Economics and Philosophy, **28** (2012) 291–310 Copyright © Cambridge University Press
doi:10.1017/S0266267112000259

A STRAWSON–LEWIS DEFENCE OF SOCIAL PREFERENCES

JELLE DE BOER
University of Amsterdam and Delft University of Technology, the Netherlands
j.s.deboer2@uva.nl

This paper examines a special kind of social preference, namely a preference to do one's part in a mixed-motive setting because the other party expects one to do so. I understand this expectation-based preference as a basic reactive attitude (Strawson 1974). Given this, and the fact that expectations in these circumstances are likely to be based on other people's preferences, I argue that in cooperation a special kind of equilibrium ensues, which I call a loop, with people's preferences and expectations mutually cross-referring. As with a Lewis-norm, the loop can get started in a variety of ways. It is self-sustaining in the sense that people with social preferences have sufficient reason not to deviate.

1. INTRODUCTION

Perhaps there will always remain people who think that every aspect of human behaviour ultimately springs from egoistic concerns. Decades of careful observations and laboratory experiments in the social sciences now increasingly render this position dogmatic.[1] Genuine social motives, or in the language of economics, social preferences, seem to be here to stay. In the behavioural economics literature, most notably, social preferences have been high on the research agenda in recent years. Here I want to examine a special kind of social preference, namely a preference to do one's part in a mixed-motive setting because the other party expects one to do so. I understand this expectation-based preference as a basic reactive

[1] The volume of literature is vast; here are some exemplary findings: Orbell *et al*. (1988); see David Sally (1995) for a review and a meta-analysis; Roth *et al*. (1991); Bohnet and Frey (1999); Frank (2004); Henrich *et al*. (2004); Bischoff (2007).

https://doi.org/10.1017/S0266267112000259 Published online by Cambridge University Press

attitude (Strawson 1974). Given this, and the fact that expectations in these circumstances are likely to be based on other people's preferences, I argue that in cooperation a special kind of equilibrium ensues, which I call a loop, with people's preferences and expectations mutually cross-referring.

In section 2 I first argue against the idea of understanding social preferences in terms of rules or norms. It should be the other way around: rules or norms should be understood in terms of individual attitudes (as David Lewis did with conventions). Sections 3 and 4 develop the case for expectation-based social preferences, which I call 'expectation-based-cooperativeness'. Sections 5 and 6 question how justification works given this Strawsonian attitude. I profess that justification works by means of a loop, through the mutual cross referring of people's social preferences and expectations. Sections 7 and 8 compare this with alternative views.

2. RULES, NORMS AND LEWIS

Most behavioural economists define social preferences in the most general sense as preferences that are other-regarding and not self-interested (section 3 will discuss various subspecies). To explain social preferences, quite a number of scholars from various disciplines allude to the existence of rules or norms to which people willingly conform. Let me briefly discuss two recent contributions.

In a discussion of social preferences, the Nobel-prize winning economist Vernon Smith wrote that humans do not typically 'choose in their strict immediate self-interest' but 'engage in social, rule-governed (. . .) interaction' instead.[2] Another critic, Bart Wilson, was even more outspoken: 'the challenge is to reconstruct the distinctive system of rules that defines social motives within a pattern of social practice'. The economist should see whether 'the subjects agree on the context of the interaction: do the subjects agree that this situation invokes the social motives of a rule of reciprocity, a rule of fairness, a rule of just desert, a rule of equity, a rule of equality, or some other rule to guide behaviour?'[3] Hence, in an interaction situation, people have to agree on which rules are to guide behaviour. Wilson lists several possible rules: a rule of reciprocity, a rule of fairness, a rule of just desert, a rule of equity, a rule of equality, and there could be more.

But clearly, assuming that the subjects have agreed beforehand which rules to follow is to deny the heart of the collective action problem. Where does this agreement come from, and why would it matter? With Hume, who considered why we should obey a promise, we may ask how it is

[2] Smith (2010: 85).
[3] Wilson (2010), both quotes are from p. 81.

https://doi.org/10.1017/S0266267112000259 Published online by Cambridge University Press

possible that an agreement is binding in the first place.[4] Saying that you agree, for example, to the rule 'I cooperate, you cooperate' only reveals your prior motive to cooperate and your expectation that the other will cooperate too. So aren't these the factors that we should attend to?

Psychologists Weber and Murnighan (2008) found that in a six-round four-person Prisoner's Dilemma (PD) the presence of at least one individual who makes the cooperative choice each and every time, whom they call a 'consistent contributor', tends to increase cooperation in other people. The influence of such a contributor, they say, 'seems to be mediated by their effect on fellow group members' perceptions of their group's social norms, specifically by prompting fellow group members to see the social norms that apply as more cooperative than they would in the [consistent contributor's] absence'. According to Weber and Murnighan, in social dilemmas a cooperative norm or a competitive norm can be at stake. What happens with consistent contributors is that they make the others act in accordance with the cooperative norm. Without them, the competitive norm would prevail.[5]

Now let us put ourselves in the shoes of those other people who take part. What does it mean to say that a cooperative norm applies in the eyes of the participants of the PD? Generally speaking, that should depend on enough people being around who seem willing to cooperate. It would be odd, even contradictory, to think that a cooperative norm is at work in some group while also thinking that the members lack any cooperative inclinations whatsoever. Weber and Murnighan say that the presence of consistent contributors influences how the other subjects perceive the choice situation: they make the others think that a cooperative as opposed to a competitive norm applies. I suggest that we try to understand this in a less roundabout manner and omit the reference to a norm. We can also say that consistent contributors give rise to a belief that there are enough people around with cooperative inclinations. They function as evidence that a sufficient number is going to cooperate. This belief may be right or wrong. One consistent contributor in a group of four during the first two rounds may tip somebody's balance in the third round. It is also easy to imagine that someone is mistaken in having this belief. But suppose that this is how it works. The crux of the matter is what the agents around one seem to have in mind. Compare this with a cooperative *atmosphere*.

[4] Hume (1739: III.II.v).

[5] Note that these 'consistent contributors' are not necessarily unconditional cooperators. In fact, one of the results of Weber and Murnighan's study is that these subjects obtained good outcomes, they did 'not suffer in the process' (1343). Apparently they have not been suckered in this six-round repeated game. That would make them irrational. This also means that others have not defected on the consistent contributors. It hasn't happened here but defecting on an unconditional cooperator would have been quite rational, or so I'll claim in section 5.

https://doi.org/10.1017/S0266267112000259 Published online by Cambridge University Press

This is nothing in itself. Whether there is a cooperative atmosphere in a group depends on the members of the group having beliefs about each other's cooperative ideas and plans. And so it is with a cooperative norm. Of course one can say that such a norm exists but it is the subjects who are around and what can be expected from them that explains the norm.

The notions 'cooperative' and 'competitive' as applied to one's surroundings are *inherently* strategic. They tell us something about what other people can be expected to do. If we know what other people can be expected to do, then we know all we need to know. Given this knowledge, talking of *norms* of cooperation or competition does not pull extra explanatory weight.

I conclude that a reference to norms and rules is of no great avail for our purposes. Trying to understand social behaviour in terms of norms or rules, without invoking the underlying individual attitudes, is in a sense a regress – back to before the times of David Lewis and Thomas Schelling.[6] I do not deny that norms can explain cooperation, but the explanation needs to run deeper: what is a norm, and how does it work? In line with Lewis and Schelling I propose to understand a pattern of cooperative behaviour in incidental mixed-motive settings in terms of individual actions, expectations, preferences, and how these interact. Of course, Lewis confined 'his attention to situations in which coincidence of interest predominates',[7] whereas we are interested in the PD, a case of mixed motives, but that need not imply that the methodology of individual actions, expectations, and preferences is going to fail us, and that norms are to be called in for help; it could also mean that a human tendency plays a role in this situation that Lewis had no need to address: our social preferences.

3. TWO TYPES OF SOCIAL PREFERENCES

With Schelling and Lewis, then, cooperative behaviour in one-shot settings should be explained, not by invoking norms or rules, but in terms of preferences and expectations.[8] Also, we have assumed that it is not

[6] Schelling (1960, 1978) Lewis (1969). For some recent developments along these lines see Sugden (1998), Cubitt and Sugden (2003), Sillari (2005, 2008), De Bruin (2005), Binmore (2008), Ross (2008).

[7] Lewis (1969: 14).

[8] That is in line with the economics literature. Of course, the preference–expectation model (the desire-belief model) is not the only game in town for explaining cooperative behaviour. Its proper use presupposes that people's preferences and expectations are sufficiently coherent and stable. This is what I presuppose indeed. In this paper I am interested in delineating the rationality of cooperation. That means that we should begin with understanding behaviour in terms of propositional attitudes (as far as that goes), and that is most simply the preferences-expectations model. I say more on this matter in section 7.

https://doi.org/10.1017/S0266267112000259 Published online by Cambridge University Press

true that people always act on egoistic preferences. Conclusion: people sometimes act on truly social preferences. That need not imply 'ad hocery' if these preferences can be formulated in a tractable way and empirically established.

Social preferences are here to stay. Now the next question is: what are they like? In the behavioural economics literature, two forms are commonly identified: outcome-based and intention-based. Outcome–based versions say that people care about the material pay-offs that accrue to the people they are dealing with in strategic settings. Simple altruism is a possibility, when somebody else's pay-off makes up an argument in my utility function (Andreoni and Miller 2002). Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) defend a distributional concern: inequity aversion. Charness and Rabin (2002) argue for Rawlsian maximin.

Intention-based theories say that interpreting the motive behind the action of one's counterpart matters, too. Rabin (1993) is a seminal paper, and a point of departure for many others. His idea is that people want to be kind (or unkind), i.e. sacrifice some of their own material pay off, to those who they think are kind (or unkind) to them. My belief that you are kind by not making the egoistic strategy choice makes up an argument in my utility function.[9] If large enough, it prompts me to make the non-egoistic choice as well. People want to reciprocate: repay kindness with kindness, and unkindness with unkindness. Players with the mutual belief that they are motivated by kindness can attain the cooperative outcome in a one-shot PD.

Ingenious experimental designs have demonstrated that both concerns play a role in cooperation, e.g. Cox (2004), Bacharach *et al*. (2007), Cox *et al*. (2008), Falk *et al*. (2008). Many people care about reciprocating somebody's motive, but the allocation of material pay-offs as such also appears to be a relevant concern.

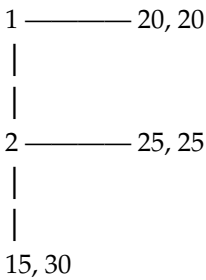
4. EXPECTATION-BASED COOPERATIVENESS AND STRAWSON

In this paper I want to concentrate on intention-based social preferences. Rabin's theory left an innovative mark in economics. Still, there are two problems with it. Firstly, the underlying psychology is quite straightforward, perhaps a little too simple in a sense. Rabin does not say where this motive of reciprocating comes from, or what its possible grounds are. Why exactly should I be kind if I think you are? Surely returning kindness or unkindness is not just a reflex? Secondly, this

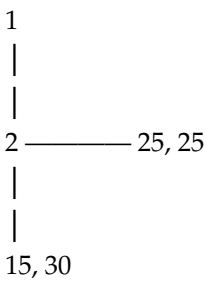
[9] This is unorthodox: psychological game theory. Rabin's forerunner is Geanokoplos *et al*. (1989). Cf. Colman (2003). In standard game theoretic Nash equilibrium individual behaviour is contingent on the other party's expectations, namely in so far as it presupposes common knowledge, but with social preferences like in Rabin's theory expectations directly *cause* the motive to make the cooperative move.

https://doi.org/10.1017/S0266267112000259 Published online by Cambridge University Press

attitude of reciprocating kindness or unkindness cannot be all that matters, especially not in a sequential PD. Because what can a second player in this situation infer about the first player who has just made a cooperative move? That he has been kind? Why, it just seems that player 1 trusts player 2; he expects him to make the cooperative choice. Perhaps it is simply this trust that player 2 responds to.[10] McCabe *et al*. (2003) presented test subjects with the following game:



They also ran this experiment, in which player 1 now has no choice other than moving down:



One important result was that twice as many played right at node 2 in the first game as compared with the second game. This provides evidence that, indeed, people who act cooperatively in such settings are not merely concerned about outcomes and distribution but also about what the other party has in mind. Why would player 2 choose 'right' and not 'down' in response to the 'down' choice of player 1, someone he does not know and will have no future dealings with? McCabe *et al*. say that if player 1 moves down he 'signals an intention to enter a reciprocal-trust relationship' and

[10] Rabin himself notes that his model omits the sequential form, and recommends this for future research (1296). Dufwenberg and Kirchsteiger (2004) elaborate on Rabin's theory for this game, and propose a new solution concept, sequential reciprocity equilibrium. Central to their theory is still, however, the idea that people simply want to repay kindness with kindness and unkindness with unkindness.

https://doi.org/10.1017/S0266267112000259 Published online by Cambridge University Press

that player 2 wants to 'reciprocate' this.[11] Less intricate, I suggest, is that player 2 simply understands that player 1 expects him to choose 'right', and that it is this expectation on player 1's part that player 2 wants to fulfil.

Let us call this motive to fulfil another person's expectation in a mixed motive setting *expectation-based-cooperativeness* (sometimes more briefly *cooperativeness*). Obviously, this attitude cannot be understood from a self-interest point of view. A sceptic might claim that people who act like this in a sequential PD are not rational.[12] They do not understand the game well enough, and more playing will make them converge towards defection. But is this game so difficult? With these games social scientists have learnt to expend considerable effort to ensure that their subjects understand the games they have to play and exclude subjects from the data if otherwise.[13] Also, it does not seem to be true that cooperation in such settings correlates with stupidity. At least, in an experiment among Cornell University students, cooperation increased with level of education; those nearer to graduation tended to cooperate *more* in one-shot PDs than their juniors.[14]

An association to rules and norms does not seem to clarify the matter either; see section 3. Surely a plausible evolutionary story can be told that people and even primates who are socially inclined outreproduce their egoistic brethren, but that mostly explains why we *have* these attitudes. 'Because I am a social primate' is not a very satisfactory answer to the question of why one chooses to cooperate.

Or perhaps we *should* understand this attitude as more or less reflexive in nature. Perhaps such cooperativeness can only *provide* reasons and does it not derive from earlier reasons. With Strawson, we could understand expectation-based-cooperativeness as a 'normal reactive attitude'. Then it wouldn't have further rational bases. There are no further justifications outside of the reactive attitudes because, as Strawson says, 'questions of justification are internal to the structure'.[15]

Strawson famously spoke of a range of reactive attitudes like resentment, gratitude, love, affection. We normally exercise these attitudes in our dealings with each other. They make up a web or a framework through which the ideas of free will, justice, desert, guilt and responsibility get their meaning. It is not this web of attitudes that should be justified.

[11] (269).

[12] For example Binmore (1995).

[13] With precautions, I believe that experimental one-shot games can isolate social preferences. However, some scholars say that only repeated games sufficiently reflect real life and that one-shot game experiments are inherently flawed, or at least often problematic. I will not enter this discussion here. For a good defence of the use of one-shot games see Camerer and Fehr (2004).

[14] Frank (2004: 172).

[15] Strawson (1974).

https://doi.org/10.1017/S0266267112000259 Published online by Cambridge University Press

That, Strawson says, is neither called for nor possible. Justification works from within the system of reactive attitudes. It is my aim in this section and the rest of the paper to examine how justification might operate, given the reactive attitude of expectation-based cooperativeness, as applied to the issue of cooperation in incidental mixed-motive games.

Strawson says that 'the reactive attitudes rest on and reflect, an expectation of, and demand for, the manifestation of a certain degree of goodwill or regard on the part of other human beings towards ourselves'. As said, I propose to follow up on this and translate this to our case by saying that the attitude of expectation-based cooperativeness is also a primitive reactive attitude. Let me characterize this by means of Hume's famous tale of the two farmers: 'Your corn is ripe to-day; mine will be so tomorrow. 'Tis profitable for us both, that I shou'd labour with you to-day, and that you shou'd aid me to-morrow.' I believe that for many people in the role of the second farmer a desire simply to live up to the expectations of the other person is part of their normal psychology. McCabe *et al*.'s (2003) experimental design, which I just mentioned, nicely isolates this motive.

Let us say that the typical circumstances of expectation-based cooperativeness are a two-person strategic situation, as in the McCabe *et al*. (2003) trust game and with Hume's farmers:

- 1 expects 2 to cooperate in a mixed motive setting.
- 2 could let 1 down, thereby creating more material gains for himself.
- Expectation-based cooperativeness belongs to 2's normal motivational repertoire.

Surely, in mixed-motive situations like a sequential PD, as with Hume's farmers, there is a difference, namely a danger of intruding free riders. Thus in such circumstances one should be sensitive to this: one should neither trust nor collaborate with those who have betrayed one or others in the past; one should not fulfil people's expectations if these are infused by exploitation. Cooperative inclinations in mixed-motive settings cannot be naïve; they must be hedged by effective methods to teach defectors a lesson and possibly shut them out of the game. (I will say more on this in section 6.) In short, I suppose that the tendency to collaborate with someone because he expects this can be effective in strategic contexts like the PD but then augmented with the capacity to track and ostracize possible free-riders.

5. A LOOP OF MUTUALLY REFERRING PREFERENCES AND EXPECTATIONS

The next question is how this attitude of expectation-based cooperativeness can work and be justified in an incidental mixed-motive game.

https://doi.org/10.1017/S0266267112000259 Published online by Cambridge University Press

When is it appropriate for me to make the cooperative choice in such a situation? According to the view that I will develop from here on, one party's preferences are based on the other party's expectations, while the other party's expectations are based on the first party's preferences – and vice versa.[16] People's preferences and expectations are cross-referential. In this way, practical inference is closed. There are two steps.

Firstly, if I am stirred by this social tendency and consider making the cooperative choice in a one-shot PD, I should have reason to suppose that my counterpart is going to cooperate, too. That seems obvious enough: I simply don't want the sucker outcome. But why do I believe that the other is going to cooperate? I believe that he is going to cooperate because he is *motivated* to do so, not for example because he trembles. If he cooperates with a trembling hand, then he really meant to sucker me, and so I will defect. The same holds if I have reason to believe that his motivation is dubious, e.g. when he only cooperates because he wants to make a good impression on somebody else. Normally I believe that the other player is going to cooperate because I assume that he is socially motivated. So my belief about somebody else's conduct depends on an assumption about his preferences.

Secondly, the question is why I should behave cooperatively in a situation in which others expect me to do so. As I have argued, it is precisely this expectation that player 2 responds to. An important part of his reason, it seems, is exactly *that* the others expect him to cooperate at node 2. The fact that the other expects me to cooperate gives me a reason to do so per se. People are inclined to cooperate at least partly *because* others expect them to cooperate. They want to live up to the expectations of their fellow players. To see this, consider an individual who you expect to cooperate with you in a PD but about whom you also know that he does not expect *you* to cooperate. He acts with a saint-like morality, since he could not care less whether you cooperate or not. I suppose this makes a difference. As this saint does not seem to bother about your conduct nor about his own interest, you may just as well go for the higher material gain. Hence I suppose that for rational people their willingness to cooperate in a one-shot PD partly depends upon the expectations that the other players have about them.

If this is the way things are with two individuals in a one-shot PD, then it is the case that their preferences and beliefs are mutually referential. My expectation about the possibility of cooperation refers to your preference to cooperate, and my preference to cooperate

[16] For this I draw on den Hartogh (2002, chs. 2 and 3; 1998).

https://doi.org/10.1017/S0266267112000259 Published online by Cambridge University Press

relates to your expectation about me. As this also holds for you, we get:

a) I prefer to cooperate, partly based on b).
b) You expect that I will cooperate, partly based on a).

If there is sufficient cooperative disposition, then we can go through this a)-b) loop endlessly. There is no need for further reason-giving as soon as you are in it. The loop is enough. It is free floating. In this loop our preferences and expectations are in mutual justificatory equilibrium.

Hence, rational individuals with social preferences who are also aware of the fact that everybody is rational and social will cycle through the a)-b) structure if pressed to explain themselves. P is the cooperatively inclined player; I is a sceptical interlocutor.

I 'Why are you willing to cooperate in a one-shot PD? Why don't you just defect?'
P 'Part of the reason is that he expects me to cooperate.'
I 'So?'
P 'Well, I do not want to defy his expectation.'
I 'But why does he expect you to cooperate then?'
P 'Because he believes that I am willing to cooperate. He trusts me.'
I 'Yes, but why are you willing to cooperate . . .' (and then he sees the circle).

6. NO NEED FOR FURTHER JUSTIFICATIONS

It is typical for an equilibrium state that it is of little importance where exactly the underlying forces – here the social preferences and the expectations – come from.[17] What matters is that they are there and upheld by their relationship of interdependence. Their causal histories or possible further justifications can be ignored.

One's expectation that the other will cooperate could be produced by an explicit exchange of promises, or by something more primitive like the other party's non-verbal signalling. It could also spring from something like a default assumption that the other is cast in the same mould as oneself.

The same holds true for one's preference to honour somebody else's expectation that one will make the cooperative choice. What matters is that your expectation that I will cooperate makes up a reason for me. I do not want to betray this confidence. This, however, is not a deeper reason; it

[17] Sober (1983), Kincaid (2008), Lewis (1969), Sugden (1989), Pettit (1996).

https://doi.org/10.1017/S0266267112000259 Published online by Cambridge University Press

is just the same reason, formulated differently.[18] But where does *that* come from? Is there no deeper reason? I claim that there isn't, or at least, deeper reasons are not needed. His expectation that I will cooperate is sufficient as a reason. It can stand on its own, and it proves valid if its conditions of satisfaction are met, viz. he expects me to cooperate indeed and does not want to sucker me.

Earlier reasoning may have led to this point, but it is not necessary that this occurs. For example, one could claim that one does not want to betray a person's confidence because that would hurt or harm this person. Or one might think of this in terms of universalizability. Or one could see it as a token of bad character. These classical moral lines of reasoning are possible but not necessary. With Strawson, the relevant justification works from within – that is along the lines of the loop for the case at hand. Why do you prefer to cooperate? Because he expects me to do so. Why does he expect you to do so? Because he thinks that I prefer to cooperate.

Practical reasoning in an interactive mixed-motive setting need not trace the causal pathways and possible justifications behind the relevant attitudes, because in equilibrium these attitudes are rationally based on just each other. In the loop, practical reasoning is closed. For example, somebody's social preference in a PD can be a product of a quite inarticulate loyalty, which has to do with belonging to some group. Perhaps the other is a fellow student, a compatriot, a partner in crime, or, just like him, a poor victim of experimentation. So such loyalty can function as a first mover when somebody sets foot in the experimental economist's laboratory to play an anonymous PD. This person doesn't see the other players but he may simply assume that they are, for example, students like him. From here, two things matter. First, he should reconsider when this assumption proves to be wrong. Second, when this assumption has not been proven wrong he should reason along the lines sketched in the interview between P and I above when pressed for an answer.

Hence, the loop can get started in a variety of ways. It is self-sustaining in the sense that people with social preferences have sufficient reason not to deviate. With a coordinative norm, a Lewis convention, deviating is harmful for everyone, including the deviator. Everyone has a reason to stick to the coordinative norm. Acting differently would be against one's self-interest. Hence the stability of this norm.

People with social preferences in a PD also have a reason to stick to the norm, now a cooperative norm, i.e. the loop. This is not a self-interested reason, but a reason that refers to somebody else, a person who expects

[18] After all, his expectation not being a reason for me is something that I do not want. These two negations can be cancelled, and that gives us the earlier formulation, while saying that his expectation not being a reason for me is of course the same as betraying his confidence.

https://doi.org/10.1017/S0266267112000259 Published online by Cambridge University Press

cooperation. One doesn't want to betray this person's confidence. That is a valid enough reason.

Of course such a norm can be undermined and a cooperative pattern can unravel. Self-interested players, who do not care if they sucker someone, can break it. So it is necessary that people only cooperate when they have a reason to think that others are like-minded, or at least, when they do not have a reason to think that the other is of a different mind. They should not tend to cooperate when there are defectors around. Now this is exactly how people seem to reason. Experiments show that the tendency to cooperate in incidental mixed-motive games is not unconditional. People do not exercise these preferences out of the blue. Subjects who are socially disposed in one-shot PDs make the cooperative choice only when they believe that the other actors will do so too. The best predictor of what people will do in such circumstances is what they think others are going to do. This seems to be a general fact. Again, this holds for various societies.[19]

People are willing to cooperate if they expect that others will cooperate, too. How exactly these expectations are formed is not our concern in this paper. In general people must have some sense of who can be trusted and who not. We can assume that such discriminatory power has evolved in human beings. Sometimes, as with any of our faculties, it gives us correct information, sometimes not. Economist Robert Frank has established that given a brief period of personal interaction most people can accurately predict what their counterparts in a one-shot PD are going to do.[20] This mechanism has of course evolved in time, in an evolutionary or repeated game, say, but that does not refute the one-shot nature of the situations that I spoke of here. For most cases we can safely suppose that the set of individuals that were relevant in moulding one's power to assess trustworthiness is disjunct from the set of individuals who figure as players in, for example, the one-shot PD.[21]

7. A MODULE SWITCHED ON: AN ALTERNATIVE EXPLANATION?

An objection to the idea of the loop is that it is overly complicated in some sense, too intricate to be true. One might claim instead that

[19] Camerer and Fehr (2006), Cox *et al*. (2008), Henrich and Smith (2004: 153, 154).

[20] Frank (2004: ch. 2: 'Can Cooperators Find One Another?'). See also economist Brosig (2002), and the work by psychologist Paul Ekman (e.g. Ekman 2003). Compare philosopher Gauthier (1986) on the disposition to cooperate in a PD under conditions of translucency.

[21] More precisely: what is known about the sets of individuals should be disjunct. This is less the case in small-scale societies. One-shot experiments are much more difficult to conduct there. See various contributions in Henrich *et al*. (2004) for the pitfalls and how the field workers tried to overcome these.

https://doi.org/10.1017/S0266267112000259 Published online by Cambridge University Press

cooperative behaviour is more primitive. People simply sense or assume that the people around them are relatively good-hearted, and this makes them behave cooperatively; otherwise they tend to defect. Their mental machinery is relatively crude, unsophisticated. Saying that people act on specific preferences and expectations is really exaggerated talk. Most of the time people switch between a limited set of habits or routines.

Such a view is not implausible; so let us investigate a recent proposal along these lines and see how it compares with the loop. Burnham *et al*. (2000) suggest that humans have a preconscious friend-or-foe (FOF) mental module for assessing the intentions of other people they are dealing with. In this theory, people cooperate because their module subconsciously registers that the others are on the friendly side. A detection of 'friends' switches on cooperation. The authors introduce this FOF module hypothesis because they too find that a significant amount of empirical material demonstrates that the standard predictions of game theory plus self-interest too often fail. The authors reject an explanation in terms of social preferences, however, because they find that this begs the question (why do people have such preferences?) and because it does not very well accommodate the fact that there is noticeable variability in behaviour across different games (it very much depends whether people behave socially or not). Instead, if people detect that somebody can be credited with goodwill and not take advantage, i.e. when he is a 'friend', they will behave cooperatively, while when the converse holds and one's counterpart is a 'foe', they will defect. There are no social preferences involved, in Burnham *et al*.'s view, because the friend-or-foe module-switching is presumed to be driven by self-interest:

> Humans have a contingent strategy, which can be usefully summarized as: if my counterpart is a friend then perform behaviour α. If my counterpart is a foe then perform behaviour β. In either case behaviour is hypothesized to be driven by self-interest, not other regarding utility. If α is a cooperative strategy, and of higher benefit than β, then the FOF detector just alerts you to likelihoods: friend (foe) detection means that you anticipate a higher (lower) probability of positive reciprocity if you follow α and therefore a larger (smaller) gain. (..) FOF detection primes you for greater expected benefits than without it. It sets you up preconsciously for making the maximizing decision.

Burnham *et al*. also note that conscious deliberation can be invoked: when there is conflicting evidence, and a friend later turns out to be a foe. The FOF module discharge is not cast in iron; it does not lead to irreversible commitments. This is an important point, to which we will return below.

I have some things to say about this FOF module and how Burnham *et al*. argue for it. Firstly, advancing the possibility of social preferences upon noting social behaviour is *not* necessarily to beg the question. It isn't,

https://doi.org/10.1017/S0266267112000259 Published online by Cambridge University Press

actually, if one can tell a story of how they work, if one can maintain a conception of these preferences on independent grounds (remember this point from the beginning of section 4).

Secondly, if someone has social preferences this does not mean that this person is supposed to behave cooperatively no matter what, meaning that their cooperation is unconditional. *Of course* this depends on the circumstances, and most notably on what can be expected from one's fellow human beings around one. Their presumed willingness to cooperate conditions one's attitude of expectation-based cooperativeness. The loop, for example, exactly accommodates this contingency.

Thirdly, I find it confusing that Burnham *et al*. say that the FOF module is driven by self-interest. Quite plausibly, if such a module exists, it has arisen because it serves some function. If the authors are suggesting this, I would have no quarrel. Reproductive success, however, need not coincide with maximizing self-interest. For those who thought so in the past, exactly the existence of various social species constituted a riddle. Evolution is about genetic, not individual success.

The crucial point is that playing 'right' at node 2 in the one-shot trust game on page 12 is empathetically *not* according to one's self-interest. Having arrived at node 2 in this game because the first person chose 'down' does not force playing 'down' on the second person. It is perfectly up to him what he decides. At node 2 rational self-interest shouts 'down'.[22] Going 'right' reveals another motivation. Is going 'right' something one does in accordance with one's FOF module? Perhaps, but it would make for an odd answer if one were asked *why* one just played 'right'. 'I chose 'right' because my FOF module drove me so' seems to be incoherent speech. Understanding and explaining human action is inherently bound up with intentionality attributes. Choosing to play 'right' is an intentional action; it involves a desire and a belief (or a preference and an expectation). If one is asked why one has played 'right' in this trust game, it seems quite obvious that the answer relates to the other person. 'I do not want to betray his confidence' or 'I care about his pay-offs too' – something along these lines. These are social preferences.

This is not to deny that there could be some sort of machinery like the FOF module operating in our heads, nicely managing our social behaviour. But it is not an alternative or rival hypothesis on all counts. After all, the output delivered by the module can be input for thought (or further thought). It is not contradictory to say about player 2 going 'right'

[22] There are scholars who try to defend the idea that going 'right' in this one-shot setting is in accordance with one's rational self interest, but they do this by giving up, in effect, the independence axiom of rational choice. For example Gauthier (1986) and McClennen (1990). I assume that Burnham *et al*. do not want to go *this* far.

https://doi.org/10.1017/S0266267112000259 Published online by Cambridge University Press

that the FOF module caused his behaviour *and* that this person wants to live up to the expectations of his fellow player.

I suggest that the loop is not overly complicated. Mutually referring preferences and expectations are appropriate for understanding what happens in trust games, because it seems that exactly under these circumstances the *why*-question is naturally posed. In such situations people typically perceive each other as creatures who do things on purpose, who act on propositional attitudes.

8. DISAPPROVAL AND APPROVAL

As we all know, *not* cooperating when one was expected to predictably evokes anger and disapproval. So isn't *that* the chief motive why people cooperate? In his book *The Economics of Rights, Cooperation and Welfare* (1984/2005), Robert Sugden argues this. He first notes that the actual provision of various public goods in a large society like the UK indicates that social norms must be maintained by something more than self-interest. Sugden explains this by the emotion of resentment: a naturally felt anger towards someone who declines to live up to one's empirically induced expectations. This resentment does not presuppose moral thought. It is a primitive response (it would also count as a Strawsonian reactive attitude). There is also another natural attitude in humans, in turn responsive to this kind of interpersonal anger, namely the tendency to lie low when others display it; and furthermore, to take anticipatory steps when it is in the air, to ward off possible resentment in others. Most people do not want another person being resentful towards them.[23]

Many people, Sugden points out, give tips to taxi drivers they will surely never meet again. Why? It is not merely because taxi drivers need the money; bus drivers need it too, Sugden argues. It is because the taxi driver *expects* a tip – an expectation he has formed by experience. Clients know this, and they also know that the driver will be resentful if he does not receive a tip, and this is something they want to avoid. Hence they tip.[24]

In later work (1998), Sugden elaborates this by distinguishing between material pay-offs like money and immaterial pay-offs that depend on

[23] Cf. the recent literature on 'strong reciprocity'. For example Fehr and Gächter (2002); see also various contributions in Hammerstein (2003).

[24] Two-person cases illustrate how it works but the phenomenon has a wider scope, Sugden argues, because in larger settings, bystanders may typically also sound their disapproval. Why exactly bystanders would want to do this, why there would be third party resentment is another matter, but it can be explained by the fact that there is a fair chance that people will be involved in future dealings with one another (Sugden 1984/2005: 156).

https://doi.org/10.1017/S0266267112000259 Published online by Cambridge University Press

other persons' expectations about one's behaviour towards them. He also demonstrates how actors with these expectations about one another can attain stable cooperation, how mutual normative expectations can develop into equilibrium.[25]

According to Sugden, then, these two attitudes together, resentment and its counterpart avoidance-of-resentment, are crucial ingredients in sustaining human cooperation in mixed motive settings.

Resentment is a powerful attitude that is based on a preceding expectation, but the problem, in my view, lies with the expectation. On what is that based? How has it come about that the taxi driver expects a tip? Suppose that the client does not tip. Then the driver will be angry because he expected a tip. By the loop account, this expectation is based on a generally presumed cooperativeness. Clients who do not tip display a specific fault, namely a lack of this attitude. The taxi driver might even point out this fault to the client. Now let us reconstruct Sugden's theory in this vein. The driver who did not get a tip is angry because he expected a tip. He expected this because people normally tip. And people normally tip in taxis, Sugden says, because they want to avoid the driver's resentment. So the taxi driver would say, in effect, that he is angry because the client did not avoid his anger. Reconstructing Sugden's argument according to the logic of the loop yields an empty reproach.

Sugden argues his case in a different way. He says that the driver's 'expectation is based on his experience of what other people normally do' (155) without an appeal to the reactive attitude. In this situation it is simply the thwarting of an experience-based expectation by the client that elicits anger in the driver. I have proposed, more specifically, that it is not merely a thwarting of his experience-based expectations about another person that causes disapproval in an individual but also that the other person fails to display the corresponding cooperativeness.

How can we distinguish between these two? We can do this by means of a case like the following. One person (A) just acts on his own, parametrically, without any reference to the expectations of other people. Another person (B) gears his actions to those of A but B knows that A is unaware of him doing so. Sugden should then say that a thwarting of B's expectations about A would still provoke resentment in B.

Here is an example. Each working day my neighbour goes to the office with or without an umbrella. I don't bother checking out the weather forecasts, I simply secretly look out of the window and copy his behaviour a few minutes later. If I see him carrying his umbrella, I take mine. If he

[25] See also the afterword in the 2005 edition of 1984, section A7. Compare Battigalli and Dufwenberg (2007) who give a formal treatment that shows that people who want to avoid guilt, which is a measure of one's belief that one lets the other party down, can attain equilibrium.

https://doi.org/10.1017/S0266267112000259 Published online by Cambridge University Press

goes out without it, I leave mine at home too. Then one day he thwarts my expectations: he does not come out of his house (he has taken a day off) and now I must judge for myself the probability of rain. This can make me feel annoyed, perhaps, but not resentful. How could I be angry with my neighbour? What is more, my poor neighbour cannot have anticipated any of this. He does not know about my expectations about his behaviour each morning. So, in a case like this, the expectations are there while resentment is out of place and avoidance-of-resentment even impossible.

As *not* fulfilling someone's expectations in a collaborative project may provoke resentment, positively fulfilling these expectations can yield approval, not just on the part of those having the expectations, but also in bystanders. This, the positive side, is something that Philip Pettit (1995) has emphasized. Pettit argues that people naturally want to be well regarded; they want others to have a good opinion of them. Fulfilling the expectations of a collaborator who relies on you by doing your part is what most people, directly involved or standing by, approve of. However, as Pettit notes himself, seeking such approval is unstable in a sense. A wish to be well regarded is after all not very well regarded itself. Who wants to be known as an audience pleaser? In Jon Elster's words (1983), the good opinion of others seems to be *essentially a by-product*. In the view under consideration approval is a by-product indeed; it derives from the loop: people having expectations based on others' social preferences and people having social preferences based on others' expectations. The client in the cab who has just tipped the driver before stepping out may draw some mild approval from the one who is waiting to get in: he has displayed the expected attitude.

Approval cannot stand alone. Victor Pelligra (2005) elaborates on this by invoking Adam Smith's *Theory of Moral Sentiments* (1759). To begin with, people do not just seek approval and avoid disapproval on the part of others; they also tend to see themselves through the eyes of others. This yields self-approbation and disapprobation: a capacity for self-evaluation. Then this ability becomes supplemented, Pelligra says, with an objective reference point: the *impartial spectator.* The impartial spectator has a cool gaze, he is relatively detached from the particularities of context and the heat of emotional response. He objectifies the voice of the audience. Thus the impartial spectator transforms the earlier desire to be praised into the desire to be *praiseworthy*.

I'd say that praiseworthiness still presupposes the attitude of expectation-based cooperativeness. That is what it is partly *about*. It is he or she displaying this attitude in a situation in which the stakes are high who possibly elicits genuine approval on the part of bystanders. Genuine approval of someone's actions and motives might develop into a more objectified notion of praiseworthiness, perhaps by the dynamics Smith has envisioned. The point is that praiseworthiness cannot emerge from thin

https://doi.org/10.1017/S0266267112000259 Published online by Cambridge University Press

air; it must have an object. The tipping client is praiseworthy, I believe, not because the impartial spectator tells me this – that would be an odd kind of authority argument – this client is praiseworthy for showing the disposition of cooperativeness, a character of sorts.

9. CONCLUSION

The loop can arise under strategic one-shot mixed motive situations, in which one party could let another party down, thereby creating more material gains for himself. In the loop, people act on the reactive attitude of cooperativeness. This is based on the other party's expectation. I have assumed that this attitude belongs to people's normal motivational repertoire and that people normally expect this attitude in each other. At the same time, the loop requires an evolutionary context in which the interaction of cooperativeness and the corresponding expectations can be sustained, in which possible free riders can be spotted and shut out of the game.

So I have defended a conception of social preferences that says that people's willingness to cooperate is based on other people's related expectations. The resulting pattern of behaviour can be stable. It is not a pattern that ensues from people with interlocking expectations who are after their self-interest, as with a coordinative equilibrium, a Lewis norm. It ensues from people with cross-referential expectations and the social preference of expectation-based-cooperativeness – a reactive attitude. We might call it a *moral* equilibrium, a Strawson–Lewis norm.

REFERENCES

Andreoni, J. and J. Miller 2002. Giving according to GARP: an experimental test of the consistency of preferences for altruism. *Econometrica* 70: 737–753.

Bacharach, M., G. Guearra and D. Zizzo 2007. The self fulfilling property of trust: an experimental study. *Theory and Decision* 63: 349–388.

Battigalli, P. and M. Dufwenberg 2007. Guilt in games. *American Economic Review* 97: 170–176.

Binmore, K. 1995. *Game Theory and the Social Contract, Volume I: Playing Fair.* Cambridge, MA: MIT Press.

Binmore, K. 2008. Do conventions need to be common knowledge? *Topoi* 27: 17–27.

Bischoff, I. 2007. Institutional choice versus communication in social dilemmas: an experimental approach. *Journal of Economic Behaviour & Organization* 62: 20–36.

Bohnet, I. and B. Frey 1999. The sound of silence in prisoner's dilemma and dictator games. *Journal of Economic Behaviour and Organization* 38: 43–57.

Bolton, G. and A. Ockenfels 2000. A theory of equity, reciprocity and competition. *American Economic Review* 90: 166–193.

Brosig, J. 2002. Identifying cooperative behaviour: some experimental results in a prisoner's dilemma game. *Journal of Economic Behaviour and Organization* 47: 275–290.

Bruin, B. de 2005. Game theory in philosophy. *Topoi* 24: 197–208.

Burnham, T., K. McCabe and V. Smith. 2000. Friend-or-foe intentionality priming in an extensive form trust game. *Journal of Economic Behaviour and Organization* 43: 57–73.

https://doi.org/10.1017/S0266267112000259 Published online by Cambridge University Press

Camerer, C. and B. Fehr 2004. Measuring social norms and preferences using experimental games: a guide for social scientists. In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, ed. J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr and H. Gintis, 55–95. Oxford: Oxford University Press.

Camerer, C. and E. Fehr 2006. When does 'economic man' dominate social behaviour? *Science* 311: 47–52.

Charness, G. and M. Rabin 2002. Understanding social preferences with simple tests. *Quarterly Journal of Economics* 117: 817–869.

Colman, A. 2003. Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioural and Brain Sciences* 26: 139–198.

Cox, J. 2004. How to identify trust and reciprocity. *Games and Economic Behaviour* 46: 260–281.

Cox, J., K. Sadirai and V. Sadirai 2008. *Experimental Economics* 11: 1–24.

Cubitt, R. and R. Sugden 2003. Common knowledge, salience and convention: a reconstruction of David Lewis' game theory. *Economics and Philosophy* 19: 175–210.

Dufwenberg, M. and G. Kirchsteiger 2004. A theory of sequential reciprocity. *Games and Economic Behaviour* 47: 268–298.

Ekman, P. 2003. *Emotions Revealed*. New York, NY: Henri Holt.

Elster, J. 1983. *Sour Grapes. Studies in the Subversion of Rationality.* Cambridge: Cambridge University Press.

Falk, A., E. Fehr and U. Fischbacher 2008. Testing theories of fairness – intentions matter. *Games and Economic Behaviour* 62: 287–303.

Fehr, B. and S. Gächter 2002. Altruistic punishment in humans. *Nature* 415: 137–140.

Fehr, B. and K. Schmidt 1999. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114: 817–868.

Frank, R. 2004. *What Price the Moral High Ground? Ethical Dilemmas in Competitive Environments*. Princeton, NJ: Princeton University Press.

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

Geanokoplos, J., D. Pearce and E. Stachetti 1989. Psychological games and sequential rationality. *Games and Economic Behaviour* 1: 60–79.

Hammerstein, P. (ed.) 2003. *Genetic and Cultural Evolution of Cooperation*. Cambridge, MA: MIT Press.

Hartogh, G. den 1998. A conventionalist theory of obligation. *Law and Philosophy* 17: 351–376.

Hartogh, G. den 2002. *Mutual Expectations. A Conventionalist Theory of Law*. The Hague: Kluwer.

Henrich, J. and N. Smith. 2004. Comparative experimental evidence from Machiguenga, Mapuche, Huinca, and American populations. In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, ed. J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr and H. Gintis, 125–167. Oxford: Oxford University Press.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr and H. Gintis (eds.) 2004. *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. Oxford: Oxford University Press.

Hume, D. 2000 [1739]. *A Treatise of Human Nature*, ed. D. F. Norton and M. J. Norton. Oxford: Oxford University Press.

Kincaid, H. 2008. Structural realism and the social sciences. *Philosophy of Science* 75: 720–731.

Lewis, D. 1969. *Convention. A Philosophical Study*. Oxford: Basil Blackwell.

McCabe, K., M. Rigdon and V. Smith 2003. Positive reciprocity and intentions in trust games. *Journal of Economic Behaviour and Organization* 52: 267–275.

McClennen, E. 1990. *Rationality and Dynamic Choice. Foundational Explorations.* Cambridge: Cambridge University Press.

https://doi.org/10.1017/S0266267112000259 Published online by Cambridge University Press

Orbell, J., A. van de Kragt and R. Dawes 1988. Explaining discussion induced cooperation. *Journal of Personality and Social Psychology* 40: 45–51.
Pelligra, V. 2005. Under trusting eyes: the responsive nature of trust. In *Economics and Social Interaction. Accounting for Interpersonal Relations*, ed. B. Gui and R. Sugden, 105–124. Cambridge: Cambridge University Press.
Pettit, P. 1995. The cunning of trust. *Philosophy and Public Affairs* 25: 202–225.
Pettit, P. 1996. Functional explanation and virtual selection. *British Journal for the Philosophy of Science* 47: 291–302.
Rabin, M. 1993. Incorporating fairness into game theory and economics. *American Economic Review* 83: 1281–1302.
Ross, D. 2008. Classical game theory, socialization and the rationalization of conventions. *Topoi* 27: 57–72.
Roth, A., V. Prasnikar, M. Okuno-Fujiwara and S. Zamir 1991. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: an experimental study. *American Economic Review* 81: 1068–1095.
Sally, D. 1995. Conversation and cooperation in social dilemmas: a meta-analysis of experiments from 1958 to 1992, *Rationality and Society* 7: 58–92.
Schelling, T. 1960. *The Strategy of Conflict.* Cambridge, MA: Harvard University Press.
Schelling, T. 1978. *Micromotives and Macrobehaviour*. New York, NY: W.W. Norton & Co.
Sillari, G. 2005. A logical framework for convention. *Synthese* 147: 379–400.
Sillari, G. 2008. Common knowledge and convention. *Topoi* 27: 29–39.
Smith, A. 1759. *The Theory of Moral Sentiments*, ed. D.D. Raphael and A.L. Macfie. Oxford: Oxford University Press.
Smith, V. 2010. What would Adam Smith think? *Journal of Economic Behaviour and Organization* 73: 83–86.
Sober, E. 1983. Equilibrium explanation. *Philosophical Studies* 43: 201–210.
Strawson, P. 1974. Freedom and resentment. In *Freedom and Resentment and Other essays.* London: Methuen.
Sugden, R. 1984/2005. *The Economics of Rights, Cooperation and Welfare*. New York, NY: Palgrave MacMillan.
Sugden, R. 1989. Spontaneous order. *Journal of Economic Perspectives* 3: 85–97.
Sugden, R. 1998. Normative expectations: the simultaneous evolution of institutions and norms. In *Economics, Values, and Organization*, ed. A. Ben-Ner and L. Putterman, 73–100. Cambridge: Cambridge University Press.
Weber, J. and J. Murnighan 2008. Suckers or saviors? Consistent contributors in social dilemmas. *Journal of Personality and Social Psychology* 95: 1340–1353.
Wilson, B. 2010. Social preferences aren't preferences. *Journal of Economic Behaviour and Organization* 73: 77–82.

https://doi.org/10.1017/S0266267112000259 Published online by Cambridge University Press