IESE Business School

University of Navarra

Occasional Paper

OP no 03/3

September, 2002

ALTRUISTS' CHANCES IN THE PRISONER'S DILEMMA

Miguel Soler *

* Professor of Finance, IESE

IESE Occasional Papers seek to present topics of general interest to a wide audience. Unlike Research Papers, they are not intended to provide original contributions in the field of business knowledge.

IESE Business School - Universidad de Navarra
Avda. Pearson, 21 - 08034 Barcelona. Tel.: (+34) 93 253 42 00 Fax: (+34) 93 253 43 43
Camino del Cerro del Águila, 3 (Ctra. de Castilla, km. 5,180) - 28023 Madrid. Tel.: (+34) 91 357 08 09 Fax: (+34) 91 357 29 13

ALTRUISTS' CHANCES IN THE PRISONER'S PRISONER'S DILEMMA

Abstract:

Unlike in the Iterated Prisoner's Dilemma game, where various different strategies can be considered to enforce cooperation, including the famed Tit-for-tat strategy, in the one-shot case cooperation is said to have no chance. The author of this paper argues that the non-iterated case should be treated as a game between different types of players and that, even if we assume that the players do not have any information about each other's actual decision, a "discriminating altruist" (i.e. a player who decides to cooperate or not depending on her estimate –naturally subject to error– as to what the other player is likely to do) could adapt her strategy to her estimate of the decision likely to be taken by the other player. The opinion of experts from seven different branches of the social sciences is cited to confirm the existence of such a discriminating capacity in human players.

The paper shows how, in a population made up exclusively of discriminating altruists and "rational" egoists, the former (seeking to cooperate only with those they estimate to be discriminating altruists) can have a better expected value than the latter (who defect unconditionally) if the % of error in estimating the nature of the other player is small enough for the circumstances. This proves that the immoral recommendation of economic "rationality" is not borne out, as claimed, even by the invincibility of the dominant egoistic decision. The paper concludes by proposing a more general moral strategy, based on not taking into account the possibility of exploiting the other player when weighing up the alternatives.

Keywords: prisoner's dilemma, rationality, self-interest, altruism.

ALTRUISTS' CHANCES IN THE PRISONER'S DILEMMA

1. The prisoner's dilemma game: A problem with an inadequate solution?

"The police have arrested two foreign journalists on charges of spreading defamatory stories about the country. The prosecutor offers each prisoner the following deal:

- The full penalty for the crime is three years (-3);
- If both confess, each will receive a reduction of one year, giving a total of two years in prison (-2);
- If no evidence is obtained, both will be charged and convicted for a minor offence, each being sentenced to only one year in jail (-1);
- If one confesses and the other does not, the one that confesses will be released immediately (0), whereas the other will suffer the full penalty for the main crime."

It is quite surprising that what people (but not animals!) actually do in *Prisoner's Dilemma* (abbreviated "PD") type situations is exactly the opposite of what Game Theory, a branch of mathematics most commonly used by economists, recommends: "Reasoning leads to predicting 'defect (confess) all' as the outcome of the game, in stark contrast to numerous experimental studies in which players are consistently observed playing cooperative strategies in Prisoner's Dilemma type games" (Eichberger 1993, p.206). The recommended defection is not only suboptimal (if all defect, they obtain worse results than if all cooperate, i.e. do not confess) but also clearly unethical (according to the Golden Rule, we should cooperate since we want others to cooperate with us).

In (Thaler 1992) we are told, "A similar situation applies to what economists call *public goods* (those that, like public TV, once provided, you cannot prevent people from freely using) [...] Again, economic theory predicts that when confronted with public goods, people will "free ride" (not contribute) [...] The predictions derived from this assumption of rational selfishness are, however, violated in many familiar contexts."

Many mathematicians embarked, in the 1960s, on a frantic search for a way out of such a paradoxical situation. A solution was found, in the late 1970s, for the *Iterated Prisoner's Dilemma*, where a PD game is played repeatedly, an indefinite number of times, in such a way that each player's strategy can vary in light of the other's conduct in previous rounds. In his famous computer tournament of strategies, Robert Axelrod (1984) found that the best strategy was the very simple *Tit-for-tat*:

– C*ooperate* in the first move;

– Then *reciprocate* the other's conduct in each subsequent move (meet cooperation with cooperation and defection with defection).

However, the non-iterated game remains unsolved: suboptimal defecting continues to be the decision recommended by the theory. In this paper I will try to show that a theoretical justification for cooperating can be found, even without iteration, if it is merely allowed that the players have certain abilities common to real human beings.

The story translates into a **payoff matrix** (payoff Player 1 \ payoff Player 2):

		Player 2	
		C	D
Player 1	C	-1 \ -1	-3 \ 0
	D	0 \ -3	-2 \ -2

Where "C" and "D" stand for "cooperates (does not confess)" and "defects (confesses)".

What makes this a Prisoners' Dilemma type game is the order of the payoffs (calling "(C, D)" the payoff for "cooperating while the other defects", and so on):

$$(C, D) < (D, D) < (C, C) < (D, C)$$

In my example: $-3 < -2 < -1 < 0$

Other usually required conditions are:

– Each player has full information about: who her opponent is, the strategies available to her, and her payoffs, but

– She is completely ignorant about the other's actual decision;

– There will not be any possibility of *retaliation* in the future (subsequently, this is not a *cooperation game* because there is no way to enforce an agreement).

I will make the following additional assumptions about the players:

– They are *human* beings who, as we will see,

– Can *predict,* subject to some error, each other's behavior.

We will discuss whether applying that human faculty affects the Prisoner's Dilemma's condition that the players do not actually know each other's decision.

2. The paradoxical "rational" solution

To solve the PD, Game Theory applies the economic criterion that players are *Homo economicus*, a species that behave "*rationally*" (in quotation marks, because this is a special sense), trying, in the case of a game, to maximize their payoff regardless of the nature of the other player.

We can analyze a game by applying the criterin of the ***best response,*** selecting the strategy that gives the best payoff for each strategy adopted by the other player. In our game:

I- The other player either will cooperate or will not cooperate (defect) –a logical truth;

II- If the other player defects, the best I can do is to defect (2 years instead of 3);

III- If the other player cooperates, the best I can do is to defect (0 years instead of 1);

IV- So, in any case, the best I can do is to defect.

Each player has a ***dominant*** *strategy*, one that is best regardless of the strategy employed by the other. As the situation is *symmetric* (the same for both players), two "rationals" will decide the same: to defect, each getting 2 years in prison.

There is no *dilemma* (you know which of the two solutions is best), but there is a *paradox* (the logical solution seems contrary to common sense): in real life people behave "irrationally" and obtain a better result. It also seems contrary to common sense that the decision should not depend either on the value of the payoffs (consider the case, 0 < 1 day < 2 years < 2 years + 1 day : would you close the doors to the mutual benefit of cooperation for the sake of not risking 1 day?) or on the trustfulness of the other player (would you defect in front of your best friend?).

Actually, being a "rational" is a way of being conservative, as may be seen if we use an alternative criterion, the ***maximin value*** (also called *security level*): the decision is taken in order to "minimize your maximum loss". In the PD, you should decide to defect in order to avoid (C, D) if the other defects, and equally to avoid (C, C) if the other cooperates. In fact, it is a theorem of Game Theory (Eichberger 1993, Lemma 3.1) that: every *dominant strategy* is a *maximin strategy*.

I will depart from the interpretation, stated in (Danielson 1992, p. 14), that: "*Strong game theory* makes the following recursive assumptions:

i) All agents are identical –that is, symmetrically rational– and

ii) all agents are fully and freely informed about assumptions i) and ii)";

because, even allowing that two "rational" players cannot decide otherwise than to defect both, the Prisoner's Dilemma still raises the question, as yet unsolved, of whether a "rational" is the best possible player in the single PD, and this cannot be decided unless other players, with other strategies, are admitted to the game. In this view, **"solving" the Prisoner's Dilemma means finding players with an alternative strategy capable of obtaining a better payoff than a "rational"**, rather than finding fault with the "rational" reasoning.

The "rational" solution is a *Nash* ***equilibrium***: none of the players on her own has any incentive to change her decision. But this solution is not a *Pareto* ***optimality***: it is possible to have a better payoff for one without worsening the situation of the other (in fact, bettering hers, too). That the optimality does not coincide with the equilibrium is a mathematical curiosity that the PD shares with other games.

As the solution to "defect both" is an equilibrium, the players will not change their decision to "cooperate both" (the optimum) unless some additional constraint forces them to do so. As the conditions of the Prisoner's Dilemma do not allow external constraints, the only possibility is that the prisoners *constrain themselves*. The people who, as observed, cooperate in PD type situations must evidently impose upon themselves the obligation to cooperate. An obligation that one imposes upon oneself voluntarily is a moral duty.

3. The moral solution and its weaknesses

To be moral, you should take into account the interest of the other (*altruism*), regardless, many say, of your own interest (*pure altruism*). A "rational" is an *egoist* because she considers only her own interest.

A pure altruist will cooperate because this is "the best for the other". The most commonly used moral approach is to apply the *Golden Rule* – positive (or negative):

"Do (not do) unto others as you would have them (not) do unto you"

I should cooperate because that is "what I want the other to do unto me". Again we have the optimum (C, C) if both apply the rule. The rule is supposed to advise pure altruism, which many say is the only really moral conduct.

Two pure altruists obtain a better result than two "rational" egoists, but a pure altruist will be *exploited* and will "lose" when faced with an egoist; so, from Game Theory's point of view, egoism is the winner of the match.

Anyway, in a "population" (a group of players) composed of two subgroups (each consistently pursuing the same strategy, so we can identify them by the strategy they follow), one of "rational" egoists ("RE") and the other of pure altruists ("PA"), can the benefit the altruists obtain (when paired) compensate for the loss they suffer (when exploited by the egoists)? Let us see what happens in a population consisting of 80% PA and 20% RE (we should apply **Bayesian Decision Theory** and compute their *expected utility* –the expected utility of one player in one single Prisoner's Dilemma type match, with given probabilities of players being of a certain type, is *equivalent to the mean value* to be obtained by this player in successive independent matches[1] with all the other members of the population, with frequencies equal to the mentioned probabilities):

$$\text{expected utility PA} = 0.8(-1) + 0.2(-3) = -1.4$$

$$\text{expected utility RE} = 0.8(0) + 0.2(-2) = -0.4$$

A RE player obtains a better result on the average, but is the RE strategy really better than that of the PA? It is far from crystal-clear: the PA, by cooperating inside such a

population, obtains a better payoff (-1.4) than if all were to defect (-2): the PA is a better player if you are interested only in approaching the optimum and are not worried about being exploited. Sometimes, though, the exploited strategy loses in the strong biological sense of becoming extinct.

In addition, pure altruism can yield an incorrect solution in some games, suggesting that it commits a similar logical error to the "rational" solution.

4. The (pure) altruist's dilemma: reviewing the golden rule

Consider the game with the following payoff matrix (C and D for "cooperates" and "defects"; positive payoffs)[2i]:

		Player 2 C	Player 2 D
Player 1	C	3\3	4\1
	D	1\4	2\2

It is an "Adam Smith situation", in which an "invisible hand" (the payoff matrix) guides the players, who pursue their own interest, to do what is the common good (which is something that does not happen in the Prisoner's Dilemma).

The Golden Rule would advise:

– I should do the same as I want the other to do unto me;

– the best that the other can do unto me, in this game, is to defect;

– subsequently, I should defect too.

Two moral persons, applying the standard Golden Rule, would defect both, against the common good. With such inverted preferences (each prefers the best for the other) the Dilemma, though "objectively" Adam Smith, "subjectively" is a PD, and the altruists face the same impossibility to reach the optimum. The only possible logical error could be that, if one allows that the two agents know that they are identical, then the players should know that they are going to make the same decision (due to the symmetry, only (C, C) and (D, D) are really possible), which implies that each player's decision is related, in this sense, to the other's; and if you know that the other is going to do the same as you, wouldn't it be better to cooperate?

The solution would be to apply Kant's Categorical Imperative and take a "universally valid" decision, valid if taken in the same way by everybody. Where the Golden Rule falls down is in taking into account the effect of the other's conduct "unto me" instead of unto the common good. By correcting the rule to suppress that subjective view we would obtain a **Categorical Golden Rule**:

"Do (not do) unto others what you would have everybody (not) do"

Thus modified, the rule can give the correct advice: as you want everybody to cooperate (this yields a better result than if everybody defects), you too have to cooperate. As you would like what you want everybody to do to be a universal law, the modified Rule is obviously another version of Kant's Imperative. It is imperative not because someone orders you to do so, but because "the (moral) way for everybody to do what is best for all is for everyone to oblige herself to do so".

Could the "rational" also modify their rule for the PD, to make it valid when both take the same decision? They should decide to "take the decision that gives the best reward when taken by both", but cooperating ((C, C) is the best solution to be taken by both players symmetrically) will not comply with the equivalent maximin criterion, and so they will no longer be "rational". They are trapped in (D, D) even if they know that both are going to decide the same.

Returning to altruism, the Categorical Golden Rule no longer has to prescribe pure altruism, because as this extreme form of altruism is detrimental to all types of altruists (as we will see), you may well not want everybody to follow the rule of pure altruism, but rather the more appropriate rule of reciprocal altruism.

5. Reciprocal altruism as the solution to the dilemmas

Contrary to the Iterated PD, where a great number of different strategies are possible (Axelrod 1984, chapter 2), in the one-shot case only three strategies are feasible:

– unconditional defection (REs),
– unconditional cooperation (PAs) and
– conditional cooperation

Conditional cooperation, under PD assumptions, can only be based on an estimate about the other's decision. We need a strategy that, at the same time, provides cooperation and avoids exploitation: if we could guess what the other is going to do, the "intelligent" strategy would be to do the same. Like a "**reciprocal altruist**", one who tries to cooperate with a cooperator and to defect when faced with a defector.

But, if the reciprocal altruist approves of defecting when the other defects, **why** not defect also when faced with a cooperator? Because that would be **unjust**: suppose that the other prisoner has declared before and you know that she has not confessed, what would you do?

If you find it unjust to defect when you know that the other has cooperated, you should not accept reasoning III in the PD either: "If the other cooperates, the best I can do is to defect". If you want **to be just**, you should say, instead: "**If the other cooperates, I wish to cooperate too**". (There are some exceptions: in a *duopoly*, cooperation is not only not morally good, but also illegal).

Economists will say that "not trying to defect because that would be unjust" is a moral reasoning and that they merely seek to determine what is the best rationally, not morally. It can be argued, however, that it is not true that economists ignore moral constraints: they accept the constraint "do not rob"; so they should also accept the constraint

"reciprocate cooperation", because if one player cooperates, she makes the benefits of cooperation, (C, C) – (D, D), available to both. If the other defects, she obtains the additional benefit (D, C) – (C, C), but at the cost of depriving ("robbing") the former of her benefit. The economists' answer is that cooperation must be enforced by some authority, but why ignore the human capacity to oblige oneself?

Anyway, as it is a condition of the PD that it is not possible to know the other person's decision in advance, how can it be possible to be a reciprocal altruist in the PD? The answer is that not actually knowing what the other will do is one thing, whereas not having at least a rough idea what the other is likely to do is quite another.

6. The human capacity to predict the other's conduct

The capacity to predict the opponent's behavior is allowed by game theorists in many games, although it is deemed unnecessary in the PD. Game theorist Eichberger says (1993, pp. 67, 83 and 84): "A player with a dominant strategy need not speculate about the behavior of opponents [in the prisoner's dilemma] [...] In general, however, what is the best response for one player will depend on what the others do, that is, best responses will vary with the opponents' behavior [...] in other words, a player must predict correctly the behavior of opponents."

If you have a dominant strategy it seems obvious that you should not be interested in predicting the other's decision. I will try to show, however, that the ability to predict the opponent's behavior can cause the purported dominant strategy to be no longer the winning one. But let us see first how different sciences confirm such ability.

Anthropologist Robert Foley (1995, p.167[3]) tells us that the need to make such predictions was just one of the causes of the development of the human brain (emphasis added in this and later quotations): "The complexity of social relations has spurred the *evolution of the brain*: it is difficult *to predict the conduct* of another individual, especially if such conduct is based on her own predictions".

Primatologist Sarah T. Boysen (1996, pp. 180-1) has done experiments showing that the ability to deceive seems to be exclusive to humans: "In light of the animals' success and demonstrated flexibility with number concepts, we sought to explore the possible use of deception by chimpanzees [...] Both animals' *inability* to acquire what could be framed as a fairly straightforward *discrimination* problem nearly defied credulity".

Sociobiologist Edward O. Wilson (1980, p. 277) also stresses this view: "*Deception* and hypocrisy are [...] *very human* devices for conducting the complex daily business of social life [...] The all primate frankness would destroy the delicate fabric of social life that has built up in human populations". Deception is needed precisely because of the human ability to predict each other's conduct.

Psychologist Thomas Suddendorf (1999, pp. 234-40) explains how predicting behavior is part of a more fundamental human capacity and, when it appears, "During the fourth year children seem to change dramatically in the way they see the world, others and themselves. Parents observe that their children begin to make their own plans, [...] *consider other people's minds*, deceive and lie, [...] I propose that the child has developed a 'metamind'.

"[...] The concept of 'metamind' embraces the reflective self-reference of 'inner eye' (an introspection organ), the social and abstract components of 'theory of mind' (a *mind-reading* organ) [...] 'Theory of mind' refers to the explanation and *prediction of behavior* based on the attribution of mental states such as intention, knowledge or belief. [...] Others' beliefs, intentions, and knowledge are great *predictors of their behavior*, and *cooperation* as well as deception become far more effective if one knows what is on the other's mind."

Philosopher Robert M. Gordon (1996, pp. 167-8) tells us: "By replicating the facial expressions of others, we would tend to 'catch' the emotions of others (Meltzoff and Gopnik 1993), a process I call facial *empathy* [...] can assist us in interpreting, *predicting*, and explaining behavior". Facial empathy allows humans to read the other person's mind and predict her decision, just by seeing her face.

There is a new field called "experimental economics" that sets out to confirm people's behavior in real life. Has the capacity to predict the other's conduct been experimentally confirmed for the PD? "Robert Frank (1988), an economist, set up an experiment to find out. He put a group of strangers in a room together for just half an hour, and asked them each to predict privately which of their fellow subjects would cooperate and which would defect in a *single prisoner's dilemma* game. They proved substantially better than chance at doing so. They could tell, even after just thirty minutes' acquaintance, enough about somebody to *predict his cooperativeness*" (Ridley 1996, p. 82).

Let me, then, introduce a "**discriminating altruist**" (**"DA"**), a reciprocal altruist who uses her human capacity to discriminate (with some % of error, but without needing to know the other's actual decision), defecting with egoists and cooperating with altruists. Notice that there is an essential difference between my view and the *contractarian* one (a modern version of which has been presented, philosophically, by David P. Gauthier (1986a) and developed into games of artificial morality by Peter Danielson (1992, chapter 4)): they assume that their conditional players know whether the other is going to cooperate or not in order to do the same. My solution using discriminating agents respects the players' ignorance about each other's decision and may, therefore, be considered a solution to the PD.

If the DA were simply more just than the "rational" egoist, it would not be a better solution from Game Theory's point of view. But the DA obtains better results when confronted with another discriminator and tries to avoid being exploited by egoists. Can her success in the former case compensate for her occasional failures in the latter?

7. And the winner is... (but, what does it means to be a winner?)

According to Bayesian decision theory, a strategy is the *winner*, in a confrontation with others, if its expected utility is the greatest. If one strategy always wins or draws when the circumstances (payoffs, % of players, % error) are changed, we may call it the *absolute winner* for the game. (I will take the population view for a better understanding, but the conclusions obtained from populations are directly applicable to the single match PD, because of the abovementioned equivalence between expected and mean values).

Is "rational" egoism an absolute winner, in this sense, when contronted with pure altruism? Yes, because it always obtains a better result:

– confronted with a RE: RE obtains –2 and PA –3
– confronted with a PA: RE obtains 0 and PA –1

and this will not change with payoffs, as always (D, D) < (C, D) and (D, C) < (C, C).

And what about a RE matched against a DA? Is there a maximum error below which discriminating can be a better solution? We need to look at the results under different circumstances. (To simplify, I assume that the error in confrontation with a DA is the same as with a RE. This is not the case in real life, where human egoists use their machiavellian capacity to deceive; that can change the %, but not the conclusions).

Below is a matrix showing the most relevant points for the results of matches between DA and RE in our usual PD[4] (Additional values can be easily obtained through lineal interpolation; 0% and 100% have the meaning of confronting one single member of the other class):

Expected utilities **DA** / **RE** (= , if draw)

\ % DA **% error**	100	50	0
0	-1 / -2	-1.5 / -2	**-2** / =
25	-1.25 / -1.6	**-1.75** / =	*-2.25* / -2
50	-1.5 / -1.0	-2 / -1.5	*-2.5* / -2

It appears, from the above figures, that there is no absolute winner between RE and DA in the one-shot Prisoner's Dilemma. One or the other wins depending on the circumstances. The most interesting points, or combinations, are those (in bold) where there is a draw; they are the "frontier" between the two potential winners.

Actually, it can be seen from the formulas that, in any PD, with 0% error the absolute winner will always be DA, and with 50% error[5] RE: it must be a frontier in between 0% and 50% of error, only at such unrealistic extremes can there be an absolute winner.

Some conclusions:

– The **"rational" egoist's *dominant strategy* is not an absolute winner**, in the PD, when confronted with a player with the capacity to discriminate (a human being, for example).

- DA wins if she is, in discriminating, smart enough for the circumstances[6].

  How high her % error can be depends on:

  - The % DA in the population (the admissible error increases with the increase in the % DA);
  - the payoff matrix (as the expected utilities vary with it). It can be seen that the lower the penalty for being exploited (or the lower the premium for exploiting others), the higher the % error for DA still winning.

- In the presence of altruists, all players can obtain a better payoff than when all defect (-2), due to the fact that altruism opens the door to the benefits of cooperation: in this sense, DA is a better strategy than RE (but so is PA), except for very low % DA and high % error, which are the only circumstances in which cooperation is irrational because it yields worse results (in italics) than when all defect.

What goes wrong with the "rational" reasoning about the dominant strategy when DA wins? The reasoning does not fail, in a sense, as it provides good advice in order not to lose any match: a RE never loses any match with a DA. But it is also true that two DAs obtain a better payoff when paired than a pair of REs, thus compensating for their occasional failures when confronted with a RE. **The "rational" reasoning falls short in not considering the better result that other players can obtain between them**.

Although it cannot be said that the discriminating prisoner knows the other's actual decision (otherwise there would be no error), it may be alleged that discriminating with error is a middle way between perfect discrimination (which has the same effect as knowing the other's decision) and total ignorance about the other's decision, as required for the logical validity of the "rational" reasoning. And it is true that if it were required, in order to have a PD, that the players should not know at all each other at all, then discrimination would be impossible and RE would be an absolute winner. But that would be an unusual restriction that does not appear in other games, or in situations usually considered to be of the PD type. And if (in order to have a strict PD case) discrimination was not allowed, then the prisoner's dilemma would be a dilemma for... chimpanzees, not for humans.

8. A moral strategy for real life

A problem with the discriminating altruist is that she bases her decision on a single estimate (about whether the other is a DA or not), without making any use of the information actually available about the game's payoffs, which has a great influence in deciding who wins, as we may see with the following example. We shall compare the % error required to draw in the basic case and in two variations (always with % DA = 75% and the same other usual values; 0.003 is about 1 day):

basic case:	(C, D) = -3	error to draw = 30.0%
reduced penalty:	(C, D) = -2.003	error to draw = 41.4%
increased penalty:	(C, D) = -6	error to draw = 14.5%

The difference in allowed error is enormous: while in the reduced penalty case you still win with an error of 40%[7], you cannot afford to fail even 15% in the increased penalty case. How could we take account of this information in our decision?

One might be tempted to say that the solution is to decide to act either as a RE or as a DA according to who is to be expected to win, considering the matrix of expected values, given the estimated % of population and % of error. But that would be an inconsistent form of behavior: you would be neither a RE nor a DA, neither just nor unjust. In fact, you would be seen as an opportunist; people would discriminate against you, defecting when confronted with you, in which case the best you can do would be to defect always. If you wish to receive cooperation from the discriminating altruists, you have to seek justice.

Nevertheless, although a DA will be just with perfect discrimination, cooperating with all cooperators, with error she may defect with many altruists. As in real life there is error, how can the DA avoid being unjust? Should we be pure altruists if we seek justice? It is illuminating to take a brief look at what happens when a mix of PAs and DAs (who between them cooperate, except when the DAs err) are confronted with REs. Let us recall that with 50% DA and 50% RE, there is a draw with 25% error; if we now replace half the egoists with pure altruists (50% DA, 25% RE and 25% PA), the results would be (with the same 25% error):

expected utility DA = -1,3125
expected utility RE = -1,25
expected utility PA = -1,875

RE wins this time! For a reciprocal altruist a pure altruist is worse than an egoist. The reason is that the RE exploits the PA, while the DA does not. The PA is the ideal of morality, but she favors egoism and puts cooperation in danger of extinction.

Experimental economics has shown that revenge is an essential element of human conduct, even if it implies a net cost for the punisher: "Most players prove very willing, and even eager, to impose fines on co-players who lag behind in their contributions. Everyone seems to anticipate this, and *even in a game of one round*, less defection occurs than usual. [...] A lot of players show great eagerness to punish defectors. Participants seem to experience a primal pleasure in getting even with free riders. They seem to be more interested in obtaining personal revenge than in increasing their overall economic performance" (Sigmund, Fehr and Nowak 2002, pp. 84-5, about revenge in Ultimatum and Public Goods games). Experiments show that we are genetically inclined to be DAs (in the PD, revenge consists of defecting with a presumed defector) rather than PAs; the fact that PAs endanger cooperation, as we have found, may explain why.

In real life we are confronted with a mix of individuals varying gradually from RE to PA. How, then, can we be neither unjust nor exploited? What a moral, discriminating player could do when confronted with many types of players and using all the available information is to make an estimate, directly, about the probability that the other player will cooperate; then decide according to the Bayesian recommendation to maximize her expected utility, but with one modification: a reciprocal altruist, seeking justice, must **revalue the payoff (D, C)** –defect while the other cooperates– and value it equal to (D, D)[8]; she should then "cooperate" if her expected utility, so modified, is greater than (D, D). In our example, assuming an estimated 75% probability of receiving cooperation:

expected utility from cooperating = 0.75 (-1) + 0.25 (-3) = -1.5

As this is better than (D, D) = -2, she should cooperate (without considering the unmodified expected utility from defecting = 0.75 (0) + 0.25 (-2) = -0.5). Notice that, as required, such "(D, C)-modified Bayesian decision" depends on both payoffs and probabilities.

If, with such reasoning, you decide to defect and the other cooperates, you should not feel morally guilty for having been unjust, because you have not been trying to obtain the unjust (D, C) but simply to avoid the risk of suffering (C, D). Anyway, to be consistent with her beliefs and to avoid the "moral risk" of receiving the unjust (D, C), a reciprocal altruist should, when possible, inform the other prisoner if she decides to defect. Only then will she deserve to be called a "fair player".

The previous section has shown that such a moral strategy can even win against the "rational" strategy if discrimination is good enough (in relation to the % of other moral players) and the penalty for being exploited is not too high. But it has also has shown that there may be cases where such a discriminator will cooperate because she expects a result better than (D, D), and will lose, because the RE obtains an even better result thanks to the DA's errors in discriminating. In such circumstances, if losing were really bad (because it endangered altruism), cooperation would be untenable.

9. A common ground between economy and ethics

We have seen that economic "rationality" is really rational only in "Adam Smith situations", where the payoff matrix induces players pursuing their self-interest to decide what also happens to be the common interest and the optimum; morality is of no help in such cases.

At the other end of this spectrum there are "ethical situations", where the agent cannot obtain any economically valuable profit: economic rules may not be of any help here and only ethical rules may apply. To save the life of an anonymous woman who is drowning but whom you will never see again requires pure altruism. The Golden Rule may apply here, "I would like others to do the same for me"; other circumstances may require a totally disinterested rule, such as "love thy neighbor" (although the benefits of generalized altruism are inevitable).

In between these two extremes there are "common ground situations" that have economic payoffs but that are not Adam Smith situations (the Prisoner's Dilemma being the prototype). In such situations, if the players follow their self-interest, the result is suboptimal. What, in such circumstances, can the players do to be able to take the decision that is the best for each one of them when considered jointly? As we have seen, the secret is to constrain their decisions with a criterion of justice, renouncing the attempt to appropriate more benefits than are earned by one's own effort by deceiving and exploiting others, and trying, instead, to cooperate with other agents who are also willing to cooperate; if, that is, it is not too risky, because one should, at the same time, try to escape exploitation, which discourages the common effort. Economics and ethics must work together on this common ground; they cannot afford to ignore each other.

Economists should rethink our recommendation: "rationality" is a simple strategy and that is an advantage, provided it yields good enough decisions. That is not the case in the Prisoner's Dilemma's type of problems. In such common ground situations we should introduce more sophisticated, more human and less immoral agents. I suggest that the economic agents should accept, for reasons of justice, the moral constraint "do not exploit others' cooperation", just as they accept the constraint "do not rob", and that a practical way to do so would be, when analyzing the option of defecting, not to take into account (as explained earlier) the benefits of a potential exploitation.

This is not a theoretical discussion: "If people are not rational maximizers of self-interest, then to teach them that such behavior would be logical is to corrupt them. Indeed, this is just what Robert Frank and many others have found: those students who have been taught the nostrums of neo-classical economics are more likely to defect than, for instance, astronomy students" (Ridley 1996, pp. 146-7). After millenniums of development of the altruism gene, we are now spreading the egoism *meme*. Economists are teaching people to avoid the Pareto optimality and take the path of the suboptimal Nash equilibrium: if being anti-ethical is not bad enough, I hope to have shown that, in many circumstances, it is not even the strategy that has the best expected utility.

Notes[9]

1 We can see these matches, quantum-mechanics style, as being played simultaneously in parallel universes, the one-shot game being the concrete PD in which the *multiverse* collapses for the observer.

2 From Schick (1998).

3 In the Spanish translation.

4 The general formulas (abridging: a = % DA/100, e = % error/100, CC = (C, C), etc.) are:
expected utility DA = a(CC(1-e)(1-e) + CD(1-e)e + DC·e(1-e) + DD·e·e) + (1-a)(DD(1-e) + CD·e)
expected utility RE = a(DD(1-e) + DC·e) + (1-a)DD
The formulas for the frontier points can be obtained by equaling the two expected values. With some algebra, it can be deduced:
Drawing values: $a = (DD - CD)\, e \,/\, \{(CC - DD) + 2\,(DD - CC)\, e + (CC - DC + DD - CD)\, e^2\}$

5 With 0% error one DA alone can draw; two on win; with 50% error, the % DA should be > 100% (impossible) to draw; all this for any payoff complying with the conditions of the PD.

6 A discriminating altruist could increase her chances of winning by cooperating only when she is reasonably sure to win (considering the values in the matrix); but to be absolutely sure to win she should "defect always". A DA must content herself with being an *almost* absolute winner if she wishes to have an expected utility *significantly* better than a RE (*fuzzy* concepts). To minimize one's own risk is a way to be an egoist.

7 It would be even smaller if we had not maintained the value (D, C). Taking (D, C) = – 0.003, 1 day, then the % error to draw = 49.95%! DA wins with any minimal capacity to discriminate, as common sense suggests.

8 Such a modification implies that the game is no longer, subjectively, a PD, but this does not preclude that, if such a player has a better expected utility than a RE, she is a winner in what, for a RE (the strategy to win), is a PD.

9 References

Axelrod, Robert. 1984. *The Evolution of Co-operation.* Penguin Books, England.

Boysen, Sarah T. 1996. "More is less: The elicitation of rule-governed resource distribution in chimpanzees", in Anne E. Russon, Kim A. Bard and Sue Taylor Parker, Eds. *Reaching into thought: the minds of great apes.* Cambridge University Press, Cambridge.

Danielson, Peter. 1992. *Artificial Morality: virtuous robots for virtual games*. Routledge. London-New York.

Eichberger, Jürgen. 1993. *Game Theory for Economists.* Academic Press, London.

Foley, Robert. 1995. *Humans before humanity.* Blackwell Publishers Ltd., UK. (Retranslated from the Spanish version: *Humanos antes de la Humanidad.* Edicions Bellaterra 2000, Barcelona).

Frank, Robert H. 1988. *Passions within Reason.* Norton, New York.

Gauthier, David P. 1986a. *Morals by Agreement*. Oxford University Press, Oxford.

Gordon, Robert M. 1996. "Sympathy, Simulation, and the Impartial Spectator", in Larry May, Marilyn Friedman, and Andy Clark, Eds. *Minds and Morals: essays on cognitive science and ethics.* A Bradford Book. The MIT Press, Cambridge-London.

Melzoff, Andrew and Gopnik, Alison. 1993. The role of imitation in understanding persons and developing a theory of mind. In S. Baron-Cohen, H. Tager-Flusberg and D.J. Cohen, Eds., *Understanding Other Minds.* Oxford University Press, Oxford.

Ridley, Matt. 1996. *The Origins of Virtue.* Penguin Books (1997), England, etc.

Schick, Frederic. 1997. *Making Choices. A Recasting of Decision Theory*. The Press Syndicate of the University of Cambridge, Cambridge.

Sigmund, Karl; Fehr, Ernst and Nowak, Martin A. 2002. *The Economics of Fair Play*. Scientific American, January 2002.

Suddendorf, Thomas 1999. "The rise of metamind", in Michel C. Corballis & Stephen E. G. Lea, Eds. *The Descent of Mind*. Oxford University Press Inc., New York.

Thaler, Richard H. 1992. *The Winner's Curse. Paradoxes and Anomalies of Economic Life.* Princeton University Press, Princeton, New Jersey.

Wilson, Edward O. 1980. *Sociobiology: The Abridged Edition.* The Belknap Press of Harvard University Press, Cambridge. □