GOVERNANCE AND THE EFFICIENCY
OF ECONOMIC SYSTEMS
GESY

Discussion Paper No. 193

Why and How Identity Should Influence Utility

Philipp C. Wichardt*

January 2007

*Philipp C. Wichardt, Economic Theory 3, University of Bonn, Adenauerallee 24-26, D-53113 Bonn, Germany. philipp.wichardt@uni-bonn.de.

Financial support from the Deutsche Forschungsgemeinschaft through SFB/TR 15 is gratefully acknowledged.

Sonderforschungsbereich/Transregio 15 · www.gesy.uni-mannheim.de
Universität Mannheim · Freie Universität Berlin · Humboldt-Universität zu Berlin · Ludwig-Maximilians-Universität München
Rheinische Friedrich-Wilhelms-Universität Bonn · Zentrum für Europäische Wirtschaftsforschung Mannheim

Speaker: Prof. Konrad Stahl, Ph.D. · Department of Economics · University of Mannheim · D-68131 Mannheim,
Phone: +49(0621)1812786 · Fax: +49(0621)1812785

Why and How Identity Should Influence Utility

Philipp C. Wichardt[†]

Department of Economics, University of Bonn

This Version: January 11, 2007

Abstract This paper provides an argument for the advantage of a preference for identity-consistent behaviour from an evolutionary point of view. Within a stylised model of social interaction, we show that the development of cooperative social norms is greatly facilitated if the agents of the society possess a preference for identity consistent behaviour. As cooperative norms have a positive impact on aggregate outcomes, we conclude that such preferences are evolutionarily advantageous. Furthermore, we discuss how such a preference can be integrated in the modelling of utility in order to account for the distinctive cooperative trait in human behaviour and show how this squares with the evidence.

Key words: Cognitive Dissonance, Fairness, Identity, Reciprocity, Social Norms, Social Preferences, Utility

JEL code: A13, C70, C90, D01, Z13

[*] Acknowledgements: I am grateful to Matthew Braham, Urs Fischbacher, Anke Gerber, Georg Nöldeke, Patrick Schmitz and Wiebke Wichardt as well as to seminar audiences in Hamburg, Innsbruck and Jena for helpful comments and suggestions. A scholarship of the German Research Foundation (DFG), GRK 629 at the University of Bonn, is gratefully acknowledged. The usual disclaimer applies.

[†] Postal address for correspondence: Economic Theory 3, University of Bonn, Adenauerallee 24-26, D-53113 Bonn, Germany; e-mail: philipp.wichardt@uni-bonn.de.

1 Introduction

Putting game theory to work necessitates a sound understanding of the agents' preferences over outcomes, i.e. of the ingredients of utility. Yet, while abstract formal game theory has flourished since its introduction to economic theory by von Neumann and Morgenstern in 1944, the discussion of utility - until recently - has been less intense. As a consequence of this, many applied studies of game theory have relied simply on (an affine transformation of) the agents' material rewards as a proxy for the utilities associated with the respective outcomes; a choice which is natural (although not imperative) in view of the strong emphasis of non-cooperative game theory on the agents' rationality and self-interest. Over the last decades, though, a still mounting evidence for the frequent disconformity of thus predicted and observed behaviour has gathered - in particular for games which are not purely competitive (e.g. public goods games). The evidence indicates that one of the possible causes for this discrepancy can be found in the above mentioned focus on material incentives in the modelling of the agents' utility.

As a result of this, several amendments to the standard model of utility have been proposed. For example, it has been argued that agents care about fairness, i.e. have other-regarding preferences (e.g. Fehr and Schmidt, 1999); that agents care about and reciprocate the intentions of their opponents or counterparts (e.g. Falk and Fischbacher, 2006; Rabin, 1993); or that agents simply (and selfishly) prefer to act in accordance with their identity, i.e. the social norms and stereotypes they have internalised (e.g. Akerlof and Kranton, 2000; see also Benabou and Tirole, 2006b). All these approaches, which will be discussed in a later section, have contributed to the improvement of the empirical validity of game theory.

Against the backdrop of the increasing variety of new models of utility, however, the question arises *why*, from an evolutionary perspective, immaterial concerns should influence utility and, once this question has been answered, what this implies for its modelling. To argue why only material rewards should matter for utility is straightforward. Much simplified, the argument would be that the more the agent possesses the higher his evolutionary fitness so that material self-interest will prevail in the end. Yet, apparently pure material self-interest is not compatible with the data; so why is that?

Over the years, many attempts have been made to argue for the evolutionary advantage of a cooperative trait in human behaviour, especially in terms of reciprocity or, more recently, indirect reciprocity (e.g. Bester and Güth, 1998; Nowak and Sigmund, 1998, 2005; see Henrich, 2004, for a critical discussion). However, as Henrich (2004) points out, there are a couple of deficiencies these approaches commonly suffer from. In particular, the cost of the feature which mainly drives the selection process into the desired direction usually is not accounted for. For example, if agents know the preferences of their current opponent so as to adjust behaviour accordingly (e.g. Bester and Güth 1998) or if agents keep track of past behaviour of others in order to reciprocate only the “right” ones (e.g. Nowak and Sigmund, 1998), this usually comes for free - irrespective of the size of the population or the number in players of the game. Moreover, according to Henrich, most approaches explore only repeated two player interactions, but fail to give a plausible account of large-scale cooperation, as usually observed in human societies, or cooperation in one-shot or anonymous situations, as commonly observed in the lab (e.g. Camerer, 2003).

In the present paper, we address the question about the evolutionary advantage of immaterial concerns in the agent’s utility function from a slightly different perspective which allows us to largely avoid the frequent shortcomings indicated above. Within a simple model of social interaction, it is argued that the development, and dissemination of cooperative social norms which detract the agents’ focus from pure material self-interest (e.g. fairness norms) can be explained if (at least some of) the agents possess a preference for identity consistent behaviour that impacts on economic decision making (cf. Akerlof and Kranton, 2000). Thus, in contrast to previous attempts, we do not argue for the evolutionary advantage of cooperative behaviour or other-regarding preferences per se, but only for the advantage of a preference for identity consistent behaviour in the spirit of Akerlof and Kranton (2000) in conjunction with cooperative social norms.[1]

For the subsequent analysis, social interaction is modelled as a repeated random encounter between myopic, purely self-interested agents who play a one-shot Prisoner's Dilemma game which is complemented by a subsequent costly punishment opportunity. The idea of the punishment is to protect mainly cooperative societies against the invasion of defectors (through permanent punishment of defection).[2] However, as punishment is costly, its enforcement is not evolutionarily stable. To establish cooperation nonetheless, i.e. to enforce the costly punishment of defectors, we assume that agents can choose whether or not to identify with a social norm which prescribes the following kind of behaviour: 1. cooperation; 2. punishment of defectors; 3. (costly) monitoring of others such that, with a small probability, norm violators are recognised and (partly) separated from the norm-obedient agents. If the norm is followed, the direct personal punishment (2.) will deter defectors while the higher order social punishment (3.) will prevent the invasion of both defectors and "lazy" cooperators who do not contribute to the costly enforcement of the norm (which would again pave the way for an invasion of defectors).

Eventually, the partial separation of the agents, i.e. the induced assortativeness of the matching process (cf. Bergstrom, 2002, 2003), is what drives the selection process in favour of those who identify with and follow the norm. Yet, this separation does not come for free in our model. It relies on a costly ingredient of the norm (monitoring others), the performance of which is strictly dominated from the agents' perspective. Thus, in a sense, the norm creates another social dilemma: although socially desirable, it is not individually optimal to follow it. Consequently, without a preference for identity consistent behaviour, defectors, who do not follow the norm, prevail. However, once we introduce a fraction of agents with such a preference things change. For these agents, norm-obedience is the dominant action (if the agent

[1] See North (1990,1993) for a discussion of the social value of cooperative norms.

[2] For empirical evidence on the role of punishment see, e.g., Falk et al., 2005.

identifies with the norm). Hence, monitoring takes place and the assortative power of the matching process can take the desired effect, i.e. norm-obedient cooperators eventually dominate the society. As the preference for identity consistent behaviour is necessary for agents to obey the norm, we conclude from our argument that such a preference is evolutionarily advantageous.

Based on our previous argument, we also discuss how exactly a preference for identity consistency in conjunction with (in our example cooperative) social norms should be incorporated into a model of the agent's utility as relevant for decision making. In particular, it is argued that (and how) the influence of social norms on the agent's decision making should vary with both the agent's past experience and the general context in which the decision takes place. In essence, the more a context is evocative of cooperative social norms and the more the agent's experience supports this view, the more the agent will tend to adhere to the respective norms, as opposed to material incentives, in order to avoid self-inconsistencies (e.g. a guilty conscience). This effect is independent of the size of the population or the frequency of the interaction, and it even allows us to naturally account for the effects of anonymity (which, according to our argument, will reduce but not immediately dispel the influence of the social aspect of the decision).

The rest of the paper is structured as follows. In Section 2, we introduce a stylised model of social interaction and provide a formal evolutionary argument to motivate the development of a preference for identity consistent behaviour in such an environment. In Section 3, we argue how such a preference can be accounted for in a model of the agents' utility if we allow for more complex forms of interaction. Also in Section 3, we state the main general implications of our argument regarding observable behaviour and illustrate how they square with the evidence. Section 4 puts our discussion into the context of the existing literature. Section 5 concludes.

2 The Basic Model

In this section, we study the evolution of behaviour in a society under different assumptions about the agents' preferences. In particular, we distinguish between two cases: either agents have a preference for identity consistent behaviour, referred to as id-preference, or they do not. Although we explicitly consider evolution only to select between different behaviours for a given distribution of preferences, we ultimately will use the findings of our analysis to argue in favour of the development of a widespread preference for identity consistent behaviour itself.

In order to facilitate the exposition, the section is divided into three parts: an introduction of the formal set-up of the model (2.1); a presentation of the main results of the analysis (2.2); and, finally, a brief discussion of the results and the importance of the assumptions for their derivation (2.3).

2.1 Set-Up

Consider a society with a continuum of agents. Each of these agents, during his "life," is involved in N periods of interaction with other agents from that society. After the N periods, agents are replaced by their offspring who inherit the preferences and copy the behaviour of their predecessors. The underlying population dynamic is assumed to be payoff monotonic (Weibull, 1995), i.e. agents with higher aggregate material payoffs reproduce more rapidly.

Regarding per period interaction, we assume that agents are randomly matched in pairs to play the one-shot Prisoner's Dilemma (PD) depicted in Figure 1; specified individual payoffs indicate material rewards.[3] Moreover, after having played the PD game but before being rematched, players have

[3] Numeric payoffs are chosen only to facilitate later calculations. The PD-atypical numbers ensure that, including punishment and monitoring introduced later, no agent can obtain negative payoffs. Hence, payoffs can also be directly interpreted in terms of offspring.

the chance to enforce some (costly) punishment on their opponent.[4] More specifically, both players, knowing the outcome of the PD, have to choose between punishment (p) and no punishment ($\overline{p}$); the additional (economic) payoffs being -2 per player in the case of punishment and 0 otherwise.[5] In effect, the punishment opportunity provides a potential means to hamper the invasion of defectors into a mainly cooperative society. It allows cooperators to push the defectors' payoff below their own in the majority of the interactions (matches with a cooperator).[6] Thus, if defectors are sufficiently rare, cooperators are better off in terms of expected payoffs if they punish defectors.

		Player 2	
		C	D
Player 1	C	6, 6	3, 7
	D	7, 3	4, 4

Figure 1: The standard Prisoner's Dilemma.

Apart from the basic interaction described above, agents can identify (I) with a cooperative social norm or not ($\overline{I}$). Identification takes place prior to and independent of any later interaction. If an agent has an id-preference, identification with the norm renders any behaviour which is incompatible with the norm costly. The corresponding mental cost, denoted by c, is deducted from the agent's per period payoffs (for $c > 2$, this will ensure obedience). If the agent does not have the id-preference, per period payoffs remain unaffected by the identity decision. The behaviour prescribed by the norm is the following:

[4] Such an assumption appears justified as almost any type of social interaction offers the chance to end in a quarrel. There may not be a second opportunity for a profitable cooperation (e.g. a bargain) but there almost always is an opportunity to get into a row about the one that was, and this row usually is costly for both parties involved.

[5] If both players choose p, a per agent payoff of -4 results.

[6] Meeting a conditionally punishing cooperator, defectors obtain 5, cooperators 6.

N-1 Cooperate in the PD game.

N-2 Punish your opponent if and only if he has defected in the preceding PD game.

N-3 Be vigilant as to whether others obey the norm or not.[7]

Vigilance, i.e. the act of keeping an an eye on the behaviour of others (independent of the interaction agents themselves are involved in), is assumed to be costly. In particular, being vigilant (v) is associated with a per period material cost of ξ, $0 < \xi << 1$; choosing not to be vigilant ($\overline{v}$) is costless.

The purpose of the vigilance is to allow us to (plausibly) implement a form of higher order social punishment for deviations from the norm. More specifically, we assume that agents segregate into two classes: those who identify and comply with the norm (the I-agents) and those who do not (the $\overline{I}$-agents). The segregation takes place on the basis of observed behaviour, which we assume to be given by the identity decision as well as all actions effectively chosen in any of the interaction including the vigilance.[8] However, only the identity decision is assumed to be immediately revealed to everybody and to have an instantaneous effect on the agent's group assignment. The recognition of per period misbehaviour of any agent, by contrast, is assumed to depend on the overall vigilance of the I-agents and to have a delayed effect on the grouping.

In particular, if an I-agent in a certain period behaves in a way that is incompatible with the norm, i.e. if he defects, refrains from punishing a defector or from being vigilant, he will be found out with a probability $\alpha := \hat{\alpha} \cdot \nu$, where $\hat{\alpha} \in (0, 1)$ reflects the effectiveness of the monitoring system and $\nu \in [0, 1]$ denotes the fraction of I-agents who are vigilant. Once an agent is found out, all his future offspring is relegated to the $\overline{I}$-agents (despite the

[7] A lot of casual evidence indicates that vigilant behaviour indeed is "socially desired." Consider, for example, the ubiquitous requests on the London Underground to report any unattended luggage to a member of staff. Also, people who witnessed a crime, i.e. a break of a social norm cast in law, are commonly expected to give evidence in court.

[8] For a comment on the observability of the vigilance see Footnote 9 further below.

fact that they still may claim to identify with the norm!).[9] However, no agent is informed about whether he has been convicted before the end of period N.[10] To sum up: public non-identification with the norm always leads to immediate assignment to the $\overline{I}$-agents; endorsement of the norm only ensures placement among the I-agents if no predecessor has been convicted of a norm violation.

Given the segregation, the matching of agents is such that each period each agent is (randomly) matched with some other agent from his part of the society with probability q. With probability $1 - q$, the agent is (randomly) matched in an environment to which all agents of the society have access (cf. Figure 2). Thus, for $q > 0$ the matching is assortative in the sense of Bergstrom (2002, 2003). Yet, notice that the segregation does not generally separate cooperators from defectors but only agents who claim to identify with the norm (and have not been proved not to) from those who do not (or have been proved not to). After the N periods of interaction, agents are replaced by their offspring who inherit the preferences and copy the behaviour of their parent.

To wrap up, the overall process described above can be summarised as follows:

Step-1 Agents choose whether or not to identify with the norm in a way that is publicly observable.

Step-2 Agents decide (once and for all) how to play in the interaction; i.e.

[9] As concerns the observability of the vigilance and the potential consequences of not being vigilant, casual evidence indicates that displayed lack of concern for the general obedience to a cooperative norm will indeed be construed as lack of concern for the norm itself. This in turn may well compromise the respective person's reliability. Moreover, the parameter $\hat{\alpha}$, may also be interpreted as a measure for how likely it is to get into serious trouble from not being vigilant. If the number of possible observations, i.e. the number of interactions N, is sufficiently large, even very small values of $\hat{\alpha}$ will not affect our argument.

[10] The assumption that disobedient agents are relegated at the end of the interaction is made to simplify the subsequent analysis. If agents were relegated on the spot, an argument similar to the one to follow could be given, for example, under appropriate additional assumptions about the agents' discounting of own future (material) payoffs.

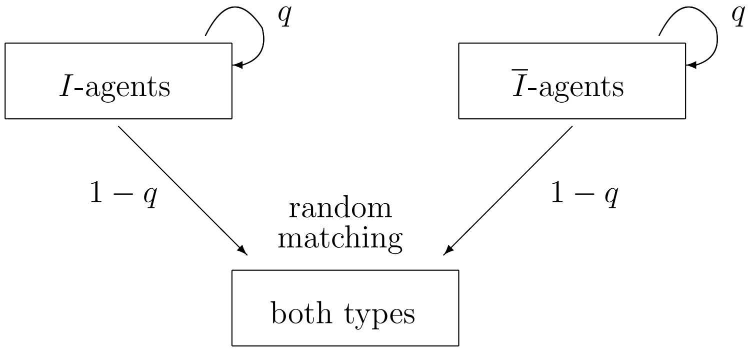

Figure 2: The matching of agents. $0 \leq q \leq 1$.

agents decide whether to be vigilant (observable), whether to cooperate or defect in the PD (observable when chosen), and whether and when to punish (observable when chosen).

Step-3 Agents are segregated according to their identity choices; yet, those for whom some predecessor has been convicted not to follow the norm are placed among the $\overline{I}$-agents irrespective of their id-choice.

Step-4 The N periods of interaction take place.

Step-5 Agents are replaced by their offspring, the number of which is determined according to the relative amount of material payoffs acquired. Each offspring inherits the preferences and copies the strategy of his parent.

Step-6 The offspring is segregated according to the parents' observed behaviour and the interaction is repeated.[11]

Notice that Step 1 and 2 – the strategy choices – mainly serve to anchor the process, e.g. if a first round after introducing the norm is taken into account. Otherwise, no actual "choice" is considered as all offspring by assumption copies the behaviour of their predecessor (cf. Step 5; see Section 2.3 for a discussion).

[11] No strategy choices are necessary as the offspring is assumed to imitate the parents.

The aim of the subsequent analysis is to study the long run evolution of society according to the above described procedure for different starting conditions, i.e. different assumptions about the presence of the id-preference and the distribution of initial behaviour in the population. In order to give a proper description of the different starting conditions, we need to specify strategies first. Each agent i's strategy can be viewed as a tuple $s_i = (s_{1i}, s_{2i})$, where s_{1i} specifies the agent's identity choice, i.e.

$$s_{1i} \in \{I, \overline{I}\},$$

and s_{2i} specifies the agent's behaviour within society, i.e.

$$s_{2i} \in \{v, \overline{v}\} \times \{C, D\} \times \{(p_C, p_D) \mid p_C,\ p_D \in \{p, \overline{p}\}\},$$

where p_C (p_D) denotes the punishment decision after observed cooperation (defection) by the opponent. The set of all (pure) strategies $s = (s_1, s_2)$ is denoted by S. The initial distribution over strategies in the population, which we consider as the starting point of our analysis, is referred to as the *starting strategy profile*; it is denoted by

$$\sigma = [\lambda_k s_k]_{s_k \in S},$$

where λ_k is the relative frequency of strategy s_k in the society, i.e. $\sum \lambda_k = 1$. Slightly abusing notation, we write $s \in \sigma$ if and only if s is played by a strictly positive fraction of agents, i.e. $s = s_k$ with $\lambda_k \neq 0$. Finally, for any repetition τ of the whole interaction (Steps 5+6), $\tau = 0, 1, ...$, the expected per period payoff of a strategy $s \in S$, given σ, is denoted by $E\pi^\tau(s \mid \sigma)$; the index τ is omitted if there is no hazard of confusion.

By assumption, the society is split up into two subsocieties according to the agents' declared identities and the behaviour of their predecessors: the I- and the $\overline{I}$-agents.[12] In order to determine the long run evolution of

[12] Recall, that in later repetitions, there may well be $\overline{I}$-agents who pretend to identify with the norm (although they will be known to only pretend).

behaviour in the society, we first attempt to determine which strategy is going to govern which part of society in the long run. Luckily, this is possible for all cases considered in the sequel. Hence, matters can be decided by means of a comparison of the expected payoffs of these strategies. We proceed with some helpful definitions.

The first thing we are interested in are those strategies that are going to dominate the respective subsocieties in the long run for a given starting strategy profile σ. In order to determine these strategies, we have to account for both the relative expected payoffs the different strategies earn (given σ) and the probability that the offspring of an agent following a certain strategy will actually remain within the respective part of society.[13] Formally, we define:

Definition 1 *Let σ be a starting strategy profile and let $s^* \in \sigma$. We say that s^* is* σ-prevailing over $\tilde{s} \in \sigma$ *among the ι-agents, $\iota \in \{I, \overline{I}\}$, if there is a repetition τ_0 of the interaction (Steps 5 and 6) such that for all $\tau > \tau_0$:*

$$E\pi_\iota^\tau(s^* \mid \sigma) \cdot P^\tau(s^* \mid \sigma) > E\pi_\iota^\tau(\tilde{s}) \cdot P^\tau(\tilde{s} \mid \sigma),$$

where $P^\tau(s \mid \sigma)$ denotes the probability that, after repetition τ, the offspring of an agent playing s remains among the ι-agents. A strategy s^* is called σ-prevailing for the I-agents *(the $\overline{I}$-agents) if s^* is prevailing over all $s \in \sigma$ for these agents.*

Apart from the question which strategy is going to prevail in which part of society, we ask whether and which part of society will dominate the other in the long run. Obviously, this question can be answered unambiguously if, from a certain repetition τ onwards, the agents of one subsociety always earn an average expected payoff which exceeds that of their counterparts from the other subsociety by (at least) a certain amount $\mu > 0$.[14]

[13] Recall that we assumed the underlying population dynamic to be payoff monotonic (cf. Step 5) and that non norm-obedient I-agents may be relegated to the $\overline{I}$-agents - despite the fact that they still claim to identify with the norm.

[14] Requiring a constant minimum distance in the payoffs is not a necessary but sufficient

Definition 2 *For any starting strategy profile σ, the I-agents are called* σ-dominant *if there is a real number $\mu > 0$ and a repetition τ_1 of the interaction, such that for all $\tau > \tau_1$ the average per period payoff earned by an I-agent $\overline{\pi}^{\tau}(I)$ is at least μ units larger than that earned by an $\overline{I}$-agent $\overline{\pi}^{\tau}(\overline{I})$. In the reverse case, we say that the $\overline{I}$-agents are σ-dominant.*

If we have been able to identify a σ-dominant part of the society as well as the respective σ-prevailing strategy, we can assess the long run evolution of behaviour in the society. The strategy that, in expectation, will govern the society in the long run is simply the one prevailing in the dominant part of society.

Definition 3 *Let σ be a starting strategy profile. A strategy s^* is called* globally σ-prevailing strategy *or σGPS for short, if it is a σ-prevailing strategy for the σ-dominant subsociety.*

For the following analysis, we assume that all agents are risk-neutral myopic payoff maximisers in the sense that they only care about their own expected per period payoffs but not about their offspring.[15] Accordingly, we restrict attention to the evolution of behaviour, given the system started with no agent playing a strictly dominated strategy - given his preferences (see Section 2.3 for a discussion of this assumption). The starting strategy profile σ is called an *undominated full support profile* if it derives from a case where both types of agents assigns positive probability to any strategy which is not strictly dominated for them. The following lemma specifies the undominated strategies for those cases considered in the subsequent propositions.

condition which ensures the absence of difficulties in the limit. We choose it as it is most convenient for the later argument.

[15] This assumption essentially ensures that for agents without the id-preference it is strictly dominated to identify with *and* follow the norm. It is necessary as we do not consider any other kind of discounting of future payoffs. Yet, rational agents would always trade off immediate gains, e.g. from not being vigilant, with the potential future consequences, i.e. a possibly lower payoff for their offspring. Thus, with an appropriate "standard" discounting of future payoffs, following the norm would still be dominated for these agents - if N is not too small.

Lemma 1 *If the mental cost from norm-disobedient behaviour is sufficiently large, i.e. $c > 2$, then the only strategies which are not strictly dominated for the agents with an id-preference are:*

$$s^* := (I, v, C, \overline{p}, p), \quad \tilde{s} := (\overline{I}, \overline{v}, C, \overline{p}, \overline{p}) \quad \textit{and} \quad \overline{s} := (\overline{I}, \overline{v}, D, \overline{p}, \overline{p}).$$

For the agents without an id-preference, the strategies

$$\hat{s} := (I, \overline{v}, C, \overline{p}, \overline{p}), \quad s' := (I, \overline{v}, D, \overline{p}, \overline{p})$$

as well as $\tilde{s}$ and $\overline{s}$ specified above are not strictly dominated (but s^* is).

2.2 Results

We now proceed to state the results of our analysis. First, we consider the standard case, i.e. a society of agents who (initially) all lack the id-preference. The result in this case is immediate and in fact the usual one, i.e. in the end everybody will defect and neither be vigilant nor punish defectors. Thus, simply introducing the norm (including a costly option to achieve assortative matching) is not enough to establish cooperation. We state Proposition 1 without proof.

Proposition 1 *Assume that none of the agents of a society has an id-preference. Then, for any undominated full support profile σ, the σ-prevailing strategies for both parts of society are such that all agents in the long run play $\overline{s}_2 := (\overline{v}, D, \overline{p}, \overline{p})$. The average per agent per period payoff is given by $\overline{\pi} = 4$.*

The next proposition considers the other extreme case, namely the one in which all agents of a society (initially) have the id-preference. It states that for norm obedient behaviour to prevail in such a society it is sufficient to require that agents are sufficiently separated. Intuitively, separation guarantees that the I-agents (who all cooperate) earn the higher expected payoff as it ensures that the relatively disadvantageous match of I-agents with potential defectors among the $\overline{I}$-agents occurs sufficiently infrequent. Given that, however, norm obedient cooperation obviously is advantageous.[16] The proof of Proposition 2 is deferred to the appendix.

Proposition 2 *Assume that all agents of the society have an id-preference and that the cost of identity inconsistent behaviour is $c > 2$. Moreover, let σ be an undominated full support profile. Then, $s^* = (I, v, C, \overline{p}, p)$ is the unique σGPS, if I- and $\overline{I}$-agents are sufficiently separated, i.e. if $q > \frac{3+\xi}{5}$. The resulting long run average per agent per period payoff is $\overline{\pi} \approx 6 - \xi$.*

Finally, we reconsider the above setting under the assumption that initially both types of agents, i.e. those with as well as those without the id-preference, are present in the society. The potential trouble in this case derives from the fact that for those agents without the id-preference, it is always strictly dominant not to be vigilant and not to punishing defectors. Yet, they may well claim to identify with the norm and, hence, start out among the I-agents. In that case, their expected per period payoff is larger than that of the I-agents with an id-preference playing s^*.

The point of the following proposition is to show that also under these diverse starting conditions agents with an id-preference playing s^*, nevertheless, will prevail. To obtain such a result, we again have to assume that the different subsocieties are sufficiently separated and that the number of interactions N is sufficiently large. Intuitively, a large number of interactions ensures that agents without the id-preference who only pretend to have identified with the norm but do not act accordingly[17] are sufficiently likely to be noticed and relegated to the $\overline{I}$-agents so that the I-agents eventually consist of norm obedient cooperators only. Separation again is necessary to guarantee norm obedient cooperators a sufficiently high payoff through (eventually) frequent matches among themselves.[18] The proof of Proposition 3 can be found in the appendix.

[16] Yet, recall, that offering assortative matching without the id-preference is not sufficient to establish cooperation if the assortativeness is costly (cf. Proposition 1).

[17] Recall that being vigilant is strictly dominated for agents without the id-preference.

[18] cf. Footnote 16.

Proposition 3 *Assume that a fraction of $1-\gamma$ of the agents of a society have an id-preference whereas the others do not, $0<\gamma<1$, and that $c>2$. Moreover, let σ be an undominated full support profile. Then, s^* is σGPS, if $q>\frac{3+\xi}{5}$ and if N is large, i.e. if $N>N^*(q,\hat{\alpha},\sigma)$, where N^* is decreasing in $\hat{\alpha}$, and increasing in q and λ_0^* (the relative frequency of s^* in σ).*[19] *Thus, in the long run the I-agents with an id-preference (playing s^*) will dominate the society and will earn an average per period payoff of $\overline{\pi}\approx 6-\xi$.*

From the above proposition we conclude that, from an evolutionary perspective which also takes into account the competition of different social groups for scarce resources, a (widespread) individual preference for norm- or identity-consistent behaviour is advantageous as it enables the development of a cooperative trait in society and, hence, leads to an increase in the respective agents' individual fitness (as compared to agents in non-cooperative societies). No assumptions about general individual concerns for the utility of other agents are necessary.

2.3 Discussion

In the above, we have studied the evolution of behaviour in a society of agents which either possess or lack an id-preference. Our results essentially show that a cooperative norm can be established in the society if and only if at least some of the agents initially have a preference for identity consistent behaviour. In these cases, norm obedient behaviour, in expectation, will prevail (under certain conditions on the interaction) and the agents with a preference for identity consistent behaviour, who follow the cooperative norm, will dominate the society in the long run.

As already indicated in the introduction, the result in favour of norm-obedient behaviour and, hence, the id-preference derives mainly from the assortativeness of the matching process. This allows norm-obedient cooperators to sufficiently separate from either do not cooperate at all or who at

[19] More specific conditions on N^* are given in the proof of this proposition. To convey a feeling for the requirements, for a uniform starting profile σ, $q\approx\frac{3}{5}$, and $\hat{\alpha}\approx 0.1$ values about $N^*=25$, which we consider to be very few interactions per lifetime, are sufficient.

least do not contribute to sustain the norm. However, for the assortativeness to gain effect, norm-obedient agents have to be vigilant - which is costly - in order to recognise and separate from agents who undermine at (least some part of) the norm. Thus, different form other approaches (e.g. Bester and Güth, 1998; Nowak and Sigmund, 1998), the cost of the mechanism which eventually drives the selection process in the desired direction is taken into account (see also Henrich, 2004). Moreover, although we also consider only a two player interaction for our main argument (a PD game), the reference to the social norm allows us to easily extend our result to cooperative behaviour on a larger scale (cf. Section 3).

Another point we want to comment on is the assumption that agents only use strategies that are not strictly dominated given their preferences. Commonly, evolutionary models analyse how a population evolves under different starting conditions - including strictly dominated strategies. Yet, they usually do so based on the assumption that all behaviour is genetically determined and not chosen deliberately. This, however, appears to be rather implausible if we look at human social behaviour. For the purposes of this paper, we therefore only assume that *agents inherit the preferences* of their predecessors but not their behaviour. As regards behaviour, we prefer to think of agents as having some (arbitrary) belief about the behaviour of others and choosing a best response against this belief - given their preferences. As a consequence of this, we exclude strictly dominated strategies a priory (they are never a best response), but otherwise allow for any undominated full-support profile to begin with (agents may have whatever beliefs they wish). Beliefs then can be thought of as being passed on from parent to offspring, who act accordingly. For the sake of argument, this transmission of beliefs within the model is implicitly assumed to work perfectly (in effect, the offspring always copies the strategy of the parent). Notice, however, that this assumption is not restrictive. In fact, also an occasional copying of other parents' beliefs, with the respective consequences for behaviour, would not affect the outcome of our analysis as we consider full support starting profiles anyway.

Furthermore, we want to emphasise that not all assumptions made are technically strictly necessary for our argument. In particular, results similar to the ones above can be obtained without requiring punishment to be part of the norm. Being observable, vigilance alone would be sufficient to establish the necessary separation between norm-obedient agents with an id-preference (who cooperate and are vigilant) and other agents (who at most cooperate but never are vigilant). Yet, introducing the punishment has two advantages. First of all, the additional punishment makes it harder for defectors to invade a mainly cooperative society. Thus, in contrast to other purely costly ingredients that could be incorporated into the norm, e.g. burning some endowment prior to each interaction, punishment effectively helps mainly cooperative societies to *save* resources. Accordingly, a norm with punishment has a comparative advantage over one without once cooperation becomes prevalent. Moreover, adding the punishment also seems to be more plausible from an empirical point of view as cooperative social norms commonly do not only prescribe cooperation itself but also punishment of different order (i.e. direct punishment of defection as well as - higher order - punishment of general norm-disobedience, e.g. unenforced punishment;[20] cf. Section 3.2). The stylised norm employed here, which combines direct punishment of defectors with indirect punishment of general norm-violators through the vigilance, allows us to capture both aspects.

Finally, it is interesting to note that neither infrequent random matching with agents from other societies nor infrequent migration of agents would affect our argument. This again is due to the fact that we considered only full support starting profiles.[21] Societies that comprise merely agents without an id-preference (considered in Proposition 1) might fail to settle for defection, though, if norm-obedient cooperators with an id-preference could invade and sufficiently separate themselves. This, however, would only further the eventual dissemination of the id-preference and, hence, strengthen our previous argument.

[20] For example, witnesses of a crime are usually obliged to give evidence in court (with few exceptions), and perjury itself commonly is liable to prosecution.

[21] The same holds true for mutations as long as no strictly dominated strategies, i.e. strategies excluded by Lemma 1, are played.

3 On How to Account for Identity in Utility

In the previous section, we have outlined an argument for the general advantage of an individual preference for identity-consistent behaviour within a simplified model of social interaction. Building on this, we now want to consider more closely, albeit less formally, how such a preference can be satisfactorily accounted for in the modelling of the agent's utility if the social environment allows for more complex patterns in the interaction (3.1). Finally, we discuss the main behavioural implications of our considerations and show how they square with the evidence (3.2).

3.1 General Discussion

So far, we have assumed that for those agents with an id-preference self-inconsistent behaviour is associated with a fixed mental cost c that is high enough to ensure obedience to the proposed norm. Assuming a fixed cost was sufficient for our argument as the matching process was sufficiently randomised. Yet, if we allow for patterns in the matching process, e.g. through occasional long episodes of interaction with members from different societies (e.g. as salesman or envoy), things change. Other societies may not follow a cooperative convention but may primarily consist of defectors. And if there is a chance of longer episodes of interaction with non-cooperative agents, a fixed cost from disobedience with an internalised cooperative norm, which ensures cooperative behaviour of an agent, may be considerably detrimental.[22] Thus, increasing correlation in the matching process, e.g. through increasing levels of inter-group interaction (which certainly accompanied the development of human societies), calls for a more flexible psychologic mechanism enforcing cooperation where appropriate and defection where necessary.

Such a flexibility can be achieved, though, if we allow the cost from

[22] Recall that own cooperation given opponent defection plus the additional punishment and vigilance results in overall economic payoffs of $1 - \xi$ for the cooperator/punisher and $+5$ for the defector, whereas the average payoff of an agent in the cooperative home society is $6 - \xi$. Hence, longer episodes of matching with a defector can easily become detrimental in terms of individual fitness.

inconsistent behaviour $c(.)$ to depend on the context (do norms apply or not?) and to gradually adjust to the player's past experience. Dependence on the context allows the agents to agree on the general invalidity of a norm in a certain context (e.g. cooperation with an opponent in a competitive game of sports, or more drastically in war, is usually considered inappropriate). In terms of our previous model, the offspring of agents who defect in a commonly agreed context, may not be relegated to the $\overline{I}$-agents.

Dependence on (recent[23]) past experience in turn allows the agents to respond in a more deliberate way to the behaviour of their opponents even if the general context is cooperative. If, for example, the agent has suffered from a repeated defection by his opponents in the recent past, he now is able to "learn" the inapplicability of the cooperative norm in the respective context and may - with time - find it easier to adjust and defect himself.[24] Returning to a more cooperative environment which allows for more positive experience, however, the cost of non-cooperative behaviour may adjust back such that the agent returns to cooperative behaviour. Again, in terms of our model, the agents of a society may, for example, agree to relegate only the offspring of repeatedly disobedient agents.[25] In that case, if the readjustment to a cooperative environment is sufficiently fast, a more flexible cost from norm disobedient behaviour appears to be preferable.

In the following, we propose a stylised model of utility which still accounts for the main parts of our prior argument (cf. Section 2) but nonetheless offers the desired flexibility.

Assume that the agent's utility, which he tries to maximise, is given by a weighted average of the economic rewards π_i and a psychologic component $c(.)$ which is related to the congruence of the agent's behaviour with the internalised norms,[26] that relative weights are fixed (idiosyncratic) constituents

[23] It appears reasonable to assume the impact of more recent experience to be stronger in order to allow for a more flexible adjustment to potential patterns in the matching.

[24] For a more extensive discussion of this point, see Section 3.2.

[25] Also, potential gains from trade may outweigh the risk of such occasional norm-violations in expected terms.

[26] This again is not to say that the agent's economic benefit from a certain outcome π_i

of the agent's identity, and that the cost from inconsistent behaviour is omnipresent and equal for all agents.[27] The clear distinction between identity dependent relative weights and a generally valid cost from norm disobedience is made for expositional purposes only. Yet, it also seems plausible as it allows us to think of $c(.)$ as some commonly agreed upon standard to which agents can subscribe at an individual degree (relative to their economic self interest captured by π_i). Thus, only those agents with a (sufficiently strong) preference for identity consistent behaviour bother about the psychologic component whereas others do not - depending on the relative weights.[28]

Flexibility in the psychologic component of utility, then, is achieved by assuming $c(.)$ to depend not only on the agent's current behaviour, s_i, and its compatibility with the norm, but also on past experience, i.e. on the history until period t, denoted by $h_i(t)$, and the specific type of interaction G; i.e.

$$c_i(.) = c(s_i, h_i(t), G),$$

where the absolute value of $c(.)$ is lower the more negative experience the agent has gathered in the (recent) past and the less the general context, G, is evocative of (in our case) cooperative social norms.[29] Putting things together, we obtain a utility function of the following form:

$$u_i(s_i, s_{-i}, h_i(t), G) = (1 - \rho_i) \cdot \pi_i(s_i, s_{-i}) + \rho_i \cdot c(s_i, h_i(t), G),$$

will vary. In fact, it will remain fixed and unaffected by any mental discomfort. However, to capture individual incentives for decision making, using a relative approach appears reasonable.

[27] This implicitly presumes that the set of social norms available is the same for all agents and that social norms can only be accepted on an *all or nothing* basis. Yet, adding one component for each norm adds nothing to our argument but notation.

[28] Notice that such a change does not affect the validity of the previous argument in favour of the existence of preferences for identity consistent behaviour. Only the formal argument becomes more involved if we allow for more widespread weights instead of assuming weights to be either equally split or completely focused on economic rewards.

[29] In general, there will be far more social conventions and stereotypes than only cooperative norms (see e.g. Akerlof and Kranton, 2000). Yet, to keep the exposition simple, we confine our analysis to cooperative behaviour and cooperative norms which have attracted so much interest in the recent past (cf. Section 4.1).

where $\rho_i \in [0,1]$ is the individual specific relative weight of the identity component in the utility, which we treat as a fixed constituent of player i's identity, and s_{-i} denotes the current strategies of the other players.[30]

It is immediate that generally cooperative agents who base their decisions on a utility function as the above are less prone to be exploited repeatedly as we assumed $c(.)$ to diminish with negative experience. Nevertheless, the reduction of $c(.)$ has to be gradual in order to prevent defectors invading the society. If economic incentives dominated too quickly, a single defector can trigger a cascade of defections and, thus, a breakdown of the cooperative convention. Each agent, having met the defector, afterwards would follow the purely economic self-interest and would defect himself (at least once), thereby prompting at least one additional defection, etc.

Summing Up If the matching process allows for patterns, e.g. through longer episodes of matching with agents from other (unknown) societies, a more flexible influence of the preference for identity consistent behaviour becomes advantageous. Such a flexibility can be achieved if we assume only the agent's relative preference for identity consistent behaviour to be fixed, but allow the respective cost associated with inconsistent behaviour to depend on the general context and (recent) past experience. In particular, if a decision context bears little connection to cooperative social norms or sharing rules (e.g. an auction / a competitive game of sports) or if the agent's past experience does not square with the norm (e.g. if the agent has been repeatedly exploited in a Prisoner's Dilemma), this cost ought to be low and behaviour should mainly be governed by material self-interest. These effects are independent of the size of the population or the number of participants of the interaction.

3.2 Implications and Evidence

In the preceding discussion, we have argued that, in a more realistic context, also norm obedient agents with a preference for identity consistent behaviour

[30] The discussion in Section 2 corresponds to the case of $\rho_i = 0.5$ and a fixed c.

should not be expected to act as uncontingent cooperators. Instead they should adjust behaviour according to their past experience and, more generally, to how evocative the respective context is of cooperative social norms. In the remainder of this section, we discuss some of the main implications of our argument for cooperative laboratory experiments. As we will see, these implications are largely consistent with the evidence.

Implications

Consider again the stylised utility function proposed in the previous subsection:

$$u_i(s_i, s_{-i}, h_i(t), G) = (1 - \rho_i) \cdot \pi_i(s_i, s_{-i}) + \rho_i \cdot c(s_i, h_i(t), G).$$

In order to make our point, we restrict attention to the case of laboratory studies of repeated Prisoner's Dilemma or repeated Public Goods games without punishment or potential segregation of agents. We confine ourselves to Prisoner's Dilemma and Public Goods games (without punishment) as for these games the players' (stage game) actions intuitively can be thought of as being ranked on a scale from 0 (purely selfish) to 1 (purely cooperative), i.e. $s_i \in [0, 1]$.[31] Moreover, in view of cooperative social norms, the actions available for these games entail *competing interests* in that incentives from material self-interest are strictly opposed to cooperative behaviour. Put differently, irrespective of any other player's behaviour, each player's material payoffs π_i are strictly decreasing in the cooperativeness of his behaviour s_i.

As regards the other ingredients of the above utility function, we assume that G, i.e. the cooperativeness of the general context, is fixed and equal for all players,[32] and that past experience h_i also is measured on a zero-one scale, i.e. $h_i \in [0, 1]$ with larger values of h_i indicating a more cooperative past experience. If the agent has no prior experience with the respective context,

[31] For example, for the Prisoner's Dilemma, defection can be associated with $s_i = 0$ and cooperation with $s_i = 1$; mixed strategies in between. Similarly, for Public Goods games, s_i can be associated with the percentage of the endowment contributed to the public good.

[32] Evidence for the general type of context effects indicated above is provided at the end of the "Evidence" part of this section.

we set $h_i(1) = 1$. Otherwise, h_i is determined by the average cooperativeness of all actions (but the agent's own one) observed in the previous round. Thus, for given G, the psychologic cost $c(s_i, h_i \mid G)$ can be written as a function $c(. \mid G) : [0,1] \times [0,1] \to \mathbb{R}_-$. Assuming $c(.)$ to be twice differentiable, our previous arguments can be translated into the following requirements:

$$\frac{\partial c}{\partial h} < 0, \; \frac{\partial c}{\partial s_i} > 0, \quad and \;\; \frac{\partial^2 c}{\partial h \partial s_i} > 0;$$

i.e. the more (less) cooperative the agent's past experience (own actual behaviour) the larger the mental cost of norm-disobedience, i.e. the smaller is c; and the less cooperative past experience is the weaker is the effect of own current behaviour (as the inappropriateness of the norm is learned).

Finally, let us assume that players are drawn from a continuum of agents for which the identity parameter ρ is distributed according to some cdf F which possesses a continuous density f with full support, i.e. ρ varies from a complete lack of concern for social norms ($\rho_i = 0$) to almost perfect norm conformity ($\rho_i = 1$) in the pool of players. Then, we can state the following general implications of our argument.

Implication 1 *For a repeated Prisoner's Dilemma or Public Goods game with competing interests but with neither punishment nor potential segregation (e.g. because the matching is fully random), averaging over many observations, our arguments predicts that:*

1. *Cooperation rates, i.e. average values of s, decrease over time if for all agents $h_i(1) = 1$; and the larger the number of agents whose behaviour can be observed or inferred in the course of the interaction, the more pronounced the effect will be.*

2. *Cooperation rates are decreasing in the size of the material rewards from non-cooperative behaviour.*

Intuitively, Implication 1.1 follows from the full-support assumption about the distribution of ρ_i in the population.[33] Thus, if the behaviour of a large number of agents can be observed, there will (in expectation) always be someone for whom material payoffs, π_i, immediately dominate. Accordingly, this agent will choose $s_i \approx 0$. In the next round, then, the past experience with (*unpunished*) defection will reduce the psychologic cost from own defection for all agents because the agents "learn" that norm obedient cooperation is inappropriate as it cannot be enforced. Hence, over time average defection rates will increase (as agents cannot avoid defectors either), and so forth; until only those with $\rho_i \approx 1$ keep on following the norm. Reduced observability regarding other players' behaviour, however, may impede this unravelling. A more formal derivation of Implication 1.1 can be found in the appendix. Implication 1.2 follows immediately from the assumptions and the specification of the utility function.

Evidence

Experimental evidence from the lab, in fact, appears to be largely consistent with the above implication as well as with the general thrust of the argument presented in this paper.

As regards Implication 1.1, for example, contribution rates in repeated Public Goods games (without punishment), where at least aggregate behaviour of others can be inferred, are commonly found to decline with repetition (see, e.g. Guala, 2005). More specifically, Duffy and Ochs (2005) analyse the evolution of behaviour in a repeated Prisoner's Dilemma game (also without punishment) both in the cases of fixed and random pairings. Not only do they find declining average cooperation rates in both cases; the effect is also found to be less pronounced in the case of fixed pairings than in the case of random parings (where more information about other agents becomes accessible with time). In fact, regarding fixed pairings Duffy and

[33] A similar statement can be shown to hold, for example, if we assume ρ to be constant but instead choose $h(1)$ to be drawn from a continuous full support distribution ranging from 0 to 1 reflecting the differences in the players' past experience. What essentially is necessary for our argument is a diverse perception of the relative strength of psychologic and monetary incentives, and a subsequent decrease of c for all agents.

Ochs write (p.14/15) that "*the decline in aggregate frequencies of cooperation over time is due to the presence of just a few player types, who very frequently choose to defect [...].*" In other words, more cooperative agents indeed appear to subsequently adjust to a non-cooperative environment.

In contrast, contribution rates in Public Goods games indeed are found to be higher if punishment opportunities are available (Fehr and Gächter, 2000). Moreover, there is evidence that it is in particular cooperators who make use of such an option in order to punish defectors; especially so, if the punishment comes at a considerable cost to the punisher himself (Falk et. al, 2005).[34] Although not explicitly mentioned among the above implications, these observations are very reassuring as they strongly support our general line of argument.

As regards Implication 1.2, Camerer (2003, p.46) points out that increasing the payoffs from defection in a Prisoner's Dileamma game (given cooperation of the opponent), in fact, leads to an increase in aggregate observed defection rates. And a similar effect is reported for Public Goods games. Here a decrease in the marginal returns from a contribution to the public good is found to be accompanied by a decrease in aggregate contribution rates (Camerer, 2003, p.46).

Last but not least, we want to emphasise that there also is extensive evidence which strongly indicates that behaviour indeed depends on the context in the way indicated above, i.e. that the framing of decisions according to a social paradigm which is reminiscent of some cooperative norm increases cooperation rates.[35] To cite just a few, gift exchange games, which are usually framed in a labour context, are well known for comparably high rates of cooperation (e.g. Brown et al., 2004; Fehr et al., 1998; Gächter and Falk, 2002). In contrast, increased anonymity in dictator games, which clearly reduces the social aspects of the decision, is found to decrease the amount of

[34] Notice that, if the cost incurred by the agent being punished are higher than those that accrue to the punisher, also defectors may find punishment attractive as it increases relative fitness.

[35] See also Wichardt (2005b) for a discussion of the importance of context dependence for the assessment of the significance of laboratory findings.

money left (Hoffman et al., 1994). Similarly, minimum acceptable outcomes in ultimatum games are found to be lower in case of randomised proposals made by a computer (low social aspect) than in case of human proposals (Blount, 1995).

Discussion

All in all, the evidence cited above is very much in line with the general implications of our model of utility and the discussion of the context-dependent influence of social norms on individual decision making. Of course, part of this evidence is also consistent with the various models of fairness and reciprocity which are briefly discussed in the next section. This, however, poses no problem to our argument. Once the importance of cooperative social norms for a certain context is undoubted, an application of these models on an *as if* basis indeed seems justified. Yet, in our view, one conclusion that can be drawn form the evidence is that any model of utility that eventually aims to capture what is commonly referred to as social preferences (such as fairness considerations) should be context dependent - at least to some degree (cf. Wichardt, 2005b). Different from the existing models, the model of utility proposed in this section satisfies this requirement. It emphasises the agent's (selfish) general desire to act in accordance with his identity, i.e. the social norms internalised by the agent. In conjunction with contextual aspects regarding the relevance of these norms, it thereby enables us to roughly assess ex ante when social preferences are likely to figure prominently in the agents' decision making and when they are not. Thus, it may help to clear up the picture regarding the general relevance of such preferences for economic decision making.

4 Related Literature

Finally, we put our analysis into the context of the existing literature. In particular, we discuss how it relates to other approaches which account for immaterial incentives in utility such as models of fairness and reciprocity (4.1) or the work of Akerlof and Kranton on economics and identity (4.2).

As a last step, we briefly indicate how it is connected to the research on cognitive dissonance in psychology (4.3).[36]

4.1 Fairness and Reciprocity

As mentioned earlier, many attempts have been made to account for the seemingly irrational traits in human behaviour such as fairness concerns and reciprocity through modifications in the concept of utility. Most prominent among these are the models of fairness (Rabin, 1993), inequity aversion (Fehr and Schmidt, 1999) and, more recently, of reciprocity (e.g. Dufwenberg and Kirchsteiger, 2004, or Falk and Fischbacher, 2006; see also Charness and Rabin, 2002).

The main feature these models aim to capture is the seeming concern of agents for the well-being of others. Accordingly, all these approaches incorporate the utility that accrues to other individuals from the interaction into the utility function of the agent. The particular specifications are different, though. Roughly speaking, they can be split into two categories according to whether also the beliefs about the intentions of other players are assumed to influence the agent's utility or not.

A prominent example for the latter case is the Fehr and Schmidt (1999) model of inequity aversion. In this type of model, the agents are simply assumed to have an additional social preference for the equal (or fair) split (which has to be balanced with pure self-interest). Such an unconditional preference to forgo economic benefits, however, is difficult to justify from an evolutionary perspective.

The models in the other category (to which all other papers cited above belong), therefore, try to circumvent the assumption of unconditional goodwill. To this end, they assume the agent's utility, as derived from the other

[36] Also other strands have been pursued to account for the observed inconsistencies with the rational agents paradigm. For a review of models aiming to resolve the inconsistencies taking complexity constraints on human cognition into account, see Rubinstein (1998). For a review of learning models, see Fudenberg and Levine (1998).

player's payoff, to be aligned with the intentions the agent believes the other player to have towards him.[37] Roughly speaking, if Player A ascribes positive intentions to Player B, he will benefit from being kind to B; if, however, A ascribes negative intentions to B, he will benefit from being mean to B. Thus, in a sense, players reciprocate the intentions of their counterparts so that "well-known" defectors and other reciprocators can be treated differently. In order to capture such considerations, related evolutionary models usually assume a (costless) labelling of agents according to past behaviour (e.g. Nowak and Sigmund, 1998); i.e. reciprocators only cooperate with good labels. Yet, apart from other difficulties, these models commonly fail to give a proper account of cooperative behaviour if the interaction is anonymous or one-shot or if the underlying population is large, so that a single agent has to keep track of the labelling of many others (cf. Henrich, 2004).

The present argument, which allows us to largely avoid these difficulties, offers an intuitive way to nonetheless motivate (e.g.) fairness concerns in the agents' utility considerations for various instances (i.e. whenever the respective norms are evolutionary advantageous). Moreover, through the reference to social norms, it enables us to (roughly) assess ax ante when such considerations are of less relevance to the agents' decision (depending on the salience of the respective norms). Accordingly, we do not view our approach as a general substitute for any of the others - including the rational paradigm - but rather as an intuitive attempt to put them into perspective.

4.2 Economics and Identity

The first ones to emphasise the importance of identity for economic analysis were Akerlof and Kranton (2000). In their seminal paper, they conclusively outline the behavioural consequences that arise if an individual distaste for identity threatening acts, both by himself and by others, is taken into account. In a later paper, Akerlof and Kranton (2005) extend their discussion and emphasise the positive effects of a strong association with a group on

[37] These models all draw on the notion of psychologic games as introduced by Geanakoplos et al. (1989).

the degree of cooperative behaviour with/within that group.[38] Concerning utility they write that: “a person’s identity describes gains and losses in utility from behaviour that conforms or departs from the norms for particular social categories in particular situations” (Akerlof and Kranton, 2005, p. 12) - which, in fact, is very much in line with the spirit of our analysis.

More recently, the discussion of identity has also been taken up, for example, by Benabou and Tirole (2006a, 2006b). In the 2006a paper, they demonstrate nicely how a social signalling aspect combined with an individual concern for prosocial behaviour, which may be interpreted in terms of identity, affects contributions to social goods in an intuitive way. In their later paper, Benabou and Tirole (2006b) provide an illuminating account of how various puzzling economic as well as social phenomena (e.g. taboos) can be rationalised under the assumption that decision makers tend to infer past motivations, i.e. information about their identity, from past choices.

4.3 Cognitive Dissonance

Cognitive Dissonance is a psychologic phenomenon the discussion of which can be traced back to Festinger (1957).[39] The term cognitive dissonance refers to the cost an individual incurs if he, out of his own volition, behaves in a way that is incompatible with (or threatens) his overall perception of self-integrity - his identity; e.g. to smoke despite a health-conscious self-image or to defect despite an internalised cooperative norm. The discrepancy between ideal and actual behaviour, according to psychology, causes a kind of mental distress called cognitive dissonance which agents will tend to avoid.

Clearly, the concept of cognitive dissonance is closely related to the Akerlof and Kranton discussion about economics and identity as well as to our analysis. In particular, the apparent evidence that self-inconsistent behaviour indeed causes mental distress very much supports the assumption that such

[38] See Wichardt (2005a) for a discussion of how this argument extends to the case where the individual’s identity is based on the association with more than one group.

[39] See Harmon-Jones and Mills (1999) for a more recent treatise. See Akerlof and Dickens (1982), for a discussion of the economic relevance of this phenomenon.

metal costs are also taken into account in the agent's economic decision making. Nonetheless, we are not aware of any other evolutionary approach to this issue. Yet, it is of course very reassuring that psychologists are able to identify today what we just claimed to be evolutionarily advantageous.

5 Concluding Remarks

In this paper, we have outlined an evolutionary argument in favour of an individual preference for identity-consistent behaviour, where identity refers to the social norms and values internalised by the individual. Based on this argument, we have proposed a stylised model of utility which basically assumes that what is relevant for economic decision making beyond material payoffs is obedience with social norms. Moreover, the degree to which such additional considerations influence utility and, hence, individual decision making is assumed to reflect the salience of these norms in the respective context as well as the individual's general focus on these aspects (determined by his or her identity). As we have argued, the general implications of such a specification of utility are very much in line with the evidence.

In our view, one advantage of the current, indirect approach to justify a cooperative trait in human behaviour (via identity and norms) is that it puts a stronger emphasis on the social aspects of human interaction (e.g. social norms). As we did not set out to prove the evolutionary advantage of cooperative behaviour per se, we were able to avoid many of the shortcomings commonly connected to such attempts, such as (e.g.) a costless mechanism separating "the good" from "the bad" or a restrictedness of the argument to two person interactions or small groups (cf. Henrich, 2004).[40] Moreover, the current approach allows us to accommodate aspects of both the rational

[40] cf. Section 1. Another point emphasised by Henrich (2004) is that most existing approaches fail to explain why especially humans are so much more cooperative than almost all other animals. Although a discussion of this point is beyond the scope of this paper, it is interesting to note that the complexity of the social mechanism employed in this paper, i.e. the norm including different levels of (costly) punishment, may be interpreted as necessitating skills we would be inclined to only ascribe to us (humans) but not to other species.

paradigm and the more recent models of fairness and reciprocity according to the context of the decision. In particular, the pure rational agents model will prove most valuable in neutral, purely competitive circumstances (e.g. auctions).[41] As soon as social norms (or stereotypes) become more pronounced, however, additional incentives related to identity have to be accounted for. This is when, in our view, the models of fairness and reciprocity, which usually fit the data quite well in these instances, come into play.

Yet, it is only when we know *why* the agents care about (e.g.) fairness that we are able to ex ante assess the relative strength of such considerations in a certain context. And it is only when we know what influences the agents' beliefs about the intentions of their opponents - e.g. awareness of the social aspects of a decision and the relevant norms - that we are able to make reliable (though rough) predictions about how this will affect the agents' decision making. The present discussion, which intended to shed some light on these issues, may help to clarify which model is most appropriate under which circumstances or whether even new models accounting for different social norms are necessary.

Appendix

Notation In the sequel, relative frequencies of the different strategies in the society are denoted by $\lambda^*,\ \tilde{\lambda},\ \overline{\lambda},\ \hat{\lambda},\ \textit{and}\ \lambda'$, where λ^* refers to strategy s^* and so forth.

Proof of Proposition 2

From Lemma 1, we know that for $c > 2$ any undominated full support profile σ will assign positive probability to the strategies s^*, $\tilde{s}$, and $\overline{s}$ only (as all agents are assumed to possess an id-preference). Hence, s^* is the σ-prevailing strategy for the I-agents. What remains to be shown is that, under conditions specified in Proposition 2, the I-agents are the σ-dominant for any undominated full support profile σ.

[41] Complexity considerations regarding decision making are deliberately neglected here.

In order to do so, we first show that $\overline{s}$ will prevail among the $\overline{I}$-agents. Given the matching procedure the expected per period payoff of $\overline{s}$ is given by

$$E\pi(\overline{s} \mid \sigma) = q \cdot (4 \cdot \frac{\overline{\lambda}}{\overline{\lambda} + \tilde{\lambda}} + 7 \cdot \frac{\tilde{\lambda}}{\overline{\lambda} + \tilde{\lambda}}) + (1-q) \cdot (5\lambda^* + 4\overline{\lambda} + 7\tilde{\lambda}).$$

The expected per period payoff of $\tilde{s}$ is given

$$E\pi(\tilde{s} \mid \sigma) = q \cdot (3 \cdot \frac{\overline{\lambda}}{\overline{\lambda} + \tilde{\lambda}} + 6 \cdot \frac{\tilde{\lambda}}{\overline{\lambda} + \tilde{\lambda}}) + (1-q) \cdot (6\lambda^* + 3\overline{\lambda} + 6\tilde{\lambda}).$$

A payoff comparison yields that $E\pi(\overline{s} \mid \sigma) > E\pi(\tilde{s} \mid \sigma)$ is equivalent to

$$q > \frac{\lambda^* - \overline{\lambda} - \tilde{\lambda}}{1 + \lambda^* - \overline{\lambda} - \tilde{\lambda}} \; .$$

As

$$\frac{\lambda^* - \overline{\lambda} - \tilde{\lambda}}{1 + \lambda^* - \overline{\lambda} - \tilde{\lambda}} < \frac{1}{2} \; ,$$

this is always satisfied given the restriction on q. Thus, $\overline{s}$ is σ-prevailing among the $\overline{I}$-agents for all undominated full support profiles σ.

To show that the I-agents are dominant and, hence, that s^* is the unique σGPS, it suffices to show that s^* earns higher expected payoff than $\overline{s}$ if only these two strategies are present. In this case expected per period payoffs are given by

$$E\pi(s^* \mid \sigma) = 6q + (1-q) \cdot (6\lambda^* + \overline{\lambda}) - \xi,$$

and

$$E\pi(\overline{s} \mid \sigma) = 4q + (1-q) \cdot (5\lambda^* + 4\overline{\lambda}).$$

Again, using that $\lambda^* = 1 - \overline{\lambda}$, a payoff comparison shows that $E\pi(s^* \mid \sigma) > E\pi(\overline{s} \mid \sigma)$ is equivalent to

$$q > 1 - \frac{2 - \xi}{1 + 4\overline{\lambda}} \; .$$

As we have not developed any restrictions on $\overline{\lambda}$ during the repetition of the interaction, $\overline{\lambda}$ may get close to 1. However, even for $\overline{\lambda} = 1$, the above condition for q is satisfied if $q > \frac{3+\xi}{5}$, as is required in the proposition. Thus, for any undominated full support profile σ, s^* is the unique σGPS under the conditions specified in the proposition. q.e.d.

Proof of Proposition 3

In order to prove Proposition 3, we first show that, under the conditions specified in the proposition, s^* is σ-prevailing among the I-agents and that $\overline{s}$ is σ-prevailing among the $\overline{I}$-agents for any undominated full support profile σ. From the proof of Proposition 2 it then follows that s^* is σGPS as we imposed the same restrictions on q. The properties of N^* are derived in the course of the main argument.

We begin with the I-agents. There are three types of strategies among these agents, namely s^*, $\hat{s}$, and s'. The expected per period payoffs for these are given by:

$$E\pi(s^* \mid \sigma) = q \cdot \tfrac{6\lambda^*+6\hat{\lambda}+\lambda'}{\lambda^*+\hat{\lambda}+\lambda'} + (1-q) \cdot (6\lambda^* + 6\hat{\lambda} + 6\tilde{\lambda} + \lambda' + \overline{\lambda}) - \xi$$

$$E\pi(\hat{s} \mid \sigma) = q \cdot \tfrac{6\lambda^*+6\hat{\lambda}+3\lambda'}{\lambda^*+\hat{\lambda}+\lambda'} + (1-q) \cdot (6\lambda^* + 6\hat{\lambda} + 6\tilde{\lambda} + 3\lambda' + 3\overline{\lambda})$$

$$E\pi(s' \mid \sigma) = q \cdot \tfrac{5\lambda^*+7\hat{\lambda}+4\lambda'}{\lambda^*+\hat{\lambda}+\lambda'} + (1-q) \cdot (5\lambda^* + 7\hat{\lambda} + 7\tilde{\lambda} + 4\lambda' + 4\overline{\lambda})$$

What we have to show is that for N sufficiently large it holds that $E\pi^\tau(s^* \mid \sigma) > E\pi^\tau(s \mid \sigma) \cdot P^\tau(s \mid \sigma)$, for $s \in \{\hat{s}, s'\}$. As at least $E\pi^\tau(s^* \mid \sigma) > E\pi^\tau(\hat{s} \mid \sigma)$, we need to consider the relegation probabilities. For all repetitions τ, the probability that the offspring of an agent playing one of the above strategies remains among the I-agents can be estimated as follows:

$$P^\tau(s^* \mid \ \sigma) = 1, \quad P^\tau(\hat{s} \mid \ \sigma) < (1-\alpha_\tau)^N, \quad \textit{and } P^\tau(s' \mid \ \sigma) < (1-\alpha_\tau)^{2N},$$

where $\alpha_\tau = \hat{\alpha} \cdot \frac{\lambda^*_\tau}{\lambda^*_\tau+\hat{\lambda}_\tau+\lambda'_\tau}$ with λ_τ denoting the fraction of agents playing the respective strategy in repetition τ of the interaction. Thus, the condition

$E\pi^\tau(s^* \mid \sigma) > E\pi^\tau(s \mid \sigma) \cdot P^\tau(s \mid \ \sigma)$, is satisfied if for all τ

$$A := \frac{q \cdot \frac{6\lambda^*+6\hat{\lambda}+\lambda'}{\lambda^*+\hat{\lambda}+\lambda'} + (1-q) \cdot (6\lambda^* + 6\hat{\lambda} + 6\tilde{\lambda} + \lambda' + \overline{\lambda}) - \xi}{q \cdot \frac{6\lambda^*+6\hat{\lambda}+3\lambda'}{\lambda^*+\hat{\lambda}+\lambda'} + (1-q) \cdot (6\lambda^* + 6\hat{\lambda} + 6\tilde{\lambda} + 3\lambda' + 3\overline{\lambda})} > (1-\alpha)^N;$$

or equivalently

$$\frac{ln(A)}{ln(1-\alpha)} < N.$$

If N is sufficiently large such that s^* always does better than $\hat{s}$ among the I-agents, it follows that $\frac{\lambda^*_\tau}{\lambda^*_\tau+\hat{\lambda}_\tau+\lambda'_\tau} \uparrow_\tau$, $\frac{\hat{\lambda}^*_\tau}{\lambda^*_\tau+\hat{\lambda}_\tau+\lambda'_\tau} \downarrow_\tau$ and $\frac{\lambda'_\tau}{\lambda^*_\tau+\hat{\lambda}_\tau+\lambda'_\tau} \downarrow_\tau$. Hence, as $\xi < 1$, it holds that

$$1 > A > \frac{q \cdot \frac{5\lambda^*_0+5\hat{\lambda}_0}{\lambda^*_0+\hat{\lambda}_0+\lambda'_0}}{q \cdot \frac{5\lambda^*_0+5\hat{\lambda}_0+2\lambda'_0}{\lambda^*_0+\hat{\lambda}_0+\lambda'_0} + (1-q) \cdot 3} =: B,$$

and that for all τ

$$1 - \alpha_\tau \leq 1 - \alpha_0 = 1 - \hat{\alpha} \cdot \frac{\lambda^*_0}{\lambda^*_0 + \hat{\lambda}_0 + \lambda'_0} < 1.$$

Thus, the first condition on N^* which is independent of τ is given by:

$$N^* > \frac{ln(B)}{\ln(1-\alpha_0)}.$$

Similarly, the requirement that $E\pi^\tau(s^* \mid \sigma) > E\pi^\tau(s' \mid \sigma) \cdot P^\tau(s' \mid \ \sigma)$ is satisfied if for all τ:

$$C := \frac{q \cdot \frac{6\lambda^*+6\hat{\lambda}+\lambda'}{\lambda^*+\hat{\lambda}+\lambda'} + (1-q) \cdot (6\lambda^* + 6\hat{\lambda} + 6\tilde{\lambda} + \lambda' + \overline{\lambda}) - \xi}{q \cdot \frac{5\lambda^*+7\hat{\lambda}+4\lambda'}{\lambda^*+\hat{\lambda}+\lambda'} + (1-q) \cdot (5\lambda^* + 7\hat{\lambda} + 7\tilde{\lambda} + 4\lambda' + 4\overline{\lambda})} > (1-\alpha)^{2N};$$

or equivalently

$$N > \frac{ln(C)}{2\ln(1-\alpha)}.$$

Accordingly, we obtain as a second condition for N^* which again is independent of τ:

$$N^* > \frac{ln(D)}{2\ln(1-\alpha_0)} \ ,$$

with

$$D := \frac{q \cdot \frac{6\lambda_0^* + 6\hat{\lambda}_0 + \lambda'}{\lambda_0^* + \hat{\lambda_0} + \lambda_0'}}{q \cdot \frac{5\lambda_0^* + 7\hat{\lambda}_0 + 4\lambda_0'}{\lambda_0^* + \hat{\lambda}_0 + \lambda_0'} + (1-q) \cdot 4} \ .$$

Hence, if

$$N^* > max\left\{\frac{ln(B)}{\ln(1-\alpha_0)}, \frac{ln(D)}{2\ln(1-\alpha_0)}\right\},$$

then s^* is σ-prevailing among the I-agents for any undominated full support profile σ. Moreover, N^* is a function of $q, \hat{\alpha}$ and the initial distribution of strategies given by σ. The signs of the derivatives of N^* given in the proposition follow immediately from the above conditions.

We now turn to the $\overline{I}$-agents. For these, we have to consider $\tilde{s}$ and $\overline{s}$, the expected per period payoff of which is given by:

$$E\pi(\tilde{s} \mid \sigma) = q \cdot \tfrac{6\tilde{\lambda} + 3\overline{\lambda}}{\tilde{\lambda} + \overline{\lambda}} + (1-q) \cdot (6\lambda^* + 6\hat{\lambda} + 6\tilde{\lambda} + 3\lambda' + 3\overline{\lambda})$$

$$E\pi(\overline{s} \mid \sigma) = q \cdot \tfrac{7\tilde{\lambda} + 4\overline{\lambda}}{\tilde{\lambda} + \overline{\lambda}} + (1-q) \cdot (5\lambda^* + 7\hat{\lambda} + 7\tilde{\lambda} + 4\lambda' + 4\overline{\lambda})$$

We show that $E\pi(\tilde{s} \mid \sigma) < E\pi(\overline{s} \mid \sigma)$ for all repetitions. As all offspring of the $\overline{I}$-agents will again be part of the $\overline{I}$-agents, this is sufficient to prove that $\overline{s}$ is σ prevailing for the $\overline{I}$-agents. Now $E\pi(\tilde{s} \mid \sigma) < E\pi(\overline{s} \mid \sigma)$ can be rewritten as

$$q \cdot \frac{\tilde{\lambda} + \overline{\lambda}}{\tilde{\lambda} + \overline{\lambda}} + (1-q) \cdot (-\lambda^* + \hat{\lambda} + \tilde{\lambda} + \lambda' + \overline{\lambda}) > 0.$$

Using that $\lambda^* = 1 - \hat{\lambda} - \tilde{\lambda} - \lambda' - \overline{\lambda}$, this in turn can be simplified to

$$q > \frac{\lambda^* - 0.5}{\lambda^*} = 1 - \frac{1}{2\lambda^*} \ ,$$

which is always satisfied as we required $q > \frac{3+\xi}{5}$ in the proposition. Thus, by the last step of the proof of Proposition 2, it follows that s^* is σGPS for any undominated full support profiles σ (given the requirements of the proposition). q.e.d.

Derivation of Implication 1.1

We prove the statement under the simplifying assumption that all players can actually observe all other decision made the whole pool of players in any round. Let c_n denote the cost of norm disobedience in period n of the interaction By assumption $c_1 = c(., 1, G)$ is equal for all agents and unaffected by any negative past experience. Thus, an agent who maximises his utility will choose s_i such as to equate (if possible):

$$\frac{-\pi'(s_i)}{c_1'(s_i)} = \frac{\rho_i}{(1-\rho_i)};$$

otherwise, boundary solutions obtain.

As we assumed ρ to be distributed according to some continuous full support distribution F, it follows that $h(2) < 1 = h(1)$ for all agents as at least some players will not fully cooperate in the first round but can neither be punished of that nor be avoided in the later rounds. From this it follows that $c_2(s_i) < c_1(s_i)$, for all s_i, as we assumed $\frac{\partial^2 c}{\partial h \partial s_i} > 0$. Due to the distributional assumption on ρ, the process unravels so that also $c_3 < c_2$ and so forth, until only $h(t) \approx 0$.

Obviously, if the number of agents that can be observed is small, the process will be slower or may even fail to start. Averaging over many few-agent-interactions, though, it will still be visible (due to the distributional assumptions made). Notice, however, that if c_1 is already reduced before the start of the interaction, either for all agents or only for those who then choose to defect in period 1, the process of unravelling may fail to start. q.e.d.

References

Akerlof, G., and W. Dickens, 1982, "The Economic Consequences of Cognitive Dissonance," American Economic Review 72, pp. 307-319.

Akerlof, G., and R. Kranton, 2000, "Economics and Identity," Quarterly Journal of Economics 115, pp. 715-753.

Akerlof, G., and R. Kranton, 2005, "Identity and the Economics of Organizations," Journal of Economic Perspectives 16, pp. 9-32.

Andreoni, J., 1990, "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving," The Economic Journal 100, pp. 464-477.

Benabou, R., and J. Tirole, 2006a, "Incentives and Prosocial Behavior," American Economic Review 96, pp. 1652-1678.

Benabou, R., and J. Tirole, 2006b, "A Cognitive Theory of Identity, Dignity and Taboos," Princeton University, mimeo, October.

Bester, H., and W. Güth, 1998, "Is Altruism Evolutionary Stable," Journal of Economic Behavior and Organisation 34, pp. 193-209.

Bergstrom, T., 2002, "Evolution of Social Behavior: Individual and Group Selection," Journal of Economic Perspectives 16, pp. 67-88.

Bergstrom, T., 2003, "The Algebra of Assortative Encounters and the Evolution of Cooperation," International Game Theory Review 5, pp. 211-228.

Blount, S., 1995, "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences," Organizational Behavior and Human Decision Processes 63, pp. 131-144.

Brown, M., A. Falk, and E. Fehr, 2004, "Relational Contracts and the Nature of Market Interactions," Econometrica 72, pp. 747-780.

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

Charness, G., and M. Rabin, 2002, "Understanding Social Preferences With Simple Tests," Quarterly Journal of Economics 117, pp. 817-869.

Duffy, J., and J. Ochs, 2003, "Cooperative Behavior and the Frequency of Social Interaction," mimeo.

Dawkins, R., 1976, *The Selfish Gene*, Oxford University Press, Oxford, new edition 1989.

Dufwenberg, M., and M. Kirchsteiger, 2004, "A Theory of Sequential Reciprocity," Games and Economic Behavior 47, pp. 268-298.

Falk, A., E. Fehr, and U. Fischbacher, 2005, "Driving Forces Behinde Informal Sanctions," Econometrica 73, Notes and Comments, pp. 2017-2030.

Falk, A., and U. Fischbacher, 2006, "A Theory of Reciprocity," Games and Economic Behaviour 54, pp. 293-315.

Fehr, E., and S. Gächter, 2000, "Cooperation and Punishment in Public Goods Experiments," American Economic Review 90, pp. 980-994.

Fehr, E., G. Kirchsteiger, and A. Riedel, 1998, "Gift Exchange and Reciprocity in Competitive Experimental Markets," European Economic Review 42, pp.1-34.

Fehr, E., and K. Schmidt, 1999, "A Theory of Fairness, Competition, and Cooperation," Quarterly Journal of Economics 114, pp. 817-868.

Festinger, L., 1957, *A Theory of Cognitive Dissonance*, Evanston, IL: Row Peterson.

Fudenberg, D., and D. Levine, 1998, *The Theory of Learning in Games*, MIT Press, Cambridge, Massachusetts.

Gächter, S., and A. Falk, 2002, "Reputation and Reciprocity: Consequences for the Labor Relation," Scandinavian Journal of Economics 104, pp. 1-26.

Geanakoplos, J., D. Pearce, and E. Stacchetti, 1989, "Psychological Games and Sequential Rationality," Games and Economic Behaviour 1, pp. 60-79.

Guala, F., 2005, *The Methodology of Experimental Economics*, Cambridge University Press, Cambridge.

Harmon-Jones, E., and J. Mills (eds.) 1999, *Cognitive Dissonance - Progress on a Pivotal Theory in Social Psychology*, American Psychological Society, Washington, DC.

Henrich, J., 2004, “Cultural Group Selection, Coevolutionary Processes and Large-Scale Cooperation,” Journal of Economic Behavior and Organization, pp. 3-35.

Hoffman, E., K. McCabe, K. Shachat, and V. Smith, 1994, “Preferences, Property Rights and Anonymity in Bargaining Games,” Games and Economic Behavior 7, pp. 346-380.

North, D., 1990, *Institutions, Institutional Change and Economic Performance*, Cambridge University Press, Cambridge, UK.

North, D., 1993, “Economic Performance Through Time,” in T. Persson (ed.), *Nobel Prize Lectures, Economics 1991-1995*, World Scientific Publishing Co., Singapore, 1997.

Nowak, M., and K. Sigmund, 1998, “The Dynamics of Indirect Reciprocity,” Journal of Theoretical Biology 194, pp. 561-574.

Nowak, M., and K. Sigmund, 2005, “Evolution of Indirect Reciprocity,” Nature 437, pp. 1291-1298.

Rabin, M., 1993, “Incorporating Fairness into Game Theory and Economics,” American Economic Review 83, pp. 1281-1302.

Rubinstein, A. 1998, *Modeling Bounded Rationality*, MIT Press, Cambridge, Massachusetts.

von Neumann, J. and O. Morgenstern, 1944, *Theory of Games and Economic Behavior*, 1st edition, Princeton University Press, 2nd edition: 1947.

Weibull, J., 1995, *Evolutionary Game Theory*, MIT Press, Cambridge, Massachusetts.

Wichardt, P., 2005a, "Identity and Why We Cooperate With Those We Do," http://ssrn.com/abstract=748004.

Wichardt, P., 2005b, "Norms, Cognitive Dissonance, and Cooperative Behaviour - A Comment on Laboratory Experiments in Economics," http://ssrn.com/abstract=782244.