Review of Social Economy

ISSN: 0034-6764 (Print) 1470-1162 (Online) Journal homepage: http://www.tandfonline.com/loi/rrse20

Cooperation and fairness: the flood–Dresher experiment revisited

Tom De Herdt

To cite this article: Tom De Herdt (2003) Cooperation and fairness: the flood–Dresher experiment revisited, Review of Social Economy, 61:2, 183-210, DOI: 10.1080/0034676032000098219

To link to this article: https://doi.org/10.1080/0034676032000098219

REVIEW OF SOCIAL ECONOMY, VOL. LXI, NO. 2, JUNE 2003

Routledge
Taylor & Francis Group

Cooperation and Fairness: The Flood–Dresher Experiment Revisited

Tom De Herdt
University of Antwerp
Tom.deherdt@ua.ac.be

Abstract In this paper we set out to deepen our understanding of the importance of fairness in decision-making within the context of Prisoners' Dilemma games. A review of the "historic" Flood–Dresher experiment provides a useful empirical basis, as it allows us to look in considerable detail at how the experimental players made up their minds. We try out several game-theoretical readings of the experimental results, and find some value in Adam Smith's age-old concept of rules of conduct. We find that fairness considerations are much more than mere excuses for taking a free ride or pointers to focal points. They seem to play a considerable role both at a conscious and at a less-than-conscious level.

Keywords: cooperation, fairness, prisoners' dilemma, rules of conduct

In a seminal article entitled Fairness as a Constraint on Profit Seeking, Kahneman, Knetsch and Thaler convincingly explain several market anomalies by introducing the assumption that economic behavior is influenced by standards of fairness—even in the absence of a long-run interest to apply such standards (Kahneman *et al.* 1986). The argument implies that individual behavior is merely indirectly related to outcomes. Social norms seem to mediate between behavior and outcomes. In what follows, we try to explain the anomalous outcome of a "historic" prisoners' dilemma experiment by applying the same assumption.

Experiments are appropriate methodological tools for considering in detail the articulation between behavior, local norm formation and outcomes. Once we question the assumption of economic rationality, studying real-life actors in an artificially controlled environment can be an interesting option for exploring alternative assumptions. We begin by reviewing a little (quasi-)experiment, conducted in January 1950 and known as the "Flood–Dresher experiment" (Flood

Review of Social Economy
ISSN 0034 6764 print/ISSN 1470–1162 online
http://www.tandf.co.uk/journals
DOI: 10.1080/0034676032000098219

1958; Luce and Raiffa 1967: 101; Axelrod 1984: 216; Poundstone 1992). As an experiment, it would not stand up to today's methodological requirements. However, it does provide an interesting basis for exploring the way in which behavior, incentives and rule-following are articulated. The argument is *not* that players do not act intentionally; it is rather that they become so obsessed with attaining a consensus over fair distribution that we can rightly question the "rationality" of their behavior.

After a discussion of the results of the experiment, we put forward some alternative theoretical ideas on the relationship between behavior, incentives and rules. According to some scholars, social norms are enforced by "informal rewards and punishment" (e.g. Lundberg and Pollack 1993). Others gave substance to this assertion by referring to the role of emotions in sustaining norms (Frank 1987; 1988; Elster 1989; 1999; Becker 1996; Platteau 1994). Most of these authors refer directly or indirectly to Adam Smith. Indeed, Smith's work, and particularly his *Theory of Moral Sentiments (TMS)*, is relevant in this context in that it provides a framework which allows us to go considerably beyond material interests. At the end of the first section, we first interpret the Flood–Dresher experiment by reviewing some contemporary game- theoretic interpretations. Subsequently, we develop our own interpretation by making extensive use of Adam Smith's concept of rules of conduct. The third section summarizes the main ideas of the paper.

THE FLOOD–DRESHER EXPERIMENT

In an attempt to verify the usefulness of Nash's identification of a (non-cooperative) Nash equilibrium, mathematicians Merril M. Flood and Melvin Dresher invited two friends, economist Armen Alchian (UCLA) and mathematician John Williams (RAND), to play the following game:

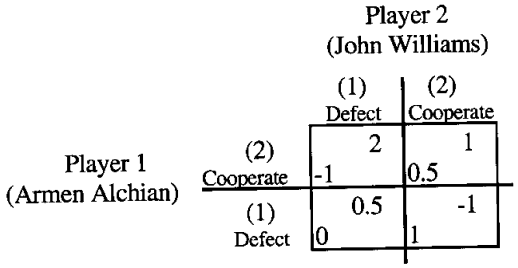

Figure 1: One-period game in Flood–Dresher experiment
Source: Poundstone (1992: 106)
Armen Alchian' payoffs (representing dollar cents) are in the southwestern corner, John Wiliam's payoff in the northeastern corner.

Armen Alchian's payoffs (representing dollar cents) are in the southwestern corner, while John Wiliams's payoffs are in the northeastern corner. Each player is summoned to maximize his payoff. But payoffs depend as much on own decisions as on the other's choices. Each player has to choose either (1) or (2) *in the absence of knowledge about the other's choice*. Thus, four different allocations are possible. If, for example, Alchian chooses (1), he gains either 0 or 1, depending on whether Williams plays (1) or (2) respectively.

This game had to be played exactly 100 times, and the players knew this in advance. The players could not communicate with one another, neither before nor during the game. But they were informed about each other's moves and about the resulting payoffs after each round. They were also asked to keep a log of personal comments during the game. The results and a copy of the log are presented in Table 1.

Note that the game structure resembles the classical prisoners' dilemma (PD), except for the asymmetry in the payoffs. In this perspective, strategy (2) corresponds to the "Cooperative" strategy, while (1) implies "Defection". In fact, mathematician Albert Tucker used a "dressed-up version" of this game in May 1950 during a lecture on game theory for psychology students. On this occasion, he invented the now world-widely known tale of the prisoners to clarify the dilemma inherent in the game (Rapoport and Chammah 1965: 24; Luce and Raiffa 1967: 94; Axelrod 1984: 216; Poundstone 1992:117). The Flood–Dresher experiment is hence a PD-experiment *avant la lettre*.

In what follows, we shall first discuss the results of the experiment. As the experimental subjects were asked to keep a log of running comments (written down after having decided on a strategy in a particular game, but before the other player's choice was revealed), there are clues as to the reasons for the (less and less frequent) breakdowns in cooperation. This will enable us, in a third and a fourth section, to interpret the players' intentions and strategies.

EXPERIMENTAL RESULTS

Game theory suggests two solutions to a game such as that presented above. First, "hard" rational choice predicts that, if the game shown in Figure 1 is played only once, it would be rational from players' individual points of view to play D, irrespective of the opponent's moves. The principle of backward induction further indicates that a definite number of iterations will not change the rational choice prediction that players should play D in each round.[1] Second, following David

1 For a critique of the backward-induction argument, see Elster (1989) and Dupuy (1992).

Table 1: The Flood–Dresher experiment: moves and comments[o]

Game	*Moves*		*Armen Alchian's comments*	*John Williams's comments*
	AA	*JW*		
1	D	C	JW will play D—sure win. Hence if I play C—I lose.	Hope he's bright.
2	D	C	What is he doing ?!!	He isn't but maybe he'll wise up
3	D	D	Trying mixed ?	Okay, dope.
4	D	D	Has he settled on D ?	Okay, dope.
5	C	D	Perverse!	It isn't the best of all possible worlds.
6	D	C	I'm sticking to D since he will mix for at least 4 more times.	Oh ho! Guess I'll have to give him another chance.
7	D	C		Cagey, ain't he ? Well...
8	D	D		In time he could learn, but not in ten moves so:
9	D	D	If I mix occasionally, he will switch—but why will he ever switch from D ?	
10	D	D	Prediction. He will stick with D until I change from D. I feel like DuPont.	I can guarantee myself a gain of 5, and guarantee that Player AA breaks even (at best). On the other hand, with nominal assistance from AA, I can transfer the guarantee of 5 to player AA and make 10 for myself, too. This means I have control of the game to a large extent, so player AA had better appreciate this and get on the bandwagon. With small amounts of money at stake, I would (as above) try (by using C) to coax AA into mutually profitable actions. With large amounts at stake I would play D until AA displayed some initiative

Table 1: (Continued)

Game	*Moves*		*Armen Alchian's comments*	*John Williams's comments*
	AA	*JW*		
				and a willingness to invest in his own future. One play of C by AA would change me from D to C, where I would remain until bitten. On the last play it would be conservative for me to switch to D, but I wouldn't do so if the evidence suggested that AA was a nice stable personality and not in critical need of just a little extra cash.
11	D	C		Probably learned by now.
12	C	C		I'll be damned ! But I'll try again.
13	C	C		That's better.
14	C	C		Ha!
15	C	C		(bliss)
16	D	C		
17	C	D		The stinker.
18	C	D		He's crazy. I'll teach him the hard way.
19	D	D	I'm completely confused. Is he trying to convey information to me ?	Let him suffer.
20	C	D		
21	C	C		Maybe he'll be a good boy now.
22	D	C		Always takes time to learn.
23	C	C		Time.
24	C	C		
25	C	C		
26	D	C		
27	C	D		Same old story.

Continued

Table 1: (Continued)

Game	*Moves*		*Armen Alchian's comments*	*John Williams's comments*
	AA	*JW*		
28	D	D	He wants more C's by me than I'm giving.	To hell with him.
29	C	D		
30	D	D		
31	D	C	Some start.	Once again.
32	C	C	JW is bent on sticking to C. He will not share at all as a price of getting me to stick to C.	—-, he learns slow!
33	C	C		On the beam again.
34	C	C		
35	C	C		
36	C	C		
37	C	C		
38	D	C		
39	C	D		The —-.
40	D	D		
41	D	C		Always try to be virtuous.
42	C	C		Old stuff.
43	C	C		
44	C	C		
45	C	C		
46	C	C		
47	C	C		
48	C	C		
49	D	C	He will not share	
50	C	D		He's a shady character and doesn't realize we are playing a 3rd party, not each other.
51	D	C		
52	C	C		He requires great virtue but he doesn't have it himself.
53	C	C		

Table 1: (Continued)

Game	*Moves*		*Armen Alchian's comments*	*John Williams's comments*
	AA	*JW*		
54	C	C		
55	C	C		
56	C	C		
57	C	C		
58	C	C	He will not share.	
59	C	C	He does not want to trick me. He is satisfied. I must teach him to share.	
60	D	C		A shiftless individual – opportunist, knave.
61	C	C		
62	C	C		Goodness me! Friendly!
63	C	C		
64	C	C		
65	C	C		
66	C	C		
67	D	C	He won't share.	
68	C	D	He'll punish for trying!	He can't stand success.
69	D	D		
70	D	D	I'll try once more to share-by taking.	
71	D	C		This is like toilet training a child – you have to be very patient.
72	C	C		
73	C	C		
74	C	C		
75	C	C		
76	C	C		
77	C	C		
78	C	C		
79	C	C		

Continued

Table 1: (Continued)

Game	*Moves*		*Armen Alchian's comments*	*John Williams's comments*
	AA	*JW*		
80	C	C		Well.
81	D	C		
82	C	D		He needs to be taught about that.
83	C	C		
84	C	C		
85	C	C		
86	C	C		
87	C	C		
88	C	C		
89	C	C		
90	C	C		
91	C	C	When will he switch as a last minute grab of D ? Can I beat him to it as late as possible ?	
92	C	C		Good.
93	C	C		
94	C	C		
95	C	C		
96	C	C		
97	C	C		
98	C	C		
99	D	C		
100	D	D		

º Comments precede the moves made on the same line.
Source: Poundstone 1992: 108–116.

Kreps' lead, one could argue that just a small amount of uncertainty about the other player's *real* interest would change this choice completely. Suppose that one player suspects the other to be (slightly) non-selfish. One can then demonstrate that the outcome will be cooperation from the early beginning to almost the end (e.g. Kreps 1990: 536–543).

However, neither prediction corresponds in the least with the actual outcomes. In fact, the DD-outcome was obtained in only 14 percent of the rounds, while CC was chosen in 60 percent. Furthermore there were many fluctuations during almost the entire experiment, *prima facie* difficult to explain. Before proposing our own interpretation, we shall first consider in detail how the game evolved.

During the first round, Alchian plays as Nash would predict: expecting Williams to bet on the sure win (D). Playing C would be stupid. Williams, on the other hand, bets on the unstable CC-outcome, commenting "Hope he's bright". Even after the result (DC), Williams does not change his conception of "brightness," he hopes that Alchian will "wise up". Next follows a sequence during which Williams responds to Alchian's non-cooperative behavior with D, not so much because he changed his mind, but rather because he wants Alchian to face the consequences of his choice of play. Alchian is confused: he does not understand William's first two cooperative moves. Apparently with a view to gaining more knowledge, he again plays C in the fifth round. Williams interprets this as an attempt to try the cooperative route, and he responds by playing C for the next two rounds. In round 8, Williams reverts to D, which he keeps up for the subsequent three rounds without abandoning his idea of "brightness:" "In time he could learn, but not in ten moves", he comments. His option for D is part of a strategy to "teach" Alchian about brightness, and the strategy would appear to work: as of round 11, we observe a sequence of CC-outcomes. One D by Alchian in the sixteenth round frustrates Williams to such an extent that he, too, plays D in rounds 17 and 18, even when Alchian reverts to C.

The sequences 21–26, 31–38, 41–49, 51–67, 71–81 and 83–99 may be regarded as almost identical, albeit with ever-longer cycles: in every instance, Williams begins by breaking the DD-equilibrium. Alchian responds by also playing cooperatively, but with time he invariably reverts to defection. Some "punishment" follows, after which Williams goes back to playing cooperatively. Judging by Williams's comments and behavioral reactions, he responds rather emotionally to Alchian's non-cooperative moves ("He's crazy," "To hell with him," "This is like toilet training a child—you have to be very patient"). On the whole, however, these emotions make sense within Williams's *project* (which is expressed quite explicitly in his comment at round 10), i.e. "to coax Alchian into mutually profitable actions." One could argue that, given the ever-longer cycles, this project more or less succeeded. At the other side of the table, and at least up to round 19 but probably up to round 31, Alchian is confused more than anything else. From round 32 onwards, his comments rather reflect frustration with his inability to gain as much as Williams: "He will not share at all as a price of getting me to stick to C."

Finally, during the last rounds of the experiment, the mechanism of backward induction appears to come into play. Williams comments as early as round 10 that

he expects the mechanism to take effect at some point. He is also aware that he will have to pre-empt Alchian's anticipated defection: "On the last play, it would be conservative of me to switch to D, but I wouldn't do so if the evidence suggested that Alchian was a nice stable personality and not in a critical need of just a little cash". Alchian pre-empts this strategy by playing D in round 99— "as late as possible," as he commented at round 91.

CONSIDERING THE EXPERIMENT AS A SUPERGAME

In an experiment set up by Selten *et al.* (1997) almost half a century later, players were asked to design a complete strategy for a twenty-period supergame *prior* to play. In terms of the strategic game situation, the experiment resembles Robert Axelrod's famous "computer tournament" twenty years earlier (Axelrod 1984) as well as his follow-up research, in which people's decisions were modeled as computer algorithms (1997). In contrast to these computerized simulations, however, players in "real" experiments such as those conducted by Flood and Dresher are able to *adapt* their strategy in response to previous moves and outcomes. Be that as it may, looking upon the experiment as "one large multi-move game" is precisely what John Nash suggested to Flood and Dresher when they confronted him with their experimental evidence against the non-cooperative Nash-equilibrium (Flood 1958: 16). What can be learned, then, from comparing the Flood–Dresher experiment with these computer simulations?

Decomposing the Players' Strategies

To begin with, Selten *et al.* observed that each strategy consisted of three phases. In phase one, a cooperative goal or "ideal point" is chosen on the basis of fairness considerations. In phase two, a "measure-for-measure" policy is designed in order to get the opponent to move towards this ideal point. Only the final phase is dominated by the "normal" considerations of backward induction. Note that the first two phases as identified by Selten can also serve to decompose Axelrod's "winning" tit-for-tat strategy into (1) a cooperative start and (2) either cooperation or defection, depending on the other player's previous move. Selten's third phase does not apply to Axelrod's computer tournament, as the latter was supposed to continue eternally.

On the surface of it, one can also distinguish between three phases in Williams's play: in the first phase, the comment "Hope he's bright" clearly suggests that Williams is proposing what he believes to be the ideal point. Then follows a complex phase during which Williams seemingly (and initially rather unsuccessfully) tries to convince Alchian that this point is indeed ideal. The final phase, which sets in at approximately round 98, is dominated by

backward-induction reasoning. These three phases are present in the same order in Alchian's game, though they seem to be preceded by a phase of confusion during the first twenty rounds, apparently occasioned by the fact that Williams did not play the anticipated D during the early stages of the game.

Alternatively and in our view more appropriately, one might distinguish not so much between phases that are clearly demarcated in time as between *behavioral patterns*, whereby the players are seen to switch from one pattern to another depending on the dynamics of the game and how they interpret them. On the one hand, they play as fairness requires. We call this *norm-setting behavior*. On the other, they play in order to *teach* the other player. We call this *norm-teaching behavior*. Players seem to switch from norm-setting to norm-teaching behavior when they perceive the other to be deviating from the norm. Conversely, they switch back again when their teaching appears to have deflected them too much from the cooperative goal.

Fairness as an Opportunity

A second point of similarity between the two types of experiment might show up in the supposed "ideal point" that the players choose at the beginning. On the face of it, Selten may be correct in asserting that experimental subjects "make no attempt to predict the opponent's reactions and nothing is optimized. Instead of this, a cooperative goal is chosen by fairness considerations and then pursued by an appropriate design of the strategy" (Selten *et al.* 1997: 517).

Figure 2 illustrates which "ideal points" were identified by Williams and Alchian respectively. The figure represents the space of payoffs for both Alchian (x-axis) and Williams (y-axis) for the "large multi-move game". The corner points correspond to the "pure" strategies, while the shaded quadrilateral represents the set of possible outcomes for the one-hundred-times iterated game shown in Figure 1.

Wiliams's ideal would have been to play (C, C) a hundred times in succession. This ideal has nothing to do with the maximum payoff (200c$ in case Alchian would be "paying" 100c$). Wiliams's reasoning is made explicit in his comment during round 10. His benchmark is the non-cooperative "Nash" equilibrium (0, 50). Compared to this outcome, (50, 100) would clearly be appreciated by both parties. True, Williams's reasoning can be applied to any other point north-east of (0, 50), that is to any other point on the bold parts of the lines y = 133,3-100/150b and y = 300-4x. However, the choice for the (50,100)—outcome is an attractive "ideal point" in two respects: first, it is aesthetically attractive to choose a pure strategy. Such an aesthetic appeal is not without interest, given the primitive means of communication between the players. Second, and more important perhaps, it could also be said to be a "fair" focal point: if, as in Figure 3, we draw a 45º-line through the Nash equilibrium (0, 50), this line runs exactly through the (C, C) outcome. In

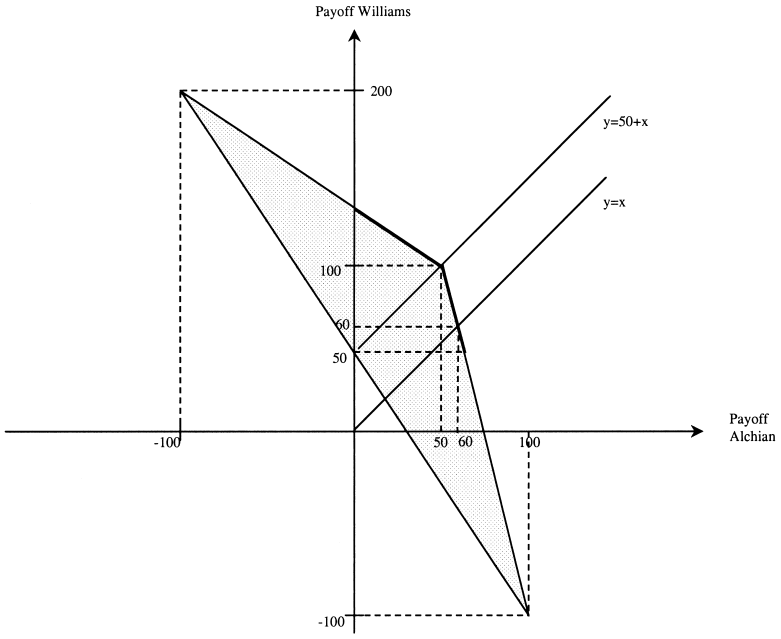

Figure 2: Flood–Dresher experiment: space of payoffs of the supergame°
Source: own construction, based on experimental payoffs (Poundstone 1992: 108–161, Flood 1958: 15).

°The shaded area represents the space of all possible outcomes when the game presented in figure 1 is repeated during 100 rounds. It is also the area between the four extreme points representing the payoffs of the four possible pure strategies (100x(C, C), 100x(D, C), 100x(C, D), 100x(D, D), or (50, 100), (0, 50), (–100, 200) and (100, –100) respectively).

Wiliams's mind, a pure (C, C)-strategy would split the surplus beyond the Nash-equilibrium into two precisely equal parts.[2] This is undoubtedly what Selten would call a "cooperative goal chosen by fairness considerations".

Though we have fewer clues as to what is Alchian's ideal (after round 31), we can safely assume that he did *not* consider it to be the pure (C, C)-strategy. It is very probable that the focal point he had in mind is located at the cross-section of the lines y = x and y = 300-4x, or the point (60, 60).[3] This point could have been

2 One might wonder what would happen if the "aesthetic" and the "fair" focal points do not coincide. A small change in the cooperative outcome (C, C) from (50, 100) to (50, 90) would be sufficient to study this.

3 Or, put more carefully, if the outcomes were to tend towards this focal point, Alchian would have been unable to invoke "sharing" as an excuse to play D.

reached by playing a mixed strategy, whereby (C, C) is played four times and (D, C) just once (though not necessarily in that order). It is an attractive option, as it would have resulted in an equal payoff to both players notwithstanding the initial asymmetry. To Alchian, anything else should be considered unfair, and he also frequently accuses Williams of being unwilling to share. Of course, it is not a very aesthetic focal point, which probably explains in part why Williams is so insensitive to Alchian's message.

The question arises whether we, as outsiders, might agree more with one player than the other. Both options are, in any case, focal points on the Pareto-frontier. As long as we respect the principle of non-comparability of utilities, we must remain indifferent to them. Interestingly, however, this principle seems of no concern to either player, however well acquainted with utility theory. Both players defend the relevance of *their* proposed ideal point, arguing that any deviation from it would be unfair. Perhaps it is because the debate is about money and not about an abstract notion of individual utility that an agreement is deemed to be possible.[4] Be that as it may, the impossibility of interpersonal comparability is *not* what causes disagreement; this disagreement would rather appear to be occasioned by their diverging views on whether inequality in initial endowments should be compensated for. If it should, then Alchian is right. If not, Williams is. Equality as such is not the issue. The issue is rather what kind of inequality one is willing to take responsibility for and to compensate for through redistribution.

Thus, while we have sufficient reasons to assume, just as in the Selten *et al.* case, that each player had an ideal point in mind, it is clear that, in the Flood–Dresher case, matters are more complicated, as there seems to be more than one ideal point. But in addition to this particular difference between the two types of experiment, the Flood–Dresher experimental outcomes also suggest that Selten may have drawn too sharp a contrast between "optimization" on the one hand and "fairness considerations" on the other. In the context of a real-actor supergame, players might also be assumed to "optimize" in a different way. In PD-like situations, each player's individual payoff depends on *both* players' decisions, while the repetitive nature of the game implies that players might influence the opponent's decision-making. In this context, the fairness of the cooperative proposal is clearly a *sine qua non*. Besides having the potential to convince the other party that it is an interesting option, a fair allocation is also rather unique, and thus the choice for such a point carries the pecuniary advantage of low transaction (bargaining) costs.

---

4 This argument was used by Sen (1990) in explaining male bias in intra-household resource allocation: the contribution made by the income-earning husband is in any case much more salient than that by the female homemaker.

Thus, players use their behavioral options as a communication device to arrive at "mutually profitable actions" (Williams, round 10), in order to make *more* money than the Nash-equilibrium would allow them to. But if a player uses his behavioral options primarily to *convince* the opponent, the link between players' interests and what they happen to choose becomes loose. In such situations, the players' revealed preferences should not be confused with what they consider to be their interests (in the supergame) (Sen 1979 [1976]).

Nor should their interests be confused with "myopic" round-by-round self-interest maximization. Note that this "myopic" self-interest can no longer play an explicit role in the players' strategies: if it could, it would be difficult to convince the other party that one is playing fair. Nevertheless, we cannot but notice the existence of what we might call an *elective affinity* between each player's "myopic" interests and their perception of what fairness implies in the context of the game. Though each proposed ideal point is arguably fair, either player seems to be curiously biased towards that point which serves his interests best. How to explain this self-serving bias (Babcock and Loewenstein 1997: 116) is a question in its own right.

The Art of Convincing

There is a third aspect of similarity between the two types of experiment, namely the phase/pattern of what Selten called a "measure-for-measure" policy. Indeed, the experiments converge in that both contain a response to situations where the other player is perceived to be out of line. However, as strategies in the Flood–Dresher experiment are constructed "live," the behavioral pattern one of the players refers to as "teaching" is able to become far more complex than simple "tit for tat". For example, players allow themselves to "punish" more than just once, purportedly in order to get their message across. Furthermore, they sometimes ignore defection on the part of their opponent. Finally, the "teaching" pattern is broken quite deliberately by Williams when he starts to play C again after a series of Ds.

Figure 3 clarifies the cost of the investments each player has made to arrive at the ideal point each had in mind. We calculate this cost in relation to the "sure" Nash-equilibrium (eternal (D, D)). This "sure" strategy predicts a payoff of 50c$ for Williams and breakeven (0 c$) for Alchian (see also Wiliams's remark during round 10). Because he is so focussed on teaching, Williams stays below this "sure" payoff up to round 51. In fact, he ends up with barely more than he would have obtained by playing a prudent eternal D (66$). Armen Alchian does much better than Nash would *ex ante* predict (40$ as compared to 0), but this appears to be the case *notwithstanding* his frequent attempts to teach Williams. Indeed,

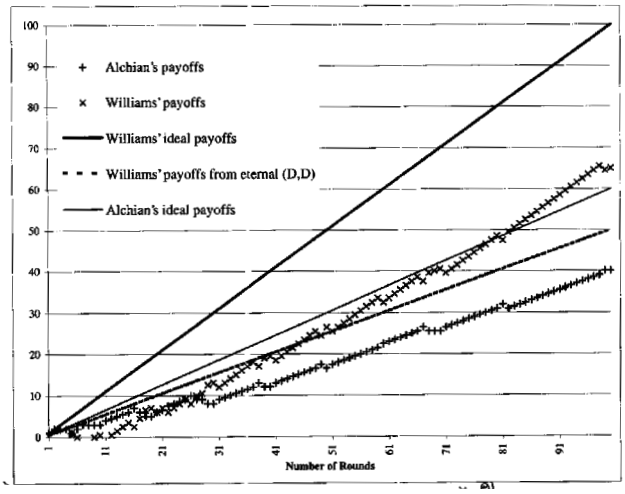

Figure 3: Effective and projected outcomes of Flood–Dresher experiment
Source: own calculation, based on experimental outcomes (Poundstone 1992: 108–116).

each time he wants to communicate to Williams that they should *share* (by playing D), he pays an opportunity cost, since Williams merely perceives this "teaching" as either opportunism or folly. Accordingly, Williams strikes back with D, allegedly to discipline Alchian, while remaining deaf to the message that Alchian is trying to convey. By comparing to what would have happened if Alchian had not tried to get Williams to share (a project he seems to undertake from round 23 onwards), we can calculate this opportunity cost at 7,5$,[5] or 20 percent of the payoff actually earned.

We can infer from this that many of the players' moves can indeed be interpreted as investments, as *costs* they are each willing to incur in order to influence the other's decisions, so as ultimately to arrive at the ideal point they have in mind. However, we also noted that this investment in teaching one another was both risky and costly; too risky and costly perhaps to make the investment worthwhile.

It follows that "investment" might provide insufficient an explanation for the players' behavior. As fair allocations are focal points, they minimize the transaction costs of coming to an agreement (Young 1998). But the before-discussed game experiment illustrates quite clearly that this cannot entirely explain the players' insistence on obtaining a "fair" outcome. If it were the case

5 Eternal cooperation from 23 onwards leads to 46, 5, as compared to the actual result of 39.

that the attraction of fair proposals lay merely in their being points where interests can easily meet, the players in the above-discussed experiment would have stopped "teaching" once they had become aware that the other was not so easily persuaded. However, the fact remains that they did not stop. There seem to be two (complementary) explanations for this. First, as suggested by the players' rather emotional reactions, they seemed to value punishment in its own right, as a way of "getting back" at their opponent, or in the words of Alchian: "I'll try once more to share—by taking" (round 70). Second, once engaged in this rather emotional battle, both players somehow seem to ignore the possibility of there being two focal points, which would imply that their attempts at convincing each other might be in vain. In any case, there is no indication that they ever came to doubt their own definition of fairness.

Similarities and Problems

To conclude this section, even if the Flood–Dresher experiment is played by "real" persons rather than computer algorithms, some interesting similarities show up if one interprets the experiment as a supergame. To begin with, the strategies deployed by the experimental players can be decomposed into three behavioral patterns between which players may switch as they see fit. In this case, they seem to switch between (1) norm-setting behavior, (2) norm-teaching behavior and (3) end-of-game behavior. Only in the last phase do they play as predicted by the principle of backward induction. To be sure, real players can switch between (1) and (2) as they see fit. We also observe that—in the case of "real" players- this teaching is far more elaborate and complex than the simple, clear and relatively forgiving tit-for-tat behavior often encountered in computer tournaments. Sometimes players were also much less forgiving and sometimes they simply ignored defection, interpreting it as a mistake. Never the less, these three behavioral patterns correspond closely to the three phases in each algorithm, as identified by Selten. Further, just as in the computer algorithms designed by Selten's students, the "norm" corresponds to a "cooperative goal chosen by fairness considerations". In the context of a supergame, such a norm is an evident objective once one decides to try and convince the other to play cooperative. In the 'live' game too, the fairness of a cooperative goal has clearly an instrumental value.

At the same time, however, there is more to the Flood–Dresher experiment than a supergame-reading would suggest. True, the explanation of the players' behavior in terms of the three before-mentioned behavioral patterns allows one to reconnect behavior and self-interest. This is an "enlarged" self-interest, defined at the level of the supergame. Never the less, such an interpretation remains somewhat incomplete. While all of the previous remarks seem to tie in with

a rather instrumental interpretation of fairness, such an instrumental view is not entirely consistent with observed reality. For one thing, we noted a curious *elective affinity* between players' myopic (round-by-round) interests and how they defined the "fair" cooperative outcome. Second, the interpretation is incomplete in that players enjoy some freedom in deciding when to switch between patterns. More specifically, sometimes "teaching" seems to have been reduced to a mere rationalization of rather blind "revenge". Third, players were also reluctant to reassess their initial opinion, up to the point of endangering their ultimate aim of attaining a better outcome than that offered by an eternal (D, D). The question arises: why were the players so deaf to each other's arguments?

CONSIDERING THE EXPERIMENTAL PLAYERS AS FOLLOWERS OF RULES OF CONDUCT

We shall not attempt to summarize the different ways in which social (or moral) norms have been discussed in the rapidly evolving (New Institutional Economic) literature. Instead, we shall fall back on Adam Smith's notion of "general rules of conduct" and contrast it with the more recent literature wherever we feel this might be useful. We begin with a short summary of Smith's insights on this matter. This will allow us to reconsider the set of problems we identified at the end of the previous section.

Smith argues that human beings have an interest in following "general rules of conduct" or "general rules" since they constrain *self-love*. As Akerlof and Dickens put it, people loving themselves also like to think of "themselves as 'smart, nice people'. Information that conflicts with this image tends to be ignored, rejected or accommodated by changes in other beliefs" (1982: 308).[6]

According to Smith, the effect of "self-love" in distorting our views of the world and of our own interests is the strongest precisely "when it is of most importance that they should be otherwise", namely "when we are about to act" and "when the action is over". He argues that just *before* acting,

> the eagerness of passion will seldom allow us to consider what we are doing, with the candour of an indifferent person. The violent emotions which at that time agitate us, discolour our view of things; even when we are endeavouring to place ourselves

6 Akerlof and Dickens derive this assertion from Festinger's theory of cognitive dissonance reduction, which says that "an individual who holds two or more cognitions (i.e. attitudes and beliefs) that are psychologically inconsistent will experience an uncomfortable state of tension, called dissonance. The individual will then be "driven" to reduce dissonance by changing one or more of the cognitions so that they are no longer inconsistent" (Quattrone and Tversky 1986: 39). The belief in one's own smartness and nice-ness may be a particularly "tough" cognition and indeed a benchmark for testing the reality-value of other cognitions.

> in the situation of another, and to regard the objects that interest us in the light in which they will naturally appear to him, the fury of our own passions constantly calls us back to our own place, where everything appears magnified and misrepresented by self-love.
>
> (1979 *TMS*: 157)

After we have acted, we can examine ourselves and our conduct more coolly, but self-love might again delude our mind:

> It is so disagreeable to think ill of ourselves, that we often purposely turn our view away from those circumstances which might render that judgement unfavourable. He is a bold surgeon, they say, whose hand does not tremble when he performs an operation upon his own person; and he is often equally bold who does not hesitate to pull off the mysterious veil of self-delusion, which covers from his view the deformities of his own conduct. Rather than see our own behavior under so disagreeable and aspect, we too often, foolishly and weakly, endeavour to exasperate anew those unjust passions which had formerly misled us … and thus persevere in injustice, merely because we once were unjust, and because we are ashamed and afraid to see that we were so.
>
> (Smith 1979 *TMS*: 158)

In other words, individuals who try to make case-by-case choices risk being worse off than individuals who pattern their behavior on some general rule of conduct. The problem of case-by-case decision-making is not so much that we have limited deciphering capacity, as argued for example by Herbert Simon (Simon 1981; see also Williamson 1987; North 1990), but rather that we are deluded by our self-image of being nice and smart. In the same vein, people who try to learn only from their own behavior run a greater risk of drawing the wrong conclusions than those who define the appropriateness of their behavior by observing the conduct of others (too).

> Some of their actions shock all our natural sentiments. We hear everybody about us express the like detestation against them. This still confirms, and even exasperates our natural sense of their deformity. It satisfies us that we view them in the proper light, when we see them in the same light. We resolve never to be guilty of the like, nor ever, upon any account, to render ourselves in this manner the objects of universal disapprobation.
>
> (1979 *TMS*: 159)

Thus, "morally appropriate" behavior and "self-interested" behavior should, in Smith's mind, not be considered to be opposites. Indeed, it is our moral sense that protects us from myopia. Somewhat paradoxically, we seem to be able to better pursue our interests when we can observe ourselves from a distance, i.e. from the

position of an impartial spectator. It is this impartial spectator who informs us about our "real" interests.

Note that Smith was careful to point out that the "ideal" (impartial) spectator we try to identify with should be distinguished from the *immediate* bystander. The latter's praise (or blame) can be very *partial*, in the double sense of being based on incomplete knowledge and biased by his or her own interests. An *ideal* (impartial) spectator, on the other hand, is supposed both to know everything and not to be affected by the consequences of our actions. He shares the latter quality with people we might qualify as *strangers*. If an actor can suppose that an indifferent observer would praise his action, he may qualify this action as "praise-worthy". If we want to know whether we are really worth what other people say we are, we have no option but to imagine ourselves in an *indifferent* spectator's position. If we were to use this internalized spectator as a

> looking-glass by which we can, in some measure, with the eyes of other people, scrutinise the propriety of our conduct … we can be more indifferent about the applause, and, in some measure, despise the censure of the world; secure that, *however misunderstood or misrepresented*, we are the natural and proper objects of approbation.
>
> (1979 *TMS*: 112, my italics)

Note also that, according to Smith, the individual's conception of "appropriate behavior" should not so much be considered as a trait that is instilled during "upbringing" and almost unthinkingly transmitted from one generation to the next. Rather, it is the product of an active process of observation of others' behavior, confrontation of these observations with one's own behavior and others' perceptions of it, and generalization of these observations towards other "similar" circumstances. Individuals gradually learn to recognize that certain situations are similar and that they are thus subject to similar rules of conduct—however remote these similarities might be in the minds of others.

Be that as it may, in practice there remain several degrees of freedom between what *is*, all things considered, appropriate action and what a specific individual might define as such in a specific situation. To begin with, given the differences in social position that individuals occupy in society and the differences in their personal histories, each will have been confronted with different choice situations and different "real bystanders" and hence develop a rather specific *man within the breast.* In contemporary terms, we could say that individuals develop different *identities* (Sen 1985: 348–349). Consequently, it is quite probable that individuals adopt different definitions of appropriate behavior.[7] Further, general rules of

7 See, in this context, also Smith (*TMS*: 145). In other words, it is rather problematic to speak of a society's culture as if it were a single, consistent whole.

conduct are by definition general, whereas specific choice situations are ultimately unique. This implies that there may always be some degree of freedom involved in interpreting a determinate choice situation.[8] Consequently, there can be significant differences in what people find the most appropriate course of action *in a specific situation.* Finally, most rules allow for exceptions. Following Elster's lead, one could argue that almost every social norm is accompanied by *adjunct norms*, by rules, as it were, which "form […] a penumbra around the main norm, a grey area that leaves room for manoeuvring" (Elster 1989: 110). Whether or not this grey area strengthens or erodes the original norm depends, among other things, on the criteria used to define the loophole (Ainslie 1992). The excuse must be "good," in the sense that the criteria allowing deviation from a general rule must be so salient and "beyond individual control" that the individual can still credibly say to herself that her behavior is praiseworthy.

This results in a relatively open perspective on "general rules of conduct", as the individual retains some essential freedom vis-à-vis "society". In this sense, Smith's account seems to open up a third way of thinking about norms and socially embedded behavior, somewhere in-between unrealistic myopia of an over-socialized *homo sociologicus* and the unrealistic insensitivity of the under-socialized *homo economicus* (Granovetter 1985, 1992; Sabel 1991; Williams 1988; Elster 1989; Platteau 1994). Yet, it would also be consistent with Smith to argue that the freedom that individuals *do* have for making sense of the world as they know it will probably be exploited by their self-love. While the general picture is one of "smart and nice guys" playing "reasonable," myopic self-love fills in the details.

These elements can now be applied in trying to answer the problems we identified at the end of the previous section.

Fairness and Self-Love

The curious existence of two ideal points in the Flood–Dresher version of the Prisoners' Dilemma turns out to be rather atypical when compared with other prisoners' dilemma experiments—though perhaps not with the real-life cases these PD's are supposed to represent (e.g. Sen 1984; Bardhan 2000). Following our interpretation, the existence of these two ideal points has to do with an asymmetry in the payoffs from cooperation. Be that as it may, it is interesting that both players retained *that* definition of fairness which came closest to satisfying their *own*

8 This aspect was discussed in Smith (*TMS*: 184–185).

self-interest. Note that we use the word "retained," not "chose". Indeed, the log provides no trace of a *deliberate* choice process between the two definitions. Thus, myopic self-interest seems to play a role on a less-than-conscious level, or at least it does in this particular instance. In other cases, this assertion might be more difficult to sustain.

Never the less, the apparent *elective affinity* between definition of fairness and underlying interests seems to confirm the hypothesis that in case of ambiguity between different "ideal points," "self-love" will fill in the details and ultimately determine which is chosen. As Smith asserts in different ways and on different occasions, the process of identification with others can never be complete; the individual can never entirely surrender his own position.

The limits to the influence of self-love are certainly clear to see: whenever it becomes too obvious that the "reasonably fair" option is merely a veil for self-interest, one may assume that the other party will not feel spontaneously attracted to this focal point. Whatever the implicit reason, the explicit reason for choosing one point over another must be that it has qualities unrelated to mere self-interest. Otherwise, there is no reason why the opponent should subscribe to it—except in the rare circumstance that the latter is truly altruistic. Wholly in line with this view, Elster defines social norms as *non-outcome oriented* injunctions on which to act (Elster 1999: 145). The Flood–Dresher experiment illustrates what happens if the other party is all but convinced of the disinterested nature of the injunction: a reasonable agreement is almost impeded.

Teaching and Self-Love

Once it had become clear to a player that his opponent was not immediately prepared to go along with his definition of fairness, both players felt a need to design what Selten calls a "measure-for-measure" policy. Or, in more general and perhaps more appropriate terms, a strategy to teach the opponent about fairness. Certainly Williams and Alchian both legitimize their own defection in terms of a strategy to "teach" the other about fairness.

Do we have any reason to disbelieve them on this point? Again, one could argue that their words are but an ex-post rationalization of myopic self-interest. Indeed, we cannot reasonably say that William's strategy was also the *best* pedagogical approach to making Alchian cooperate: certain moves are perhaps more easy to interpret as myopic revenge than as reasonable teaching. Moreover, this punishment sometimes resulted in rather long periods of mutual defection. The question therefore arises whether a simple tit-for-tat might not have been a much more transparent and pedagogical method. In Alchian's case, the

correspondence between "taking" and "teaching" was even more obvious (cf. round 70) .[9]

But even so, the fact that myopic self-interest seems, at least in part, to sustain the micro-level tactics of teaching does not in itself constitute an argument against considering "teaching" as a pattern, a rule of conduct which is relatively independent from the pico-economic incentives that sustain it. For one thing, as we have already argued, a fair outcome is also considered to be of instrumental importance to enabling both players to realize something more than the (D, D)-outcome.[10] By implication, "teaching" too is of instrumental importance, even if it appears to be partly determined by more myopic considerations.

Thus, the question is rather whether and to what extent myopic behavior intervenes in determining the moment of switching between the norm-setting and norm-teaching behavioral patterns. The first pattern might be referred to as the primary norm, while the second might—in Elster's terms—be called an *adjunct norm*. Could one reasonably argue that the excuse to deviate from the primary norm is a good one? Both Williams and Alchian shift from norm-setting behavior towards norm-teaching behavior *only* in response to the other player's "unfair" move. This seems reasonable, in the sense that they allow the other party to determine for themselves whether or not to deviate. Further, the "loophole" for defection that they found (i.e. in order to teach) is restricted in time. Williams's rationalization of defection as "punishment" allows him to play opportunistically, but the "punishment" must involve a constructive element, i.e. it must allow Alchian to better his life. This includes the possibility of ignoring the other's defection—or of interpreting it as an error—and of reverting to demonstrating the co-operative outcome, so as to break a negative cycle of mutual defection. The tactic of occasionally letting bygones be bygones increases the likelihood of coming to an agreement. To conclude, the overall picture is one of reasonable actors restraining their immediate impulses in order to further an interest that is obtainable only in the long run. Nevertheless, we cannot deny the existence of myopic, "emotional" reactions hidden beneath the surface of "reasonable" norm-teaching and norm-switching. Self-love might thus have affected certain details of the play.

---

9 Tit-for-tat behavior could also be interpreted as the outcome of what is referred to as weakness of will: "in principle I cooperate, but because of some difficulty in controlling emotion I deviate from this principle as a means of 'striking back' (revenge) or 'getting even' (envy)". This would imply that reciprocity has its roots in 'nature' rather than in 'culture' (Elster 1998; De Waal and Berger 2000). This is not to say that all expressions of envy or revenge are 'nature'; as argued at length by Elster (1990; 1991), different cultural contexts can result in widely different expressions of these emotions.

10 In this respect, rational choice theory is incomplete: depending on whether we see the actor as a case-by-case player or as a person with some consistency over different games, we reach quite diverging conclusions.

Note also that our proposal to interpret defection as part of a forward-looking measure-for-measure policy is not inconsistent with other interpretations of PD-experiments (Isaac *et al.* 1994; Selten *et al.* 1997; Fehr and Gächter 1998). The alternative would be to consider the tit-for-tat pattern (or conditional cooperation) as a fair pattern in itself. Indeed, a situation where Williams unconditionally plays C and where Alchian is consequently allowed to play D without being punished is also *intrinsically* unfair. This is one of the reasons why some authors have interpreted systematic reciprocity as a social norm *in itself*: "Cooperate on condition that everyone else cooperates as well" (Sugden 1984: 775; see also Sen 1985; Elster 1989: 213; Rabin 1993; Vandevelde 1993: 73; Dasgupta 2000: 347–351, Platteau 1994). The latter interpretation is, however, less compatible with the results of the Flood–Dresher experiment. Indeed, the players were less forgiving than tit-for-tat would require. Perhaps this rendered their behavior less fair, but also made it more pedagogical. Furthermore, note that the pattern of norm-teaching loses its rationale once it appears to lead to the dead-end alley of a series of (D, D). At that moment, players revert to norm-setting by playing C. By occasionally letting bygones be bygones, one increases the likelihood of reaching a fair outcome, even though this strategy is quite costly and, in itself, rather unfair. This voluntary restarting of the game sequence can only be understood if one clearly separates the two behavioral patterns of norm-setting and norm-teaching, *both* of which play an instrumental role in furthering the players' self-interest.

Self-love and Identity

While we have just argued that the players' self-reported rationalization of defection as "teaching" should be taken seriously, we cannot ignore that this strategy was not very successful. Indeed, an interpretation of the other's failure to adhere to the "fair" outcome as a problem of lack of brightness or lack of reason is consistent with the strategy of teaching, but this is not necessarily the only interpretation. In fact, we have pointed out that a correct interpretation would involve recognizing that there are several "ideal points" and that disagreement about fairness was possibly what prevented both players from playing a more balanced game. So why were the parties so "deaf" to one another's arguments about fairness? Although we cannot answer this question unequivocally, a Smithian interpretation of the experiment would suggest that a fair proposal is not only of instrumental importance, but that it also has an *intrinsic* appeal: anything else than equal treatment questions the *identity* of the parties—Smith's *man within the breast*. There is in principle no reason why someone like you should be entitled to a bigger or smaller slice of the cake. Conversely, allocations

do not just tell you what you get, but also who you are (or who you are compared to). Consequently, anyone deviating from the "fair" allocation not only questions the allocation itself, (s)he also questions who you are. This interpretation allows us to make sense of Alchian's repeated claim that "He won't share," "I must teach him to share," etc, and his ensuing costly attempts at inducing Williams to share.

Interestingly, as the opportunity to go beyond a mere (D, D)-outcome is dependent upon human beings' ability to identify with each other and to imagine themselves in each other's positions, the effective realization of this opportunity can also be impeded by a failure to fully identify, or by people's claim of somehow being entitled to a bigger slice of the cake. This observation is, of course, not new. It goes back at least to observations made by Max Weber a century ago:

> This universal phenomenon [the belief by the privileged that their good fortune is just] is rooted in certain psychological patterns. When a man who is happy compares his position with that of one who is unhappy, he is not content with the fact of his happiness, but desires something more, namely the right to his happiness, the consciousness that he has earned his good fortune, in contrast to the unfortunate one who must equally have earned his misfortune….
>
> (cited in Scott 1995:68)

We would submit, then, that it is precisely because Williams was so convinced of his "right to his happiness"—and that Alchian's misfortune was perceived by him as an "inevitable" result of the experimental payoff-matrix- that the log of his comments is replete with emotional exclamations such as "opportunist," "shiftless individual," "knave," "this is like toilet-training a child". More importantly, this is probably what led him to over-invest in convincing the other of this right.

CONCLUSION

In this paper, we set out to deepen our understanding of the importance of fairness in decision-making in the context of Prisoners' Dilemma games. A review of the "historical" Flood–Dresher experiment provided a useful empirical basis, as it allowed us to look in considerable detail at how the experimental players had made up their minds. In the second part of the paper, we considered several theoretical readings of the game outcomes.

The reading that seems to come closest to explaining the results is that the players' behavior is situated somewhere in-between that of the *homo economicus* and that of the *homo sociologicus*. Being more sensitive than the former and less myopic than the latter, the experimental subjects of the Flood–Dresher game

seem to define and follow general rules of conduct as defined by Adam Smith. This enables them to somewhat detach themselves from immediate impulses, and of trying out the cooperative outcome even when pure rational choice-reasoning would predict otherwise. It also induces them to teach one another about what is the "right" thing to do. On the other hand, it inspires a biased view on one another's behavior, a drawback that may explain observed over-investment or over-ambition in attempts to "teach" the other.

This reading ties in closely with the proposition by Selten *et al.* that multi-move games can be considered as supergames, whereby actors' strategies should be analyzed as coherent ways of playing. More specifically, it is possible to decompose the players' strategies in both types of games in behavioral patterns between which they switch as they deem fit, in view of obtaining a cooperative goal. However, while the case can be made that both experimental subjects played a more or less coherent strategy, it is hard to comprehend why they should have focussed so strongly on the "right" or "fair" way to play the game. A "fair" outcome seems to have more than a merely instrumental value as a focal point; it seems to have an intrinsic value as well. We submit that the fairness of the outcome is so important because it provides players with an identity: what they get tells them who they are—or who they are identified with. At the same time, this interpretation should take into account that there may be considerable ambiguity in the definition of a fair outcome. Further, agreeing with Adam Smith that great significance should be attributed to self-love, we predict that people will be biased towards that definition of "fairness" that suits them best. Thus, far from automatically focussing parties on *the* "fair" outcome, the fairness of a cooperative goal merely creates the *possibility* of arriving at a co-operative agreement. This, however, is not a trivial issue in contexts where eternal defection would *prima facie* seem to be the only rational course of action .

ACKNOWLEDGEMENTS

The author gratefully acknowledges the comments and suggestions of, among others, two anonymous referees, Antoon Vandevelde, Jean-Philippe Platteau, Simon Gächter, Stefaan Marysse, Jos Vaessen, Stephen Windross and Ralf Dua. An earlier version of this paper was also presented at a seminar at the University of Antwerp, November 2000. Any remaining errors are mine.

REFERENCES

Ainslie, G. (1992) *Picoeconomics; The Strategic Snteraction of Successive Motivational States within the Person*, Cambridge: Cambridge University Press.

Axelrod, R. (1984) *The Evolution of Cooperation*, London: Penguin Books.

Axelrod, R. (1997) *The Complexity of Cooperation; Agent-based Models of Competition and Collaboration*, Princeton New Jersey: Princeton University Press.

Akerlof, G. A. and Dickens, W. T. (1982) "The Economic Consequences of Cognitive Dissonance," *American Economic Review* 72(3): 307–19.

Babcock, L. and Loewenstein, G. (1997) "Explaining Bargaining Impasse: The Role of Self-Serving Biases," *Journal of Economic Perspectives* 11(1): 109–26.

Bardhan, P. K. (2000) "Understanding Underdevelopment: Challenges for Institutional Economics from the Point of View of Poor Countries," in: *Journal of Institutional and Theoretical Economics* 156(1): 216–35.

Becker, G. S. (1996) *Accounting for Tastes,* Cambridge MA/London England: Harvard University Press.

Camerer, C. and Thaler, R. H. (1995) "Anomalies: Ultimatums, Dictators and Manners," *Journal of Economic Perspectives* 9(2): 209–219.

Dasgupta, P. (2000) "Economic Progress and the Idea of Social Capital," in: P. Dasgupta, and I. Serageldin (eds) *Social capital: A Multifaceted Perspective*, Washington DC: World Bank: 325–424.

De Waal, F. B. M. and Berger, M. L. (2000) "Payment for Labour in Monkeys," *Nature* 404(April): 563.

Dupuy, J.-P. (1992) *Introduction aux Sciences Sociales; Logique des Phénomènes Collectifs*, Paris: Editions Marketing.

Elster, J. (1989) *The Cement of Society: A Study of Social Order*, Cambridge: Cambridge University Press.

Elster, J. (1998) "Emotions and Economic Theory," *Journal of Economic Literature* XXXVI(March): 47–74.

Elster, J. (1999) *Alchemies of the Mind; Rationality and the Emotions*, Cambridge: Cambridge University Press.

Fehr, E. and Gächter, S. (1998) "Reciprocity and Economics: the Economic Implications of Homo Reciprocans," *European Economic Review* 42(3): 845–859.

Flood, M. M. (1958) "Some Experimental Games," *Management Science* 5: 5–26.

Frank, R. H. (1987) "If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience?" *American economic Review* 77(4): 593–604.

Frank, R. H. (1988) *Passions Within Reason; The Strategic Role of the Emotions*, New York, London: W.W. Norton & Company.

Gintis, H. (2000) *Game Theory Evolving*, Princeton NY: Princeton University Press.

Granovetter, M. (1985) "Economic Action and Social Structure: A Theory of Embeddedness," *American Journal of Sociology* 91(3): 481–510.

Granovetter, M. (1992) "Problems of Explanation in Economic Sociology," in N. Nohria, R. G. Eccles (eds) *Networks and Organizations*, Harvard: Harvard Business School Press: 25–56.

Isaac, R. M., Walker, J. M. and Williams, A. W. (1994) "Group Size and the Voluntary Provision of Public Goods: Experimental Evidence using Large Groups," *Journal of Public Economics* 54(1): 1–36.

Kahneman, D., Knetsch, J. L. and Thaler, R. (1986) "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," *American Economic Review* 76(4): 728–41.

Kreps, D. M. (1990) *A Course in Microeconomic Theory*, New York: Harvester Wheatsheaf.

Luce, R. D. and Raiffa, H. (1967) *Games and Decisions: Introduction and Critical Survey*, New York: John Wiley & Sons Inc.

Lundberg, S. and Pollack, R. A. (1993) "Separate Spheres Bargaining and the Marriage Market," *Journal of Political Economy* 101(6): 988–1010.

North, D. C. (1990) *Institutions, Institutional Change and Economic Performance*, New York: Cambridge University Press.

Platteau, J.-Ph. (1994) "Behind the Market Stage where Real Societies Exist," *Journal of Development Studies* 30(3–4): 533–577, 753–817.

Poundstone, W. (1992) *Prisoner's Dilemma,* New York: Doubleday.

Quattrone, G. A. and Tversky, A. (1986) "Self-Deception and the Voter's Illusion," in Elster, J. (ed.) *The Multiple Self*, Cambridge, MA: Cambridge University Press: 35–58.

Rabin, M. (1993) "Incorporating Fairness into Game Theory and Economics," *American Economic Review* 83(5): 1281–1302.

Rapoport, A. Chammah, A. M. (1965) *Prisoner's Dilemma: A Study in Conflict and Cooperation*, Ann Harbor: The University of Michigan Press.

Sabel, C. F. (1991) "Constitutional Ordering in Historical Context," in F.W. Scharpf (ed.) G*ames in Hierarchies and Networks*, Boulder, CO: Westview Press: 65–124.

Scott, J. C. (1995) *Domination and the Arts of Resistance; Hidden Transcripts*, New Haven and London: Yale University Press.

Selten, R. and Stoecker, R. (1986) "End Behavior in Sequences of Finite Prisoner's Dilemma Supergames; a Learning Theory Approach," *Journal of Economic Behavior and Organization* 7: 47–70.

Selten, R., Mitzkewitz, M. and Uhlich, G. R. (1997) "Duopoly Strategies Programmed by Experienced Players," *Econometrica* 65(3): 517–555.

Sen, A. K. (1979) [1976] "Rational Fools: a Critique of the Behavioral Foundations of Economic Theory," H. Harris, (ed.) *Scientific Models of Man; the Herbert Spencer Lectures 1976*, Oxford: Clarendon Press: 1–25.

Sen, A. K. (1984) *Resources, Values and Development*, Oxford: Basil Blackwell.

Sen, A. K. (1985) "Goals, Commitment, and Identity," *Journal of Law, Economics and Organization* 1(2): 341–355.

Sen, A. K. (1990) "Gender and Cooperative Conflicts," in I. Tinker (ed.) *PersistentIinequalities: Women and World Development*, New York: Oxford University Press: 123–149.

Simon, H. A. (1981) *Sciences of the Artificial*, Massachussets: MIT Press.

Smith, A. (1976/1790) *The Theory of Moral Sentiments*, (D. D. Raphael and A. L. Macfie, eds) Oxford: Clarendon Press.

Smith, A. (1979/1791) *An Inquiry into the Nature and Causes of the Wealth of Nations*, (R. H. Campbell and A. S. Skinner, eds), Oxford: Clarendon Press (2 vols).

Sugden, R. (1984) "Reciprocity: The Supply of Public Goods through Voluntary Contributions," *Economic Journal* 94: 772–787.

Vandevelde, T. (1993) "Eigenbelang, Rationaliteit en Ethisch Handelen," in P. Reynaert, (ed.) *Wetenschap en Waardevrijheid*, Leuven: Garant: 63–84.

Williams, B. (1988) "Formal Structures and Social Reality," in D. Gambetta (ed.) *Trust; Making and Breaking Cooperative Relations*, Cambridge: Cambridge University Press: 31–48.

Williamson, O. E. (1987) *The Economic Institutions of Capitalism*, New York: The Free Press.

Young, H. P. (1998) *Individual Strategy and Social Structure; An Evolutionary Theory of Institutions*, Princeton: Princeton University Press.