about asymmetric information, is structurally similar to this case of the finitely repeated PD-game, where rational cooperation is proven possible given a tiny perturbation of irrationality. The crucial difference, however, between the two is that, in our distrust case, the perturbation has nothing to do with irrationality.[26] In our distrust case, it enters through the narrow crack of the possibility that a different game is being played and that there is asymmetrical information between the players: that only one of them is informed which of the two games is being played. As we saw this possibility, even if tiny, is capable of sustaining an equilibrium strategy of trust almost to the end of play.[27]

V. THE POSSIBILITY OF TRUST

In a society where there is war of all against all and people expect the worst of each other, people may come to anticipate that the types of interactions in which they will find themselves are PD-structured situations, and will act accordingly. In such a society, people will lock themselves forever in mutual suspicion and distrust. From this kind of distrusting animal, no trusting, promise-fulfilling creatures can hope to emerge. The pattern of behavior they adopt is self-reinforcing. In such a society, the possibility of being exposed to trust behavior and its benefits, and thereby to learning trust behavior, is effectively blocked.

[26] Pettit and Sugden show that the backward induction need not unravel from the back in the finitely repeated PD-game for a different reason than the trembling-hand perturbation. Indeed, they argue a stronger point, which is that "rational players are necessarily not in a position to run the backward induction argument" (*op. cit.*, p. 182), since (1) any act of cooperation by either player would cause the breakdown of an assumption that is required for them to run the induction—namely, their common knowledge/belief in each other's rationality, and (2) "there are beliefs which a player might rationally hold which would make it rational for him to cooperate initially" (*op. cit.*, p. 179). For a similar argument see Bicchieri. But this solution, while not relying on a "perturbation," still relies on irrationality (in the form of a breakdown in the common knowledge, or belief, in the rationality of the players).

[27] In the literature, the effect of such tiny perturbations of incomplete information is discussed also in connection with the other well-known games that exhibit a finite-horizon paradox: the centipede game and the chain-store game. It is also tied with the phenomenon of *reputation*: I may find it beneficial to sometimes act "irrationally" in order to build up a reputation, if being thought irrational might later be beneficial to me through the influence this has on how others play. See Kreps, pp. 536-43.

This pessimistic scenario can be avoided without resorting to Hobbes's solution, deus ex machina-style, of the external Sovereign. Nor need the assumptions about human nature be made more charitable. We may, without changing too much in Hobbes's initial assumptions, stick to his unflattering picture of human beings as basically egotistical creatures, and yet see a way for these creatures to have the resources for arriving at mutual trust. All they need is to realize that an alternative setup for distrust exists: to recognize that not all distrust situations are PD-structured. With this, there will emerge the recognition that, depending on the type of situation, "distrust" is ambiguous between a harsh, exploitative sense and a milder sense. The notion of egoism, too, will be seen to be open to an "isolationist," defensive interpretation which underlies a situation of soft distrust, in addition to an offensive interpretation, which underlies situations of hard distrust.

But the main point is that in order for trust to get going, no assumption need be made that, as a matter of fact, the state of nature is *not* Hobbes's PD-structured one. All that is required is that the possibility be acknowledged that it *may not* be so structured. The mere ambiguity between hard and soft distrust is all that is needed for trust to emerge.

EDNA ULLMANN-MARGALIT
Center for Rationality and Interactive Decision Theory/Hebrew University of Jerusalem