Public Reason 5 (1): 19-30

Social Norms: Repeated Interactions, Punishment, and Context Dependence

Jonathan Grose & Cedric Paternotte[1]
University of Bristol, Ludwig-Maximilians-University Munich

Abstract. We argue that recent game theoretic approaches to social norms differ on some fundamental issues, our focus being on recent accounts by Ken Binmore and Cristina Bicchieri. After a brief introduction, we present the deepest cause for their disagreement, namely whether the action of norms should be modelled as a one-shot game, the option favoured by Bicchieri, or by a repeated game, as Binmore does. Although these choices appear to leave room for the two accounts to be complementary, we then argue that this is not possible. First, differing attitudes to modelling punishment, a central feature of all informal work on social norms, prevent any straightforward integration of the two theories. Second, the solution cannot consist in merely choosing between the two accounts, as they both fail to deal with the way in which triggered norms depend on context, in static as well as diachronic frameworks.

Key words: social norms, Game Theory, context-dependence, punishment, repeated interaction.

I. GAME THEORY AND SOCIAL NORMS

Game-theoretic approaches to social norms have flourished in the recent years, and on first inspection theorists seem to agree on the broad lines that such accounts should follow.[2] By contrast, this paper aims to show that the main two interpretations of social norms are at odds on at least one aspect of social norms, and both fail to account for another aspect.

We are sympathetic to the broad project of using game theory to model social norms. Our aim is not to undermine this project or argue that no coherent framework or frameworks is possible in principle. Rather, we aim to show, with reference to two particular issues, that at this stage it is unclear if and how differing approaches can be integrated into a unified picture. The two issues we go on to discuss are, first, the role of punishment within the models and, second, how they deal with the context dependence of norm triggering. In each case it is the question of whether or not the model should be one of repeated or one-shot interactions that is the source of problems for integration.

When it comes to different modelling approaches, the recent literature provides two presentations of social norms. Ken Binmore outlines a model that frames social norms in terms of repeated interactions (1994, 1998, 2005). In contrast, Cristina Bicchieri's model is of one-shot interactions and involves utility transformations, triggered by social context,

1] This paper stems from our earlier "Social Norms and Game Theory: Harmony or Discord?," (Paternotte & Grose, 2012) and shares its overall diagnosis but adds to it by providing a deeper analysis of several points.

2] Despite their disagreements on specific issues; see for instance the issue 9 (vol. 3) of *Politics, Philosophy and Economics*, 2006.

that make norm conformity the rational behaviour (2006, 2008). Both of these authors repeatedly refer to each other's work as belonging to a common programme but, we argue, their difference on the question of repeated interaction models makes integration of their approaches problematic.

We draw attention to the fact that differences on the question of whether we should think in terms of repeated encounters or not lie at the very foundations of the models. Such disagreement makes common approaches to punishment and context dependence deeply problematic. This is not to say that having multiple approaches is unwelcome in principle. Still, those unfamiliar with the literature could be forgiven for taking it that there are core similarities common to all game theoretic treatments of social norms. In fact, the only issue on which they unambiguously agree is that norms should be modelled as being equilibria of a game theoretic model. This leaves open the question of what is the correct game to be modelled and it is to that issue that we move in section two.

II. TO REPEAT OR NOT TO REPEAT?

First, note that it would be totally unreasonable to demand that all social interactions be modelled by the same game. Our point in this section is that a fundamental difference in the models presented by Binmore, Bicchieri and others makes it difficult to see them as compatible analyses that can sit on the shelf next to each other and be called upon depending on the strategic nature of the situation at hand. Rather, they disagree at a deep level about the way in which we should model norms and important features such as punishment.

The core difference between these approaches to social norms that we consider is the type of interaction modelled. It has been common to identify the fundamental problem of social cooperation with the prisoner's dilemma (PD). For instance, it has been suggested Hobbes' *Leviathan* presents an informal analysis of PD interactions (1991). More recently, Brian Skyrms (2004) stresses the value of thinking about social interactions in terms of the "Stag hunt" game, citing Rousseau as a precursor. Previously, Skyrms (1996) devotes chapters to the PD and to Hawk-Dove / Chicken games.

These studies reflect one aspect of norm-following that has been stressed in the literature, that being that norms pull us towards actions not in accordance with our narrow self-interest (Young 2008). In most of the above games the equilibrium outcome is at odds with that which is optimal from society's point of view. The exception being the stag hunt in which there is a coordination problem between two equilibria, one of which is Pareto superior to the other.

Bicchieri's account of social norms is broadly in keeping with these approaches. She stresses that norms operate in "mixed motive" interactions. These are loosely characterised as those in which there is a conflict of interest and the potential for mutual gain. Given this loose characterisation, her definition appears to apply to a very wide class of games. Demanding that a game contains conflict to some degree can be taken as meaning that

different players' payoffs are not perfectly correlated. Requiring the presence of a mutual benefit can be translated as requiring that there be a Pareto-dominant payoff profile or a Pareto undominated one, or even one which provides a better average payoff. Although her characterisation is imprecise we can see that the focus is on norms overcoming conflicts of interest. This becomes apparent when we examine the way in which the action of norms is modelled.

Bicchieri is very explicit that, "the problem that a social norm is solving in the first place is *never* a coordination problem." (Bicchieri 2006, 34; her emphasis) When a social norm exists, it *changes* the agents' utility functions. Preference for conformity comes about via a transformation of player *i*'s utility function which converts a mixed-motive game to a coordination game in which all (or, more accurately, enough) players prefer to conform. Once the players' utilities have been transformed, mutual norm compliance becomes an equilibrium of the transformed game. When the norm is triggered, agents' utilities are modified for every result of the game as a function of the highest loss caused by anyone's deviations from the norm. My being sensitive to the norm makes me consider as less valuable the consequences of actions that do not follow it (whether these actions are mine or others'), in proportion to the highest loss that this combination of actions entails (Bicchieri 2008, 199, Appendix 9.1).

When it comes to reasons for this preference transformation when a norm is triggered, Bicchieri makes a number of suggestions. Agents may be motivated by the desire to please others, recognition that others normative expectations are reasonable and, most importantly for our purposes, fear of punishment (Bicchieri 2006, 29).

At this point we emphasise the feature of this account on which we subsequently focus. The model is of so-called one-shot games with no modelling of repeats of the interaction. Norm-following becomes rational, equilibrium behaviour in the one-shot game once the players' utilities have been transformed. In this respect Bicchieri's model is similar to Herbert Gintis's recent model in which a "normative disposition" discounts non norm-following acts (2010).

In contrast, Binmore's account is explicitly one of repeated interactions. For him, one-shot interactions are not those for which we are prepared by either our biological or cultural heritage. The "game of life" is an indefinitely repeated game. (Binmore 1994, 25) It is this focus on repeated interactions that makes Binmore concerned with a different role for norms than that of transformation of preferences. He focuses on the coordinating role of social norms that helps us to make our actions fit appropriately with each other. In game theoretic terms this fit is framed in terms of behaviours being in equilibrium with each other, where neither player would want to unilaterally change their action. The "folk theorem" of repeated game theory demonstrates that indefinitely repeated games have multiple equilibria, that is, multiple different ways in which actions can be mutually appropriate, and thus players face an equilibrium selection problem. (Myerson 1991, §7.5) For Binmore, social norms solve this problem by making a particular behaviour salient in a particular context. Our shared cultural heritage is what allows us to coordinate on an

equilibrium. Notice that in the case of Bicchieri's (and Gintis's) norms, the coordination role of norms need not arise because the utility transformation makes norm-conformity the one rational outcome.

It should also be emphasised that Binmore's focus on repeated games makes his account totally general when it comes to the base games that can be repeated. PDs, stag hunts, hawk-dove, coordination games and so on can all fall under this approach and in each case the equilibrium selection is between equilibria of the repeated game rather than what would be the equilibria if the games were played just once.

At this point it appears that there need be no incompatibility between these approaches to social norms and scope for them to be integrated. Preference transformations attempt to capture the notion of norm-following pulling people to act against their narrow self interest and their coordination role is reflected in the difficult equilibrium selection problem faced in repeated games. Integration would take the form of the Bicchieri's games forming the base games for Binmore's repeated interactions. However, we first argue, at least on the important issue of punishment, the different accounts cannot be straightforwardly integrated. We then show that they both fail to adequately describe how context-dependent social norms are, for related but different reasons.

III. WHAT PLACE FOR SANCTIONS?

There are at least three ways in which to model game theoretically the action of punishment or social sanctions. As we will see, different approaches to norms model them in different ways. In itself this is not a problem. As we stressed in section one, there is potential value in having multiple modelling approaches. The problem here is that the different ways of thinking about punishment makes them appear incompatible with each other in the sense that they cannot be integrated. This is of particular concern because punishment of non-compliance is one of the central features of any informal account of social norms.

Returning to the three game theoretic models of punishment, one possibility is for the sanction to feature as a payoff alteration. For instance, if defection in a PD is punished, and payoffs following defection are reduced, mutual cooperation can become an equilibrium outcome. This payoff transformation option is certainly the one taken by Gintis. He emphasises the "choreographing" role of norms in (one-shot) games with multiple equilibria. However, where compliance is not already part of an equilibrium outcome, sanctions can play a formal role as an argument in the utility function, thereby tipping the balance in favour of norm-following being an equilibrium (2010). In the case of Bicchieri, punishment does not feature explicitly in her formalization. However, as raised in section two, sanctioning is cited as a reason for the utility transformation that is a central feature of her model. In fact, it is unclear precisely how punishment connects with her formal model and in this sense the role of sanctions is poorly integrated with that formal account. An obvious possibility is the payoff altering role we are considering

here. In that case the discounting of payoffs associated with non-conformity with the norm would reflect punishment received. However, if we look closely at Bicchieri's formalization it seems that this cannot be a correct interpretation, something emphasised by Daniel Hausman (2008). Remember that the discounting of norm-breaking actions is a function of the negative consequences of *all* players' deviations from norm-following. If the discounting of my payoffs were due to potential punishment *of me* then we would expect such discounting to be a function only of my own deviations from the norm. So, while it is clear that, according to Bicchieri, sanctions play a role in transforming utilities, it is unclear precisely in what way this is cashed out in the model.

A second formalization of sanctioning is to expand the game to make the punishment action an explicit move. This has the advantage of prompting the modeller to pay attention to the possible consequences of punishing for the punisher. In particular, punishing is very often taken to involve paying costs oneself.[3] We will not expand further on this option since it is not one taken by either Bicchieri, Gintis or Binmore, to whose repeated interaction model we now move.

In models of repeated interactions punishment behaviour can be modelled in a way unavailable to one-shot models. In this case some plays of the base game constitute sanctioning behaviour. Take, for instance, the indefinitely repeated PD. Permanent mutual cooperation can be the outcome of equilibrium strategies. One, but not the only, possible equilibrium strategy is the famous "Tit-for-Tat."[4] In this case cooperative behaviour is reciprocated but, importantly for its being an equilibrium, so is defection. Retaliatory defection is interpreted by Binmore as a punishment for breaking the cooperative norm. A less forgiving strategy is the so-called "GRIM strategy" (Binmore 1994, 197). This cooperates until it is defected against and then switches to the punishment of permanent defection.

A strength of interpreting punishment in terms of strategies in repeated games is that it naturally makes explicit that punishment can take many forms in terms of its duration and by what it is provoked. For instance, "Tit for two Tats" requires two successive defections before it punishes with defection. However, a weakness is that punishment loses some of its special status found in informal accounts of social norms. What we mean by this is that there is nothing distinctive about punishing as represented by a repeated game strategy compared to any other behaviour that is conditional on ones partner's actions in previous rounds of play.

Having raised some tensions in both Bicchieri and Binmore's framing of punishment we move on to the question of integrating their models since, even if the issues raised above are not fatal to their accounts, there remains a problem with bolting them together. Remember that the integration suggested at the end of section two was that the utility

3] Something Gintis and others stress in their claims for a disposition in humans towards "strong reciprocity" (Gintis 2000).

4] See Binmore 1994, §3.2.5 for powerful arguments that TFT's importance as a repeated PD strategy has been overemphasised.

transformation role of norms acts on the base game and that we can then examine the coordination role for norms when repetitions of these games create, via the folk theorem result, an equilibrium selection problem. However, we have already seen that punishment plays a role in both accounts but is characterised in different ways. For Bicchieri (and Gintis), punishment triggers norm-following in a one-shot context, potentially via it being represented as altering players' payoffs. In contrast, in the repeated game framework, punishment is a constitutive part of the strategies themselves.

A potential response to this difference is that the repeated game approach can be reinterpreted into the one-shot case by taking the expected payoffs of the repeated payoff stream as the payoffs of a new one-shot game. This approach is followed by Skyrms when he argues that the repeated PD can be rewritten as a one-shot stag hunt game (Skyrms 2004). However, this is a reframing of the repeated scenario rather than an integration of the two approaches. To reiterate, our concern is that in one case punishment features as a reason for action and in the other it is just one of the moves in the game. When attempting to apply an integrated model to actual cases of norm-following where social sanctions are present it would then be ambiguous in which of the two ways we should model those sanctions.

At this point we are left with the question of whether one interpretation is to be preferred in general or in specific cases of norm-following or whether we should accept a plurality of modelling approaches. It is not our intention to adjudicate on this point here. What we stress is that such an adjudication (including agnosticism) would not be a move towards an integrated game theoretic account of social norms. Moreover, the following sections suggest that before any possible integration, both interpretations have to be completed in order to integrate the fact that norms are highly context-dependent.

IV. CONTEXT DEPENDENCE: THE TRIGGERING AND STRATEGIC ROLES OF EXPECTATIONS

Despite the striking variance in their analysis, as shown in the case of the nature and role of punishment, all accounts of social norms suffer from at least one common shortcoming, for related reasons. The basic problem is the following. A certain situation of interaction, formalised by a game (the sets of agents' actions and payoffs), may well trigger various social norms, depending on the background context of the interaction. When an interaction is repeated, should we say that the context changes so that a new social norm may appear, or that it does not since the base game describing each interaction is unchanged? In other words, when during a repeated interaction should we expect a social norm to be stable and when unstable?

First, let us recall the characteristics of one-shot and repeated interaction approaches to social norms, respectively exemplified by Bicchieri's (2006) and Binmore's (1998, 2008). For Bicchieri, a social norm is cued in certain situations by the agents' beliefs about others' behaviour and expectations. As seen in section two, this triggers a change

in preference that makes it rational for agents to choose certain actions. For Binmore, whether a norm is triggered depends on the similarity of the situation with a known one. Agents' actions are not explained by a preference change, but by their behaving as they are used to in a similar situation. In other words, when facing new situations in the laboratory, agents either conserve preferences or habits from their outside life (Woodward 2008).

Do the starting game's characteristics constrain the existence of a social norm? Not really. Theorists usually reckon that cooperative norms can appear in any kind of cooperative dilemma, that is, of games containing an outcome that, although not an equilibrium, Pareto-dominates an equilibrium.[5] However, this expresses the theorists' interest for certain kind of games (those in which cooperative behaviour calls for an explanation) rather than a logical necessity or an empirical fact. Indeed, social norm theorists are disposed to see norms appear in almost any kind of game. Binmore does not discuss the issue, and we have seen in section two that Bicchieri holds that norms can appear in any 'mixed-motive games,' that is, games containing potential for mutual benefit and some conflict of interest; this description suits most games. Defenders of the competing 'group identity' explanation of cooperation are not any more precise. For instance, Bacharach (2006) only talks of situation of 'strong interdependence,' which is nothing else than the presence of a Pareto-dominated Nash equilibrium - although he provides no explanation why this condition is paramount. Overall, virtually any game could cue a social norm.

This makes context all the more important to social norms. The term "context" refers to anything that cannot be expressed by games' parameters. So the absence of constraints or payoff structures on the existence of social norms increases the importance of expectations or of the similarity with real-life situations in determining when social norms may appear.[6] Of course, games still play a role in determining *what* social norms may appear. On this point though, somewhat unexpectedly, one-shot and repeated interaction approaches are not as separated as it seems. Expectations matter in the former, but even if they are not explicitly mentioned in the latter, their role is merely hidden.

Suppose an agent is confronted with laboratory game A and deems it similar to real-life situation B, in which she would choose to do X. She will then do X in A for the same reason that she would have done so in B: because it is part of an equilibrium (she thinks mistakenly). But by definition, agents play according to an equilibrium if they believe that others will do the same. In a repeated game approach, expectations about others' behaviour and their expectations determine what an agent is going to do; the only difference is that they do not also lead to a preference change. Note that expectations also play this role in one-shot approaches, in addition to their triggering effect: once a game is transformed,

5] E.g. the Prisoner's Dilemma, in which the mutual cooperation outcome makes both players better off than the mutual defection one does, although only the latter is a Nash equilibrium.

6] An example of the effect of similarity between games on behaviour is found in Binmore 2010: when individuals had to play a public goods game, "recognised" it as a familiar situation called "harambee" and then played accordingly. An example of the role of expectations can be found in Bicchieri 2008.

expectations about others still influence a player's choice in the new game. In other words, expectations can have both a *triggering* role and a *strategic* role (this distinction will become important in the next section). One-shot and repeated interaction approaches to social norms both recognize the strategic role of expectations, but only the former also explicitly mentions their triggering role.

A related reason why the two approaches are closer than it may seem is that the factors that cue norm-following behaviour are not necessarily different. In the one-shot interaction approach, certain expectations trigger preference change, but there is no constraint on the expectations that agents may or may not form. This reflects a widespread practice in rational choice theory: theorists determine what is the best, or rational choice for an agent given her preferences and beliefs, without asking whether those can be considered as rational or acceptable themselves. Now an agent's expectations about others may perfectly stem from the similarity of the laboratory game with a real-life situation: because A is similar to B, I may expect others to play according to B's equilibrium. So the role of game similarity could find a space even in one-shot approaches, as one possible origin of agents' expectations.

In both approaches, similarity between games may cue a social norm and resulting expectations influence the agents' behaviour. What are the differences then? First, according to the one-shot interaction approach; other factors than similarity between games may trigger a social norm. However, the main difference lies in the link between the base game and the game that agents are actually playing. According to a repeated interaction approach, the laboratory game A is *replaced* by one representing the similar real-life situation B, in which players then play according to one of B's equilibria. According to a one-shot approach, the payoffs of laboratory game A are *transformed* into those of a game C. The difference is that there need not be a payoff transformation function that leads from A to B, and more precisely not one that corresponds to Bicchieri's description (2006, 52-54). In both cases, the characteristics of game A partly influence the game that players are *really* playing; only the way to determine the latter from the former varies.

V. CONTEXT-DEPENDENCE IN REPEATED GAMES

When agents interact only once, context-dependence is not deeply problematic, as it all depends on which expectations agents have or which real-life situations they deem similar to the one at hand. Surely, this makes social norm following behaviour hard to predict, as a theorist would need to know all possibly related real-life situations and all expectations linked to a social norm in a given population. Still, the analyses provided by both one-shot and repeated interaction accounts are clear.

What happens to a social norm when the same game gets repeated? What makes agents keep sticking to it or start following another one along the way? On this point, the two accounts described above start to differ significantly, even if none of them provide a satisfactory answer.

On the one hand, as the base game is just repeated and does not change, agents may well keep following the same norm. However, there are always multiple equilibria in a repeated game, and according to Binmore each of them could be a norm. There are also multiple ways in which the game's payoffs can be modified through preference change functions, and so just as many norms according to Bicchieri. As a game is repeated, agents observe each other's behaviour and consequently may see a change in their expectations; they may also start to understand the nature of the game they are playing and as a result adapt their behaviour to it. For these reasons, both one-shot and repeated interaction accounts of social norms may predict that a social norm that agents follow changes as time passes. What does this change depend on? One way to put the problem is: when should observations of behaviour during repeated play lead to a change in the social norm an agent follows?

It is a fact that most often, agents' behaviour varies with time when they play a repeated game. As may be expected, the repeated interaction account of social norms fits some of such cases well. For instance, the rate of cooperation in a repeated public good game tends to decrease with time (Camerer 2003, 59). In an Ultimatum game, offers can usually stabilize around one arbitrary value, although in the Dictator's game (when they cannot be refused) they get closer to zero. In these three cases, agents' strategies converge towards one of the game's Nash equilibria (no contribution in the public good game, any nonzero offer in the Ultimatum one, zero offer in the Dictator game). This is consistent with what the repeated interaction account predicts: when playing game A, agents may start to act as if they were playing game B, but as repetitions of the game accumulate, their understanding of the situation will improve and they will gradually learn to play according to A's equilibria. The effects of the context (that is, the existence of a similar real-life situation) can thus be offset by the success of an agent's strategy, in terms of its concrete payoff.

Is this interpretation consistent with all experiments? It seems so. Consider Isaac and Walker's (1988) experiment (discussed by Bicchieri, 2006, 149ff.), consisting of two separated runs of ten successive public good games. Conversation between participants was allowed either only before the first sequence, only before the second sequence, or not at all. What was observed is although the level of contributions typically declines as the game is repeated, allowing conversation before a sequence led to higher, sometimes *increasing* contributions, and that the effect of a conversation before the first sequence carried over to the second one. The increasing effect is puzzling from a repeated interaction account, because in any equilibrium of a repeated social dilemma, there should be a decrease of the contribution level at least in the last round. If agents were slowly learning not to mistake the situation for a different one, their behaviour should converge towards such an equilibrium. It may be that the learning process needs longer than a handful of repetitions to kick in.

However, such an argument threatens to render the explanation ad hoc: agents could be said to learn whenever they play according to an equilibrium (which are many

in repeated games), and not to learn yet whenever they do not.[7] Moreover, whenever agents do not play according to a finitely repeated game's equilibrium, such as when their contributions increase in the last period, one can always say that they are still behaving as if in real life, when it is hardly ever sure that an interaction will not be repeated some time in the future. Any behaviour in the last repetition could thus be seen as part of an equilibrium of the infinitely repeated game. Even if players perfectly understand that the numbers of repetitions in an experiment is finite, they might be behaving partly intuitively, based on the similarity between laboratory and real-life situations. Put differently, to be satisfying the explanation should tell us when agents act strategically (by considering the payoffs and structure of the actual interaction) and when habitually.

How does a one-shot interaction account fare? It actually faces a similar problem. Let us start with a difference: one-shot accounts based on preference transformations have no problem explaining that contribution levels in a social dilemma should not decrease in the last game. This is because the preferences of norm-following agents may be such that contributing zero is not an individually dominant strategy anymore (by contrast, without a preference change, payoffs are such that defecting is always individually beneficial). Still, it is just as difficult to say when agents should start following different social norms. The problem stems from the unknown balance between the triggering and strategic effect of expectations (as described in the previous section).

When agents can have different preferences (or type) and are uncertain about others' types, a useful game-theoretic concept is that of a perfect Bayesian equilibrium (Osborne and Rubinstein 1994, 231-37). When applied to repeated games for instance, it says that agents start with a prior belief about everyone's possible type, which they will then update by Bayesian conditionalisation as they observe others' action. Observations constantly modify agents' beliefs about others' types and expectations about their behaviour. These changes of beliefs are part of the Bayesian Nash equilibrium: two strategies can only be at equilibrium if when agents implement them, the change of beliefs they cause is consistent with the strategies' prescriptions. This models the *strategic* role of expectations, that is, the way in which they help determine the behaviour agents who maximize their expected utility.

Now recall that in Bicchieri's one-shot interaction account of social norms, expectations have a *triggering* role, that is, they can lead to a change in preferences. So a change in expectations may well cause an agent swapping types, and in particular can lead to the appearance of new types. This cannot be made part of Bayesian Nash equilibria, in which a list of possible types is set from the beginning and cannot evolve. The problem is that the theory can then explain virtually any observed behaviour: either agents follow a well-defined norm, in which case expectations play a triggering role at the start and then

7] Note that the carry over effect in itself need not be a problem for a repeated interaction account. There is a repeated game equilibrium in which agents make high contributions all the time, except towards the end. As long as the contribution level decreases at some point, it may stay high for a long time before, and there is no reason why this would not carry over between several sequences of games.

only a strategic role (preferences do not change as the game is repeated); or it does not and can be explained by the fact that the expectations changed over time and so triggered another norm.

Overall, both kinds of accounts seem able to fit any data, thanks to the liberal definitions of context. Expectations can be part of the context; as they routinely change during any repeated interaction, they may trigger a change in norms at any time and thus allow one to explain any behaviour. Learning processes determine when the context's influence stops overcoming benefit-related considerations; but in the absence of a precise definition of such processes, the effect of context can also be used to explain any behaviour.

VI. CONCLUSION

We have argued that despite surface-level similarities, game-theoretic accounts of social norms are not easily integrated. This is due to the existence of two main kinds of accounts, based on one-shot or on repeated interactions. This distinction gives rise to different treatments of the role played by punishment. Moreover, the problem is not merely to choose between them, as they both suffer similarly from difficulties to account for the context-dependence of social norms while conserving their explanatory power.

jonathan.grose@bristol.ac.uk
cedric.paternotte@lrz.uni-muenchen.de

REFERENCES

Bacharach, M. 2006. *Beyond Individual Choice: Teams and Frames in Game Theory*. Edited by N. Gold and R. Sugden. Princeton University Press.

Bicchieri, C. 2006. *The Grammar of Society - The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.

———. 2008. How Expectations Affect behavior: Fairness Preferences or Fairness Norms?. In *Rationality and Responsibility*, edited by J. Krueger. New York: Psychology Press.

Binmore, K. 1994. *Game Theory and the Social Contract. Volume 1: Playing Fair*. M.I.T. Press.

———. 1998. *Game Theory and the Social Contract. Volume 2: Just Playing*. M.I.T. Press.

———. 2005. *Natural Justice*. Oxford University Press.

———. 2006. Why do people cooperate?. *Politics, Philosophy and Economics* 5 (1): 81-96.

———. 2008. Do conventions need to be common knowledge?. *Topoi* 27: 17-27.

———. 2010. Social norms or social preferences?. *Mind & Society* 9 (2): 139-57.

Camerer, C. 2003. *Behavioral Game Theory*. Princeton University Press.

Gintis, H. 2000. Strong reciprocity and human sociality. *Journal of Theoretical Biology* 206: 169-79.

———. 2010. Social Norms as Choreographer. *Politics, Philosophy and Economics* 9 (3): 251-64.

Hausman, D. 2008. Fairness and Social Norms. *Philosophy of Science* 75: 850-60.

Hobbes, T. 1991. *Leviathan*. Edited by R. Tuck. Cambridge: Cambridge University Press.

Isaac, R., and J. Walker. 1988. Communication and Free-Riding Behaviour: The Voluntary Contribution Mechanism. *Economic Inquiry* 26: 585-608.

Myerson, R. 1991. *Game Theory, Analysis of Conflict*. Harvard University Press.

Osborne, M.J., and A. R. Rubinstein. 1991. *A Course in Game Theory.* M.I.T. Press.

Paternotte, C., and J. Grose. 2012. Social Norms and Game Theory: Harmony or discord. *British Journal for the Philosophy of Science*. doi: 10.1093/bjps/axs024.

Skyrms, B. 2004. *The Stag Hunt and the Evolution of Social Structure.* Cambridge: Cambridge University Press.

———. 1996. *Evolution of the Social Contract.* Cambridge: Cambridge University Press.

Woodward, J. 2008. Social Preferences in Experimental Economics. *Philosophy of Science* 75 (5): 646-57.

Young, H. P. 2008. Social Norms. In *The New Palgrave Dictionary of Economics,* edited by S. Durlauf, N. and L. Blume. Basingstoke: Palgrave Macmillan.