games

MDPI

Article

Social Preferences and Context Sensitivity

Jelle de Boer

Philosophy Section, Delft University of Technology, Jaffalaan 5, 2628 BX Delft, The Netherlands; j.s.deboer@tudelft.nl

Received: 28 July 2017; Accepted: 6 October 2017; Published: 13 October 2017

Abstract: This paper is a partial review of the literature on 'social preferences'. There are empirical findings that convincingly demonstrate the existence of social preferences, but there are also studies that indicate their fragility. So how robust are social preferences, and how exactly are they context dependent? One of the most promising insights from the literature, in my view, is an equilibrium explanation of mutually referring conditional social preferences and expectations. I use this concept of equilibrium, summarized by means of a figure, to discuss a range of empirical studies. Where appropriate, I also briefly discuss a couple of insights from the (mostly parallel) evolutionary literature about cooperation. A concrete case of the Orma in Kenya will be used as a motivating example in the beginning.

Keywords: social preferences; game theory; ethics

1. Introduction

Among the Orma in Kenya, cooperation in a one-shot anonymous Public Goods experiment, which has a Prisoner's Dilemma (PD) structure, correlates with wealth, i.e., number of cattle. People with cattle made relatively high contributions in this game, but not so in other games like the Ultimatum Game (UG). On the other hand, Orma without cattle—wage laborers and tradesmen—demonstrated the opposite pattern by making high contributions in the UG and low contributions in the Public Goods experiment. It appeared that the Orma generally associated the Public Goods experiment with a *Harambee*, a Swahili word for fund-raising for local public projects, such as building a school. In a Harambee one is supposed to contribute in proportion to wealth. That cooperative play among the Orma in the PD experiment has to do with the institution of the Harambee in their daily lives seems fairly obvious, but exactly how is not so clear. After all, the experiment is anonymous and one-shot, whereas the Harambee is more like a repeated game, and is therefore a different kind of strategic situation.

This case of experimental gameplay by the Orma derives from a highly influential body of research done by anthropologists and economists in fifteen small-scale societies, which started in the late 1990s and is still underway [1,2].[1] Two of the major outcomes so far are, firstly, that the behavior of most experimental subjects, wherever they live, contradicts the *Homo Economicus* model. This model predicts universal defection in the PD (with monetary outcomes) and minimal offers for proposers and subsequent acceptance by responders in the UG (with monetary outcomes). Secondly, cooperative play in these games varies quite a lot. It varies much more, in fact, than has previously been established by experimental work done in industrialized countries [4].

The aim of this paper is to help in understanding this remarkable variation in cooperative behavior. My method will be rational reconstruction. I will attempt to interpret empirical results in terms of propositional attitudes and reasons. The Orma provide an exemplary case because they manifest two distinct patterns among themselves. I will start by focusing on the Orma and the

[1] The particular study of the Orma is [3].

associated Harambee and discuss various hypotheses to explain this case (Sections 2 and 3). In Sections 4 and 5, I review parts of the literature on 'social preferences'. I will focus on the class of expectations-based social preferences and I will portray the corresponding equilibrium of these preferences and expectations by means of a figure—a loop. These mutually referring social preferences and expectations can explain a large part of the contingency of cooperation as observed by the anthropologists. Thereafter I will move beyond this and attempt to apply the central insights of this literature more generally (Section 6).[2]

2. The One-Shot Game Design

In this section I discuss a couple of alternative hypotheses that explain the cooperative behavior, before moving on to the social preferences hypothesis itself. Readers already convinced about the existence of social preferences and how they can be established, and also familiar with the methodological issues around this, may wish to skip this section and go directly to Section 3.

The first hypothesis we will consider says that the Public Goods or PD experiment may be designed and conducted as an anonymous one-shot game but that the subjects do not really perceive it in this way. It is too artificial for them; they see this situation as just another moment immersed in their ongoing world, naturally connected with their personal and shared histories and with future consequences.

In recent years there has been some discussion among economists whether one-shot games are good experimental designs at all, exactly because in real life, outside of the laboratory, the actions that people perform and the choices that they make are normally elements in a social historic fabric. Nobel prize-winning economist Vernon Smith, for example, has doubts about the usability of such games in experiments: "The abstract concept of single play invokes conditions sufficiently remote from much human experience that it may be operationally difficult to penetrate." He asks whether a one-shot game is really "devoid of a history and a future," and more specifically whether it is "devoid of reputation considerations" [12].[3]

One way in which Vernon Smith could be right is of course when the experiments fail to be realized according to their protocols. Players in a game like the PD should not be able to see what others contribute, communicate their ideas and intentions to their co-players, influence other players in subsequent experiments (contagion), or be subject to experimenter's bias. Safeguarding this can be troublesome in small-scale societies and it is interesting to read how the various researchers tried to cope with this. Sometimes they failed, but the anthropologist who studied the Orma, Jean Ensminger, appears to have anticipated the pitfalls and to have been able to control the anonymous one-shot nature of the game quite well. For example, contributions were made privately in envelopes which were then shuffled. People were not allowed to talk. Ensminger raced from village to village in order to beat travelling news and thus avoid contagion. She asked her native research assistant to turn his back when offers and contributions were made. Therefore, let us assume that there were no serious leakages and that the experiment mostly worked as planned.

However, Vernon Smith's qualm goes further than this. Because even when everything is executed as it ought to be, experimental subjects can, in their thoughts, still reach beyond the laboratory walls and let the real world enter their minds: they can still *imagine* reputation issues. Contemplating defection in the anonymous one-shot PD experiment, an individual may still feel the presence of his fellow men and think that defection will somehow be bad for his name. That would be a piece of erroneous thinking according to the logic of the game but it can occur. Now the common way to test this is to compare how the same people perform under the condition of repeated play with how they meet single play. Such tests show that the vast majority of subjects play systematically differently under these two conditions. This reveals that they understand reputation building [16]. Smith says he knows the experimental literature that attests that people play

---

[2] Other philosophical work that also discusses this and/or similar material is [5–11]. Bicchieri's (2006) book [5] is influential but in Section 3 I argue that it is ultimately on the wrong track.

[3] For the same doubts see [13–15].

differently in repeated games than in one-shot games, but he claims that "this tells you only that reputation-building is more important in repeat play of the same game than in single play, not that it is absent in one-shot games." However, now reputation and its workings seem to become a bit too elusive. How could one possibly refute the reputation effects that according to Smith may linger on in single play? Smith says that one-shot play can harbor reputation building, something that one would only expect in repeated play. However, if repeated play then clearly demonstrates a difference which we would normally attribute to considerations of reputation, then it seems there is little room left still to hypothesize these considerations of reputation in single play.

To uphold this claim would require other and independent empirical support. Now Smith indeed also defends his thesis by such work, viz., an experiment conducted by Mary Rigdon, Kevin McCabe, and himself [17]. This experiment showed that people in an anonymous repeated PD play substantially more cooperatively when the experimenters secretly cluster the cooperators as the experiment proceeds. The experimenters observe who cooperate in the early rounds of the game and subsequently match individual players according to how cooperatively they have played until then. In the control group players are randomly matched. In both conditions play is anonymous and no one knows about the sorting. "The trends are unmistakable," Rigdon, McCabe, and Smith say, "as play proceeds through the later rounds, cooperation emerges and is sustained among the sorted subjects, but there is no similar round-effect for the randomly paired subjects."[4]

These subjects in the sorting group play a series of one-shot games without having a history as in repeated play, where people can monitor each other, but there has nonetheless been a relevant common causal history. This is because what happens in early rounds appears to affect behavior in later rounds. Reflecting on this, Smith asks:

> Why should a real person see no continuation value across stage games with different but culturally more or less similar strangers? Can we ignore the fact that each person shares cultural elements of commonality with the history of others? (…) Is not culture about multilateral human sociality?[5]

This is interesting but I find it unclear how this supports the earlier idea that *reputation* considerations are effective in single play. It more seems like Smith indicates something different at work here, namely, indeed, a kind of "human sociality". The following sections aim to develop a way to understand this sociality in terms other than reputation.

3. Social Preferences, Expectations and Scripts

In this section I examine the hypothesis of conditional social preferences as an explanation of cooperative behavior. These social preferences are conditional in the sense that they are based on expectations about whether the other players in the game-theoretic setting are going to cooperate. The next question is then on what these expectations can be based in turn. I discuss and then reject a theory that claims that these expectations are ultimately based on 'scripts'.

Another way to explain cooperative behavior in an anonymous one-shot PD experiment is invoking a social preference, e.g., a preference for promoting the well-being of the other players in the game or a preference for fair dealing. People with such preferences can attain the cooperative outcome in a PD (with material payoffs). In the synthesizing and interpretative chapter of their book the anthropologists discuss the possibility that people perhaps have a general inclination to behave socially, which they understand as a disposition that has evolved in a particular physical–social environment (e.g., among the horticulturalists Quichua in the Amazon tropical forest) and which then generally applies to various types of mixed-motive settings within that environment (e.g., PD, UG, etc.).

There is a growing consensus among social scientists, including economists, that people regularly act on genuine social inclinations, and that the thesis of universal egoism is false. See the

---

[4] [17] p. 997.

[5] [12] p. 9.

recent empirical literature on 'social preferences' [18–23]. One kind of social preferences is general or categorical. However, Henrich et al. reject the idea that people have a generalized disposition to behave socially and they do this exactly on the basis of Ensminger's fieldwork among the Orma. A general social disposition in game-theoretic experiments would predict that people play cooperatively in a PD *and* make high offers in a UG. The Orma, however, are divided on this point. The Orma with cattle play cooperatively in the PD and not in the UG, whereas Orma wage laborers play cooperatively in the UG and not in the PD. This does not undermine the hypothesis that people can have social preferences of some sort but it does undermine the specific idea of general or categorical social preferences.

There is now ample evidence that people's willingness to cooperate in experiments like the PD largely depends on them having expectations that the others are going to cooperate, too. Not many people are saints and have unconditional social preferences in such settings. Most of the time social preferences appear to have a conditional nature: I will if you will. The reason is that people want to avoid exploitation. They don't want to be suckered. In other words, they only want to behave cooperatively when they are in good company, when it can be expected that the others are also going to cooperate. So an actual willingness to cooperate is reasonably accompanied by an expectation that others are going to cooperate, too [22,24,25]. It is this relatively abundant type of social preferences that is the subject of the present paper.

As said, these conditional preferences are dependent on expectations. This brings us beyond standard game theory and into the territory of *psychological* game theory. In psychological game theory, expectations have a direct causal impact on the motivation of the players. Expectations become part of utility. Seminal contributions are Rabin (1993) [18] and Charness and Dufwenberg (2006) [26].[6] Rabin calls this motivation 'reciprocating kindness' and Charness and Dufwenberg call it 'guilt aversion'. But it is important to note that their actual mathematical models are silent on the precise moral or altruistic content of the preferences. The equilibrium models of these authors simply do not really make assumptions about this content. This is a good thing in a sense; it makes the theory more general.

So generally, conditional social preferences are special because, with these, a subject's expectations about his fellow players not merely make a difference for his strategy choice but for his utility. Such expectations thus transform a game, e.g., they could transform a Prisoner's Dilemma into a Stag Hunt game, when the expectations are mutual.

The suggestion that follows from this is that the Orma who play cooperatively in the PD experiment do so because they have a social preference that is conditioned by their expectation that the other players will also cooperate. The next question is: where do these expectations come from? Remember that we are working under the assumption that the Orma understand the difference between repeated and single play. Let us then also assume that this understanding is mutual among them in the sense that the subjects also understand that the others understand that they are in a single play situation, i.e., that there is common knowledge to a sufficient degree. This means that their possible expectations that the other players are going to cooperate cannot really be based on considerations that belong to repeated play. They are mutually aware of the fact that the PD experiment is *not* a stage in a Harambee. So why would one then expect one's fellow players to cooperate in the one-shot game? On what can this expectation be reasonably based?

In her 2006 book *The Grammar of Society* [5] the philosopher Christina Bicchieri argues that 'scripts' ultimately justify people's expectations in social interaction.[7] In this section I show that this is not a good way for justifying expectations. Cognitive scientist Roger Shank introduced the term 'script'. Scripts are micro personal theories about how the social world works. They contain roles and tell us what to expect from each other. For example, the script 'restaurant' sets in motion ideas that somebody who approaches your table in a restaurant will offer a menu, then take orders, bring

---

[6] See also [27,28]. For an early contribution in philosophy see [29,30] and cf. [10].

[7] Another prominent scholar who develops an expectation-based theory is Robert Sugden [31]. See De Boer [10], Section 8, for a discussion of this work.

drinks and food, and will finally require payment. Scripts, Bicchieri argues, also help to form expectations in mixed-motives situations. Once a situation is categorized as being of a certain type, a script is activated that will involve players' interlocking roles, a shared understanding of what is supposed to happen, and even prescriptions for unexpected occurrences. For example, once a particular 'fair division' script is activated, an individual will have definite beliefs and expectations about other individuals she is interacting with, even if (or especially if) she does not know such individuals personally.[8] Bichieri also stresses that scripts mostly do their work below our level of consciousness.

> What we have instead is implicit, nondeclarative knowledge. (…) Once a schema is activated, we tend to follow the norm by default.[9]

This analysis seems to fit well with what the anthropologists say in another (and favored) interpretation of their experimental results, apart from the interpretation with the evolved general disposition that we just discussed. The anthropologists argue as follows. For the Orma, as for most people, the laboratory experiment is somewhat alien: it is detached from normal daily life. Therefore, it is not immediately clear what to do, and in such situations people scan their memories for resemblances. The mechanism is that they search for analogue situations from their daily lives that they have already experienced. The PD lab experiment cues a particular analogue from daily life and this association subsequently triggers behavior in the experiment.[10] The PD reminds the Orma of the Harambee institution. As the Harambee implies contribution dependent on wealth, the amount of one's own wealth guides individual choice in the PD experiment. In terms of Bicchieri's scripts, the PD activates the Harambee script and this then induces cooperative behavior in Orma in proportion to their stock of cattle.

We are examining the hypothesis that the Orma act on a social preference which is conditioned by the expectation that the others in the experiment will cooperate to a sufficient degree. We have assumed that the participants in the game understand the difference between single play and repeated play. The question we are now dealing with is why players would expect others to play cooperatively in a situation of single play. On the script theory the explanation would be that there is a Harambee script which tells one what to do and what to expect from others: cooperate in proportion to wealth and so will the others.

The aim of this paper is the same as Bicchieri's: to offer a rational reconstruction of social interaction. This means analyzing social behavior in terms of propositional attitudes and reasons. Using this perspective, we may always ask the 'why?' question. At this point it can be asked why a socially inclined Orma would expect others to follow a script in these circumstances. After all, it is mutually understood among the players, we have assumed, that they are currently *not* in a Harambee and that each can get away with defection. Why expect others to follow a script in these circumstances? Why expect them to fare on something like an automatic pilot when there is a quite obvious reason not to do this? Single play can be seen as a novelty and this requires that one takes back the controls.[11] A script delivers a freeze-frame image of the other players in a game. The others are actually envisioned as states of nature, not as rational agents. People who act on scripts effectively transform a game-theoretic situation into a decision-theoretic one.

4. Social Preferences and Expectations in a Loop

Let us take stock. A first part of the explanation for a high degree of cooperation in single play in mixed-motive games is that people act on genuine social preferences. Secondly, these social preferences must not be understood as categorical. This is because they are often based on the

---

8 [5] p. 94.

9 [5] p. 97.

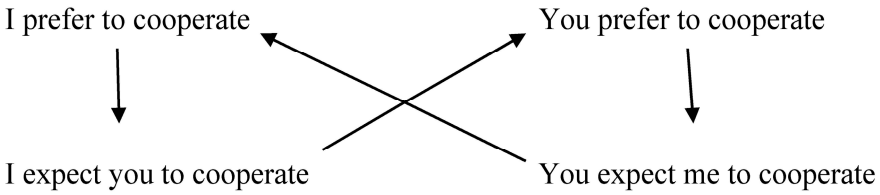
10 [1] p. 48, 49.

11 Expectation failure is a general problem for the script theory, as Roger Shank himself acknowledges, e.g., http://www.rogerschank.com/biography.html.

expectation that there is a fair chance that others in the game are also going to cooperate. Thirdly, let us assume that such expectations are rational throughout, i.e., based on a view of other people as rational creatures just like oneself, with propositional attitudes that are not seen as fixed but contingent.

In this section we examine such a view, viz., an equilibrium conception of mutually referring social preferences and expectations. After showing how it works, I will briefly compare it with a biological type of explanation of cooperation and then a decision-theoretic one.

Why would someone who is inclined to cooperate in a one-shot PD in material outcomes expect his co-players to cooperate? The general answer from the expectation-based social preferences literature is this: because he believes that the other players are similarly inclined. He *expects that the others also have a social preference*, which is of the same conditional kind. This connects the two loose ends that we started with. Suppose you and I are in a single play PD in material outcomes. Then our propositional attitudes connect like this: My preference to cooperate is based on my expectation that you will cooperate. My expectation that you will cooperate is based on your preference to cooperate. Your preference to cooperate is based on your expectation that I will cooperate. Your expectation that I will cooperate is based on my preference to cooperate. Our mutually referring propositional attitudes make up a loop. There are no loose ends; the reasoning is closed. Schematically:



The arrows indicate the 'based on' relationship which is causal and justificatory, in the ideal type case. For each arrow, what is formulated at the point of the arrow is an answer to a 'why' question about what is formulated at the base of the arrow. As mentioned, the preferences and expectations of the individuals involved in this situation make up a closed loop. By being based on each other these attitudes are in an equilibrium state. The fact that the attitudes of the individuals connect in this way gives the situation a special character. It makes it free-floating.

This model differs from Bicchieri's theory which analyses human interaction in the same terms—propositional attitudes and reasons—but which grounds the expectations in static scripts in individual minds that do not track other people's attitudes, while the loop connects people's expectations to the preferences of their companions.[12]

The scheme is general in the sense that the social preferences can be of various kinds. For example, I may have a social preference that is about inequity aversion. I want to avoid large pay-off differences between you and me (but with the clause 'only if I expect that you are not going to defect on me,' it is conditional). You, on the other hand, can have a social preference that is about fulfilling my expectations. You think that I expect you to cooperate and you want to honor that expectation, you don't want to let me down. Or I think that you are in need of some more money while your motive has to do with fairness. Many combinations are possible. What exact form a social preference has or where it comes from does not matter for the possibility of equilibrium in expectations. As long as the preferences are other-regarding and conditional on the other's assumed cooperation, the loop goes round.

To clarify a bit further the distinct character of these looping reason-based attitudes, let me contrast how they mesh with a typical biological explanation and subsequently with a decision-theoretic one. A biological explanation standardly proceeds in terms of organisms with their traits that adapt in response to a changing environment. In this view, something like social

[12] I do not deny that this model can be criticized at a deeper level. See, for example, (other) work by Robert Sugden [32] and by John Davis [33,34]. In this paper I simply build on the assumption that human behavior can be adequately explained in terms of individual preferences and expectations. This corresponds with main stream economics, most notably of course micro economics and game theory.

inclinations is seen as an outgrowth of a biological past. We are social animals. We have evolved from primates who lived in groups. It is uncontroversial that our social inclinations derive from this long past of living together. As discussed, Henrich et al. [1] considered a hypothesis in this vein, namely that a tendency to play cooperatively in mixed-motives games could have arisen from a particular physical–social environment. They went on to reject this idea on the grounds that the Orma's context sensitivity provided a clear counter-example. But let us now carry on with the biological take on this issue, which goes in terms of organisms in environments. Against Henrich et al.'s rejection, a scholar steeped in evolutionary thought might argue that the concept of 'physical–social environment' includes something more. Because one way or another, he says, it must be *something* in the environment, sufficiently broadly understood, that has produced a difference in cooperativeness among similar creatures. This is true, but if we follow this line of reasoning we must note that two things are quite special. One has to do with the organism, the other has to do with the environment.

The organism, firstly, is special because it is a creature that can learn and therefore adapt during its *own* lifetime. It does this by modifying dedicated internal states, i.e., intentional states, most notably his expectations and preferences. This addition moves the explanandum into the (overlapping) realm of decision theory. Secondly, the environment is also special. It (partly) consists of similar creatures that can *also* learn and have expectations. This further moves the explanandum into the realm of game theory.

In such a world, with agents who can think and learn and who are not too myopic, a behavioral pattern like human cooperation in a mixed-motive setting cannot be fully explained in typical biological terms of features of organisms that evolve in a changing environment. Also, comprehending the individual in this world as a rational creature but its environment still as ultimately fixed is incomplete. Humans are rational social creatures with *mutual* expectations and preferences. We all live in certain physical–social environments. But these environments may crucially differ in regard to having people around with particular expectations about one's action or not having such people around.

Note that this conditional social preference is not open to exploitation, because it essentially operates by shutting free riders out. This attitude can only evolve in evolutionary time when there is 'positive sorting': creatures with the underlying disposition must be able to find each other. Hamilton already saw that cooperation can evolve without common descent:

> kinship should be considered just one way of getting positive regression of genotype in the recipient, and that it is this positive regression that is vitally necessary for altruism. (…) in the assortive-settling model it obviously makes no difference if altruists settle with altruists because they are related (perhaps never having parted from them) or because they recognize fellow altruists as such. [35]

And, of course, kinship selection itself depends on a kind of recognition. The general thing that must happen in a risky cooperative setting is positive assortment: cooperators must be able to find each other or be brought together somehow. This can happen in a number of ways, for example by reciprocity, indirect reciprocity, signaling, environmental feedback, or by spatial structuring.[13] (Or by outsiders who are in the know, like an economist conducting an experiment, or a leader organizing people in groups). We will return to this in Section 5.

[13] See Nowak et al. [36] on positive assortment or clustering in general. On how signaling and spatial structuring can help in simulated cooperative games like the PD and the Stag Hunt, see Skyrms [37–39]; cf [40]. Economist Robert Frank has found positive evidence of sufficiently reliable signaling in human subjects in public goods games; see his "Can Cooperators Find One Another?" [41]. Cf. the work by Paul Ekman on human signaling through emotions and their corresponding facial expressions, [42]. The political philosopher David Gauthier makes use of the notion that cooperators can find each other, to a sufficient extent, in his book Morals by Agreement [43]. For some references on indirect reciprocity, see note 25.

Presently this disposition can persist, at least in part, by means of mutually referring expectations and preferences—a sophisticated version of signaling, we might say.[14] Such mutually referring mental states make human cooperation *both* rationally and evolutionarily stable. This social preference is rooted in a social disposition that has evolved over time and which is sensitive to sufficient numbers of other people being around with the same disposition. This sensitivity in rational humans consists of mutually referring expectations and preferences.

Not everybody has this social disposition or has it to the same extent—in *Homo Sapiens* there is arguably a variety of behavioral types—but people with a good degree of it have a method to foster cooperation among themselves and ostracize selfish defectors.

5. The Loop in Real Life

The suggestion is that cooperatively inclined people in a single play PD believe that their co-players are similarly inclined. But how do they know *this*? There can be various grounds for this belief. Perhaps a third party was a good source and has informed them about each other. Or the subjects have managed to reliably signal their intentions back and forth. Or they know this on the basis of experience: when they have a shared history of dealing with one another. This is what this section is about: how social preferences in an anonymous one-shot experiment relate to real-life cooperation.

A good example of the possibility of sharing a history of cooperation is, of course, the Orma's Harambee. I submit that the Harambee can evidence social preferences among the participants in two ways. Firstly, this real-life institution is an instance of *imperfect* repeated play. The group of participants is not fixed, as in typical laboratory repeated play, but naturally varies in size and in composition. People move in and out. Most importantly, for present purposes, there is no full transparency. Monitoring others and keeping track of their past behavior is limited. Under laboratory conditions the situation is relatively simple. It can be assumed that all subjects know the game form and how the game will proceed. Also, each knows that everybody else knows this, and that each knows that each knows that everybody else knows this, and so on. In the lab it can be assumed that there will be a nice symmetrical structure of mutual beliefs supporting a possible equilibrium outcome. Real life, in the meantime, is replete with information asymmetries. Some people have more information, others less. Or some people are in the know about something while others have false beliefs. More specifically, one may reasonably get the impression that the usual monitoring is lacking while it accidentally isn't. One then falsely believes that one confronts a pocket of single play. If one's fellow players understand one's perspective and still observe this individual contributing, they have reason to conclude that this individual is led by a social preference.

Such circumstances are telling: this is how we normally catch cheaters. A liar believes he can get away with a certain lie and then makes a mistake, with the result that he is noticed. The same circumstance is therefore appropriate to infer that someone is good-hearted. Such instances of epistemic asymmetries can thus demonstrate social preferences.

During repeated play these social preferences are perhaps not necessary as reputation may sufficiently bolster cooperation but that does not imply that these preferences are not around. It would be odd to assume that social preferences make themselves felt when defection pays and leave the scene when it does not. It seems more plausible that motives to cooperate are overdetermined during sequences of repeated play. Such overdetermination is often functional. In nature and also in complex artifacts redundancies are commonplace. They are especially appropriate in a turbulent or risky environment. If one mechanism fails, another can kick in.[15] This happens when repeated play is punctuated by single play, when other-regarding motivation causes cooperation at a moment that reputation has become irrelevant.

It has often been argued that it is difficult to extrapolate from results that indicate social preferences in one-shot experiments to real-life situations when real life resembles a repeated game

[14] See De Boer [44] on the relation between signaling and mutual beliefs.

[15] Compare this to the idea of "modal robustness" as developed by Alfano and Skorburg [45].

[8,12,15,46,47]. This is because, in a repeated game, self-interested preferences cum reputation considerations can sustain cooperation. This is true, but in my view it is better to characterize the Harambee as an *imperfect* repeated game. Social preferences can manifest themselves both in laboratory single play and in repeated play with epistemic asymmetries.

This is the time to return to Vernon Smith's experiment, which we discussed in Section 2, with the sorting unknown to the participants in one of the conditions. In the first round of this experiment cooperation is around 50%. In the normal randomized group cooperation in the subsequent rounds becomes less, as it usually does in such series. On the model of conditional social preferences we can understand this as follows. A number of subjects who were cooperatively inclined at the beginning meet a defector in the second round and now lose their faith in their counterparts, which causes them to defect in the following round. This then of course further reduces what was left of expectations of socially motivated cooperation. And so on. In the (unknown) sorting group, on the other hand, the opposite happens. Hence, a sizeable group starts out with social preferences accompanied by some good faith in their counterparts, and this increases in the following rounds thanks to the sorting. Through the rounds, these subjects' expectations that the other players will act on their social preferences become gradually reinforced.

A second argument that social preferences are active during repeated play is that people typically do not just answer defection by equally defecting in the next round. They do not simply turn their backs on the others. They also get angry; they *resent* what the other has done. Why? This anger is not just a blind impulse: it is directed at the other person and has a reason. The other has failed. There is something wrong with his attitude. He proves to be selfish, i.e., lacking the social preference. From such a response upon incidental defection in a repeated game we may conclude that a social preference is normally expected.

So cooperative behavior during epistemic asymmetries and occasions of disapproval upon defection indicate that social preferences are also at work during real-life repeated play. Presumably, social preferences will ripen through repeated play. How would that work? I hypothesize that this will go as follows. I put it somewhat schematically. Firstly, let us assume that at the initial stages the Harambee indeed approximates repeated play and that epistemic asymmetries are negligible. The social disposition is sufficiently widely spread in the population but its activation is just moderate or low. Repeated play will give people sufficient reason to cooperate. Also assume that cooperation occurs often so that people meet quite frequently. (Of course, meeting each other also happens outside of the Harambee, but let us keep things simple and stick to a model story in which all interaction occurs through the Harambee.) I conjecture that when people go through the rounds their social preferences towards each other will gradually develop and become stronger. This is, in itself, an arational process but we are all familiar with it. Sitting in the same class and doing the same things are conducive to sympathy. How else do people become friends? Sharing experiences and being together breed affective ties in humans. So it seems plausible that successful repeated play can foster mutual affection. Individuals will not develop these attitudes by themselves but by extrapolating, or perhaps they will also suppose these attitudes to be developing in others on the basis of subtle signaling. Later, when the Harambee becomes standard and probably somewhat larger in scale, imperfect repeated play can then show the workings of social preferences.[16]

In these ways, imperfect repeated play that is already in place can bolster social preferences in newcomers. Rounds of repeated play grow social preferences in people who are thus disposed. Newcomers notice that people are generally not taken advantage of during situations that are not fully transparent. They learn to expect a social attitude in others and this gives them a ground to

[16] For a brief historical account of the development of the Harambee, see Waithima [48], the introduction.

enact their own social attitude, if they are thus disposed. In this way, institutions can scaffold[17] social attitudes.[18]

Playing cooperatively because one believes that the other person has been cooperative in the past, since one shares a history of cooperation with other people, and the other person is a member of that group, is indirect reciprocity. One then cooperates because the other has cooperated in the past, not necessarily with oneself, but with others. To arrive at a stable norm of cooperation, like the loop, information about past behavior must have been travelling through the group. Who has been good and trustworthy, and who has been cheating? Gossip—third party information about reputation—is an important mechanism to bolster cooperation. From a certain point on, presumably, this information can translate into a generalized expectation that any other player from the same group, without knowing his exact past behavior, will be trustworthy. Of course, specific information about cheating individuals will remain highly functional, but cooperators may be relatively relaxed on information gathering and perceive unknown others as simply group members, who are part of a long history of equilibrium play and can therefore be trusted. So an Orma may play cooperatively in an anonymous one-shot PD game because he knows that the other players are Orma too, who probably have cooperated a lot in the past, as many Orma have.[19]

At the same time, it is important to see that this is not the only mechanism to sustain the psychological game-theoretic equilibrium, as portrayed by the loop. Social preferences and indirect reciprocity arguably bridge the gap between the anonymous one-shot PD and real-life cooperation. But real-life cooperation can also be facilitated in other ways. The expectations and preferences in the loop can, of course, also originate from *direct* reciprocity, when two individuals engage in repeated play with each other. A repeated game does not *require* the presence of social preferences, as argued above, but it does make them likely. Another possible source is signaling. Then one individual sends a reliable signal about one's commitment to cooperate, for example, through facial expressions that are hard to fake or by other telltale clues.[20] Another way, of course, is

---

[17] I borrow this wording from Francesco Guala who wrote: "More effort should be made in investigating how non-costly sanctions, backed up by adequate institutional scaffoldings, may be used to sustain positive reciprocity in a variety of real world settings" [6] p. 15).

[18] But doesn't this Harambee history then function much like a script (Section 3)? No, not as Bicchieri understands this. It is true that for the equilibrium something like a script can function as a first mover, so to speak, a way to enter the loop. But remember that for the equilibrium model of explaining and justifying cooperative behavior it does not matter where exactly a social preference comes from (Section 4). That is exactly the point about an equilibrium. Inside the equilibrium, inside the loop, the explanation and justification run in terms of the propositional attitudes that are in place at that very moment. What do I expect from my fellow player? I expect that she is going to cooperate. Why? I expect that she will act on a social preference. Why would she? Well, presumably she expects that I will cooperate. Why would you? I will cooperate because I am socially inclined myself and I expect my fellow player to cooperate. And now we are back where we started. The argument is a (virtuous) circle. Bicchieri's script theory omits this equilibrium part or the explanation, how the social preferences and expectations mutually refer to one another. Put differently, on the present view (1) a script would not count as a reason, a way to justify cooperation, and (2) it is a possible part of the explanation but not a necessary part, as, for example, signaling would be another possibility.

[19] To be more precise: a wealthy, cattle owning Orma may play cooperatively in an anonymous one-shot public goods game because he knows that the other players are Orma too, enough of them cattle owners, who have cooperated a lot in the past. This should explain, at least partly, the heterogeneity among the Orma themselves, with the cattle owners playing cooperatively in proportion to their wealth in this experiment, and the non-cattle-owners making low contributions. I must admit that I cannot tell this on the basis of the empirical evidence. An experiment to check this would be to inform the players about the cattle-owning status of their fellow players. According to the model with conditional social preferences, cattle owners will make substantially higher contributions in a one-shot game with other cattle owners and will do so less when they hear that they are on their own, so to speak, only with full-time laborers and tradesmen.

[20] For the importance of indirect reciprocity for human cooperation, see the pioneering work of Alexander [49]. Sugden [31] developed a model in which people keep track of each others' "standing"; cf. Nowak and

communication. Talking makes a difference and exchanging promises even more. And some people perhaps expect others to cooperate by default—only adjusting when proven wrong. So people do not *need* to share a history. Equilibrium play can get started in various ways. It is fortunate that there are various routes to make a good guess about somebody's cooperative inclinations. After all, histories must start somewhere.

6. Context Sensitivity and Robustness

In this section I will apply the scheme of the loop more broadly, to a variety of empirical findings. Economists Levitt and List [13] have pointed out that, overall, the empirical findings on social preferences are fairly mixed and should therefore be treated with caution. They cannot be extrapolated to financial markets, for example:

> It seems highly unlikely, for instance, that at the end of a day's trading, a successful trader would seek out the party that was on the wrong side of a market move and donate a substantial fraction of the day's profit to the person who lost—even though parallel behavior is routine in certain experiments.

Levitt's and List's conclusion is that social preferences may crop up in some experiments here and there but that they are really quite fragile, and very much dependent on the particularities of the situation. Such context sensitivity, they argue, is also demonstrated in framing experiments. Indeed, many experiments have shown that cooperation in a mixed-motive game can be strongly influenced by the way the game is described to the subjects. It makes a substantial difference, for example, if one's counterpart is referred to as 'partner' instead of 'opponent' [52]. People demonstrated a 31.5% rate of cooperation in a seven-round PD described as 'Wall Street Game', and a 66.1% rate when it was called a 'Community Game' [53].

These findings would be clearly troubling for a view that holds that social preferences are of a categorical or unconditional nature, because such preferences should not be dependent on the type of market one finds oneself in or on how a game happens to be described. However, it is not difficult to accommodate these findings with social conditional expectations. Levitt and List argue that socially disposed people self-select away from an environment like the trading floor, and this explains the absence of social preferences there. This is not implausible; not everybody is the same and behavioral types who can be described as profit maximizers and less concerned about other people may concentrate on trading floors, whereas a number of more social types may keep their distance from such areas. But it is even more clear, I suggest, that trading floors happen to be environments in which social preferences are little called for. The wheeling and dealing on Wall Street mostly functions without them and there are no expectations that traders would allow themselves to be influenced by such inclinations. Thus looping social preferences would predict their absence.

Naming the PD the 'Wall Street Game' likewise lowers expectations. I may well be cooperatively inclined in a neutrally described PD but the 'Wall Street' label alters this because I now expect others to defect to a higher degree. Further, I also expect that the others will probably think that I will be less socially inclined and therefore defect. And so on. Had there first been a loop, it would now unravel. The opposite goes for the 'Community Game'. Under this description, I now have less fear of being suckered because I now have stronger expectations about the other players' social preferences, and the same holds for them. This gets the loop going. 'Wall Street' and 'Community' are not just arbitrary names. They actually add information to the game. These names provide meaningful input to people's mutually referring expectations and preferences.

Whether the loop obtains among a group of people does not merely hinge on the people, on how they are; it hinges on their *interdependency*. Hence some people may cooperate just fine while a very similar group fails to do so. Cross-cultural studies are interesting in this regard. Ockenfels and

---

Sigmund's "image scoring" [50]. An exemplary experiment in behavioral economics with a public goods game associated with another game providing information on indirect reciprocity is [51].

Weismann [54] show that former West Germans contribute substantially more than former East Germans in a single play public goods experiment (which has a PD structure). Those from West Germany also have correspondingly higher expectations than the East Germans about the contributions in their own group. But the East Germans are not different people, the authors assert; East Germans are not by their natures less fair-minded or less cooperative types. Social dispositions are presumably not differently spread in these populations. "Behavior might instead depend on norms which differ between the two parts of Germany." Arguably, it is the loop norm that produces the higher level of contribution among the West Germans. Their (measured) higher expectations about their fellow players rationally presuppose that they act on a social preference. The West Germans tend to be cooperative in this experiment, it seems, because they perceive each other as fellow players, members of a group with a productive history. During this history the cooperatively disposed players have formed a generalized expectation that people with the same cultural background will generally not take advantage of them. Hence a large proportion of West Germans, and not East Germans, plays cooperatively. As the authors conjecture, the East Germans have unlearnt this during communist times. If they had a cooperative loop before, it has unraveled; their equilibrium is now mutually defecting.

Likewise, Castro [55] demonstrates that British subjects play more cooperatively than Italians. This effect completely disappears when the subjects are (knowingly) mixed. Cooperativeness among the British is apparently conditioned by what can be expected from the others.[21] We can infer that the British have high mutual expectations and cooperative inclinations among themselves. Kocher, Martinsson, and Visser [56] report a case of variation in contributions between groups of high school students that live near each other in Cape Town, South Africa, but in markedly different neighborhoods. Levels of contributions in a public goods experiment appear to be inversely related to average income in the neighborhood, with the students from the poorest area contributing most and those from the richer area contributing less. The researchers also measured expectations and found that levels of contribution strongly concorded with the expectations the players have about each other. Secondly, cooperation appeared to vary also with a six-point scale measure of trust in their schoolmates. As trust in other people is in large part a set of expectations about their motives, here too it seems likely enough that cross-referring mutual expectations and social preferences support the cooperation.[22]

Social preferences are contingent but not whimsical or unpredictably fragile. In a recent study Alexander Peysakhovich, Martin Nowak and David Rand [58] had a large group of subjects play a number of different mixed-motive games online, including prisoner's dilemma, dictator game, ultimatum game, a trust game, and two games with second-party and third-party punishment, respectively. They found that cooperative play strongly correlates across games. It also strongly correlates with helping behavior outside of these games (probed by a question to help the experimenters with providing feedback on the instructions, something the subjects could easily avoid doing). Cooperative play was also stable over time and it proved to be independent of punishing behavior. The researchers concluded that "the cooperative phenotype" shows a "substantial degree of domain generality and temporal stability."

As we saw, Levitt and List and also Henrich et al. make the case that social preferences are not domain general but context sensitive. For Henrich et al. the Orma provide the central evidence for this claim. According to the loop model, these ideas are not at odds; in fact they are complementary. People who are *generally* socially disposed act on social preferences which are *contingent* on having justified expectations about the other's social preferences being acted upon.

---

[21] Henrich et al. [1] itself is, of course, a landmark study that reports variation in contributions between cultures around the world, as is Gächter, Herrmann, and Thöni [57], but both these studies have not directly researched people's attitudes towards each other.

[22] Habyarimana et al. [59] ingeniously demonstrate that, with their experimental subjects, ethnicity mostly works as a coordinating device instead of being a marker of a sort of superiority or inferiority.

7. Conclusions

In this paper I have reviewed parts of the social preferences literature. The notion of equilibrium play between subjects who have conditional social preferences and expectations seems to give a promising explanation of the large variance in cooperative behavior in different contexts.

A theory that portrays social preferences as categorical has a loose end, rationally speaking, since these preferences should be based on expectations about other people. Without relating these expectations back to others' social preferences there would be another loose end. With looping social preferences and expectations these two ends are tied.

Social preferences are a robust phenomenon contingent on corresponding expectations. Thus, they can thrive in an environment where an institution with repeated cooperative behavior is in place. In this way, institutions in a society can scaffold pro-social attitudes among its inhabitants.

Acknowledgments: Many thanks to Jan Willem van der Rijt and the Philosophy and Economics seminar of Bayreuth University, and to Mark Alfano and Govert den Hartogh for valuable comments.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Henrich, J.; Boyd, R.; Bowles, S.; Camerer, C.; Fehr, E.; Gintis, H. (Eds.) *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*; Oxford University Press: Oxford, UK, 2004.
2. Ensminger, J.; Henrich, J. (Eds.) *Experimenting with Social Norms: Fairness and Punishment in Cross-Cultural Perspective*; Russell Sage: New York, NY, USA, 2014.
3. Ensminger, J. *Market Integration and Fairness: Evidence from Ultimatum, Dictator, and Public Goods Experiments in East Africa, In Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*; Oxford University Press: Oxford, UK, 2004; pp. 356–381.
4. Roth, A.; Prasnikar, V.; Okuno-Fujiwara, M.; Zamir, S. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *Am. Econ. Rev.* **1991**, *81*, 1068–1095.
5. Bicchieri, C. *The Grammar of Society: The Nature and Dynamics of Social Norms*; Cambridge University Press: Cambridge, UK, 2006.
6. Guala, F. Paradigmatic experiments: The ultimatum game from testing to measurement device. *Philos. Sci.* **2008**, *75*, 658–669.
7. Guala, F. Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behav. Brain Sci.* **2012**, *35*, 1–59.
8. Woodward, J. Social preferences in experimental economics. *Philos. Sci.* **2008**, *75*, 646–657.
9. Hausman, D. Fairness and social norms. *Philos. Sci.* **2008**, *75*, 850–860.
10. De Boer, J. A Strawson-Lewis defence of social preferences. *Econ. Philos.* **2012**, *28*, 291–310.
11. Paternotte, C.; Grose, J. Social norms and game theory: Harmony or discord? *Br. J. Philos. Sci.* **2013**, *64*, 551–587.
12. Smith, V. Theory and experiment: What are the questions? *J. Econ. Behav. Organ.* **2010**, *73*, 3–15.
13. Levitt, S.; List, J. What do laboratory experiments measuring social preferences reveal about the real world? *J. Econ. Perspect.* **2007**, *21*, 153–174.
14. Samuelson, L. Economic theory and experimental economics. *J. Econ. Lit.* **2005**, *43*, 65–107.
15. Binmore, K.; Shaked, A. Experimental economics: Where next? *J. Econ. Behav. Organ.* **2010**, *73*, 87–100.
16. Camerer, C.F.; Fehr, E. 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*; Oxford University Press: Oxford, UK, 2004; pp. 55–95.
17. Rigdon, M.; MCabe, K.; Smith, V. Sustaining cooperation in trust games. *Econ. J.* **2007**, *117*, 991–1007.
18. Rabin, M. Incorporating fairness into game theory and economics. *Am. Econ. Rev.* **1993**, *83*, 1281–1302.
19. Bolton, G.; Ockenfels, A. A theory of equity, reciprocity and competition. *Am. Econ. Rev.* **2000**, *90*, 166–193.
20. Andreoni, J.; Miller, J. Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica* **2002**, *70*, 737–753.
21. Charness, G.; Rabin, M. Understanding social preferences with simple tests. *Q. J. Econ.* **2002**, *117*, 817–869.
22. Camerer, C.; Fehr, E. When does 'Economic Man' dominate social behaviour? *Science* **2006**, *311*, 47–52.

23. Falk, A.; Fehr, E.; Fischbacher, U. Testing theories of fairness—Intentions matter. *Games Econ. Behav.* **2008**, *62*, 287–303.
24. Henrich, J.; Smith, N. *Comparative Experimental Evidence from Machiguenga, Mapuche, Huinca, and American Populations*; Oxford University Press: Oxford, UK, 2004; pp. 125–167.
25. Cox, J.; Sadiraj, K.; Sadiraj, V. Implications of trust, fear, and reciprocity for modeling economic behavior. *Exp. Econ.* **2008**, *11*, 1–24.
26. Charness, G.; Dufwenberg, M. Promises and partnership. *Econometrica* **2006**, *74*, 1579–1601.
27. Geanokoplos, J.; Pearce, D.; Stachetti, E. Psychological games and sequential rationality. *Games Econ. Behav.* **1989**, *1*, 60–79.
28. Colman, A. Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behav. Brain Sci.* **2003**, *26*, 139–198.
29. Den Hartogh, G. A conventionalist theory of obligation. *Law Philos.* **1998**, *17*, 351–376.
30. Den Hartogh, G. *Mutual Expectations. A Conventionalist Theory of Law*; Kluwer: The Hague, The Netherlands, 2002.
31. Sugden, R. *The Economics of Rights, Co-operation and Welfare;* Palgrave MacMillan: Hampshire, UK, 1986/2005.
32. Sugden, R. Hume's non-instrumental and non-propositional decision theory. *Econ. Philos.* **2006**, *22*, 365–391.
33. Davis, J. *The Theory of the Individual in Economics*; Routledge: London, UK, 2003.
34. Davis, J. *Individuals and Identity in Economics*; Cambridge University Press: Cambridge, UK, 2011.
35. Hamilton, W. Innate social aptitudes of man: An approach from evlutionary genetics. In *ASA Studies 4: Biosocial Anthropology*; Fox, R., Ed.; Malaby Press: London, UK, 1975; pp. 133–153.
36. Nowak, M.; Tarnita, C.; Antal, T. (2010). Evolutionary dynamics in stuctured populations. *Philos. Trans. R. Soc. B* **2010**, *365*, 19–30.
37. Skyrms, B. *The Evolution of the Social Contract*; Cambridge University Press: Cambridge, UK, 1996.
38. Skyrms, B. *The Stag Hunt and the Evolution of Social Structure*; Cambridge University Press: Cambridge, UK, 2004.
39. Skyrms, B. *Signals. Evolution, Learning and Information*; Oxford University Press: Oxford, UK, 2010.
40. Zollman, K. Talking to neighbors: The evolution of regional meaning. *Philos. Sci.* **2005**, *72*, 69–85.
41. Frank, R. *What Price the Moral High Ground? Ethical Dilemmas in Competitive Environments*; Princeton University Press: Princeton, NJ, USA, 2004.
42. Ekman, P. *Emotions Revealed*; Henri Holt: New York, NY, USA, 2003.
43. Gauthier, D. *Morals by Agreement*; Oxford University Press: Oxford, UK, 1986.
44. De Boer, J. A stag hunt with signaling and mutual beliefs. *Biol. Philos.* **2013**, *28*, 559–576.
45. Alfano, M.; Skorburg, G. The embedded and extended character hypotheses. In *Philosophy of the Social Mind*; Kiverstein, J., Ed.; Routledge: London, UK, 2016.
46. Guala, F.; Mittone, L. Experiments in economics: Extended validity and the robustness of phenomena. *J. Econ. Methodol.* **2005**, *12*, 495–515.
47. Smith, E.A. Making it real: Interpreting economic experiments. *Behav. Brain Sci.* **2005**, *28*, 832–833.
48. Waithima, A. The Role of Harambee Contributions in Competition. Experimental Evidence from Kenya; ICBE-RF Research Report No. 16/12; 2012. Available online: https://www.researchgate.net/profile/Abraham_Waithima/publication/268002279_The_Role_of_Harambee_Contributions_in_Corruption_Experimental_Evidence_from_Kenya/links/54da1c3f0cf25013d0440362.pdf (access on 9 October 2017)
49. Alexander, R.D. *The Biology of Moral Systems*; De Gruyter: New York, NY, USA, 1987.
50. Nowak, M.; Sigmund, K. Evolution of indirect reciprocity by image scoring. *Nature* **1998**, *393*, 573–577.
51. Milinski, M.; Semmann, D.; Krambeck, H. Reputation helps solve the 'tragedy of the commons'. *Nature* **2011**, *415*, 424–426.
52. Burnham, T.; McCabe, K.; Smith, V. Friend-or-foe intentionality priming in an extensive form trust game. *J. Econ. Behav. Organ.* **2000**, *43*, 57–73.
53. Liberman, V.; Samuels, S.; Ross, L. The name of the game: Predictive power of reputations versus situational labels in determining prisoner's dilemma game moves. *Personal. Soc. Psychol. Bull.* **2004**, *30*, 1175–1185.
54. Ockenfels, A.; Weimann, J. Types and patterns: An experimental East-West German comparison of cooperation and solidarity. *J. Public Econ.* **1999**, *71*, 275–287.

55. Castro, M. Where are you from? Cultural differences in public good experiments. *J. Socio-Econ.* **2008**, *37*, 2319–2329.
56. Kocher, M.; Martinsson, P.; Visser, M. Social background, cooperative behavior, and norm enforcement. *J. Econ. Behav. Organ.* **2012**, *81*, 341–354.
57. Gächter, S., Herrmann, B.; Thöni, C. Culture and cooperation. *Philos. Trans. R. Soc. B* **2010**, *365*, 2651–2661.
58. Peysakhovich, A.; Nowak, M.; Rand, D. Humans display a 'cooperative phenotype' that is domain general and temporally stable. *Nat. Commun.* **2014**, *5*, 4939, doi:10.1038/ncomms5939.
59. Habyarimana, J.; Humphreys, M.; Posner, D.; Weinstein, J. Coethnicity and trust. In *Whom Can We Trust? How Groups, Networks and Institutions Make Trust Possible*; Cook, K., Levi, M., Hardin, R., Eds.; Russell Sage: New York, NY, USA, 2009; pp. 42–64.