On Why and How Identity Should Influence Utility

Philipp C. Wichardt[†]

Department of Economics, University of Bonn

This Version: January 25, 2006

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 evolutionary advantageous. Taking our argument one step further, we also discuss how such a preference should be accounted for in the modelling of utility and show how this squares with the evidence.

Key words: Cognitive Dissonance, Fairness, Identity, Inequity Aversion, Reciprocity, Social Norms, Utility
JEL code: A13, C70, C90, D01, Z13

---

[*]Acknowledgements: I am grateful Georg Nöldeke for enlightening discussions about the modelling of evolutionary processes. Also, I want to thank Georg Nöldeke, Patrick Schmitz and Wiebke Wichardt for helpful comments and suggestions. A scholarship of the German Research Foundation (DFG) is gratefully acknowledged. The usual disclaimer applies.

[†]Postal address for correspondence: Bonn Graduate School of Economics, Adenauerallee 24-42, D-53113 Bonn, Germany; e-mail: p.c.w@web.de.

It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard for their own interest. We address ourselves not to their humanity, but their self-love, and never talk to them of our necessities, but of their advantage. (Adam Smith, 1776, pp. 23-24)

1 Introduction

For a long time, most applied studies of game theory have relied solely on (an affine transformation of) the agents' material rewards as a proxy for the utilities associated with the respective outcomes; a choice which is very natural 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 strongly indicates that one of the possible causes for this discrepancy can be found in the above mentioned restriction to 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, 2000; 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). All these approaches, which will be discussed in a later section, have greatly 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?

The aim of the present paper is to outline an answer to the question how immaterial concerns in the agent's utility function can be rationalised in the context of social evolution and to discuss its implications for the modelling of utility. Within a simple model of social interaction, it is argued that the development and survival of cooperative social norms which detract the agents' focus from pure material self-interest (e.g. fairness norms) can be explained *if* the agents possess a preference for identity-consistent behaviour that impacts on economic decision making (cf. Akerlof and Kranton, 2000). Due to the apparent advantage of such norms for the economic development of a society (cf. North, 1990, 1993), we infer from our argument that a preference for identity consistent behaviour is evolutionary advantageous. More specifically, the argument in favour of a general preference for identity consistent behaviour derives from a comparison of the long run development of the agents average earnings in a society under different starting conditions regarding the existence of such a preference.

Based on our 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 bad conscience). Thus, the relative strength of the immaterial aspects in the agent's utility will depend strongly on the agent's perception of the respective decision.

The distinctive feature of our approach is that it imputes neither general other-regarding preferences (e.g. fairness concerns) to the agent, nor a general preference for reciprocal behaviour. Instead it is argued that, whenever norms prescribing the respective behaviour are salient, the agent will tend to act *as if* he had the corresponding preferences in order to optimally balance identity and material concerns. But he will do so for purely selfish reasons and in a way that crucially depends on the context.

The difference in motivation, though, is essential. The point is that 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 that we are able to make reliable (though rough) predictions about how this will affect the agents' decision making. The current approach, which argues the case for the relevance of identity concerns and social norms, is an attempt to contribute to the discussion of these issues.

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 and discuss some general implications regarding

observable behaviour as well as experimental evidence. Section 4 puts our discussion into the context of the existing literature. Section 5 concludes.

2 The Basic Model

Within a simple model of social interaction, we study the evolution of behaviour in a society under different assumptions about the initial preferences of the agents. For our analysis we consider two cases for the agents' preferences: 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 also use the findings of our analysis in order to argue in favour of the development of a widespread preference for identity consistent behaviour itself. We begin our discussion with a description of the model of the social interaction.

We assume that the society consists of 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 number of offspring of an agent is determined according to the relative material payoffs earned by the agent during his "life."

The per period interaction can be described as follows. In each period, the agents are matched in pairs to play the one-shot Prisoner's Dilemma (PD) depicted in Figure 1 where the specified individual payoffs refer to material rewards. Moreover, after having played the PD game but before being rematched, players have the chance to enforce some (costly) punishment on their opponent.[1] More specifically, both players, knowing the outcome of

[1]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

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.[2]

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

Figure 1: The Standard Prisoner's Dilemma

Apart from the interaction, we assume that each agent of the society, prior to any matching, has to decide whether to identify (I) with a cooperative norm or not ($\overline{I}$). If the agent chooses to identify with the norm and if the agent has an id-preference, then any behaviour which is incompatible with the norm will result in a mental cost c which will be deducted from the agent's per period payoffs. If the agent does not have an id-preference, per period payoffs for this agent remain unaffected by the identity decision. The behaviour prescribed by the norm is the following:

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.[3]

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.

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

[3]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

The payoff consequences of the first to aspects of the norm are immediate. As regards the vigilance decision, we assume that being vigilant (v) is associated with a per period material cost of ξ, $0 < \xi << 1$, whereas choosing not to be vigilant ($\overline{v}$) is costless.

Moreover, 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). This 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. The purpose of the segregation is to allow norm obedient agents to avoid those who do not behave according to the norm to a certain extent, once these have been recognised by someone.

However, different from the identity decision, which we assume 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 is assumed to depend on the overall vigilance of the I-agents. More specifically, if an I-agent in a certain period behaves in a way that is incompatible with the norm but observable to others, i.e. if he defects, refrains from punishing a defector or from being vigilant, he will be found out with a probability $\alpha := \overline{\alpha} \cdot \nu$, where $\overline{\alpha} \in (0, 1)$ 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 fact that they still will claim to identify with the norm).[4] However, no agent is informed about whether he has been convicted

unattended luggage to a member of staff. Also, people who witnessed a crime, i.e. a break of a social norm cast in law, do have to give evidence in court and perjury commonly is itself liable to prosecution.

[4]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, in fact, be construed as lack of concern for the norm itself. This in turn may well compromise the respective person's reliability. Moreover,

before the end of period N.[5]

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 $1-q$, however, the agent is (randomly) matched in an environment to which all agents of the society have access (cf. Figure 2). Agents do not know where they are matched but do know q.

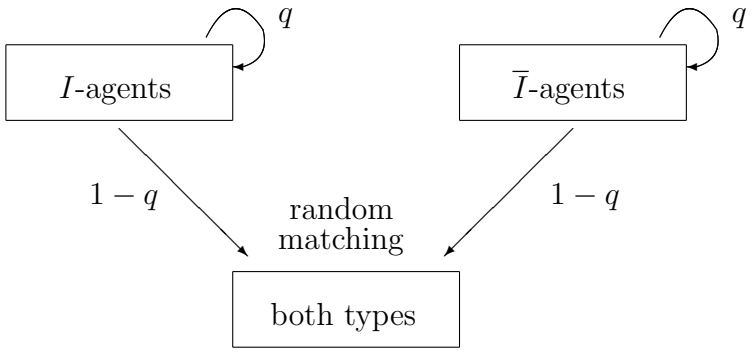

Figure 2: The matching of agents. $0 \leq q \leq 1$.

Summing up, the overall process can be described by the following six step procedure:

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

---

the parameter $\overline{\alpha}$, 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 $\overline{\alpha}$ will not affect our argument.

[5]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.

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

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 strategy of his parent.

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

In the sequel, we study the long run evolution of society according to the above described procedure for different starting conditions. More specifically, for different degrees of dissemination of the id-preference in society, we first let agents choose strategies for the initial strategic form game G which corresponds to Steps 1 to 4 from above, and then ask how behaviour and the dissemination of the id-preference in society will develop over time, given the evolutionary dynamic indicated in Steps 5 and 6.

Recall that we assumed all agents, having made their identity choice, to decide once and for all how to behave in the N periods of the interaction (Step 2). Hence, for an agent i, a strategy for the game G is a tuple $s_i = (s_{1i}, s_{2i})$, where s_{1i} specifies the agent's identity choice, i.e.

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

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

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. Furthermore, the *starting strategy profile* that obtains for the society once all agents have made their decisions 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), $\tau = 0, 1, ...$, we denote the expected per period payoff of a strategy $s \in S$, given σ, by $E\pi^\tau(s \mid \sigma)$; the index τ is omitted if there is no hazard of confusion.

As mentioned earlier, immediately after the strategy choices for G the society is split up into two subsocieties, the I- and the $\overline{I}$-agents. Hence, the reference to the subsocieties is made according to the agents first round identity choices. However, after any repetition τ of the interaction the offspring of those agents who only declare to identify with the norm but do not actually comply with it may be relegated to the $\overline{I}$-agents, although they still play $s_1 = I$. In other words, 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). Nonetheless, we stick to this terminology also for later repetitions as each of the two subsocieties is unambiguously referred to through the reference to the identity choice of those agents it originated from.

For the purposes of our argument, we study the evolution of behaviour in each part of society and compare the long run average expected payoff in

both parts relative to each other in order to assess the long run evolution of behaviour in the whole society. Accordingly, 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.[7] Formally, we define:

Definition 1 *Let σ be a starting strategy profile of G 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 $\mu > 0$.[8]

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,*

[7] 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.

[8] Requiring a constant minimum distance in the payoffs is not a necessary but sufficient condition which ensures the absence of difficulties in the limit. We choose it as it is most convenient for the later argument.

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.*

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

Definition 3 *Let* σ *be a starting strategy profile of* G*. A strategy* s^* *is called* globally σ-prevailing strategy *or* σ*GPS 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. Accordingly, we restrict attention to the evolution of behaviour, given the system started with no agent playing a strictly dominated strategy. 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.

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* G *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 with an id-preference; and

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

as well as $\tilde{s}$ *and* $\overline{s}$ *specified above for the agents without an id-preference. The relative frequencies of the above defined strategies in the society are denoted as the corresponding strategy, i.e. we write* $\lambda^*,\ \tilde{\lambda},\ \overline{\lambda},\ \hat{\lambda},$ *and* λ'.

To begin with, let us 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. We state it 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. The formal 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* σ*GPS, 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. Nevertheless, 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 following proposition tells us (roughly) that the agents with an id-preference playing s^* nevertheless will prevail if the different subsocieties are sufficiently separated and if 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[9] 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. The formal proof of Proposition 3 can be found in the appendix.

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 of G, if $q > \frac{3+\xi}{5}$ and if N is large, i.e. if $N > N^*(q,\overline{\alpha},\sigma)$, where N^* is decreasing in $\overline{\alpha}$, and increasing in λ_0^* and q.[10] 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 $\overline{\pi} \approx 6-\xi$.*

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

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

[10]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 $\overline{\alpha} \approx 0.1$ values about $N^* = 25$, which we consider to be very few interactions per lifetime, are sufficient.

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, as we have argued, norm obedient behaviour, in expectation, eventually will prevail (under certain conditions on the interaction) and the agents with a preference for identity consistent behaviour will dominate the society in the long run.

Notice that, neither infrequent random matching with agents from other societies nor infrequent migration of agents would affect the preceding argument as we considered only full support starting profiles anyway. 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, only furthers the eventual dissemination of the id-preference. Hence, we conclude:

Conclusion From an evolutionary perspective which also takes into account the competition of different social groups for scarce resources in the above described way, a (widespread) individual preference for norm- or identity-consistent behaviour is advantageous as it enables the development of a cooperative trait in society. No assumptions about general individual concerns for the utility of other agents are necessary.

3 On How to Account for Identity in Utility

Having outlined the general advantage of an individual preference for identity-consistent behaviour, we now want to consider more closely, albeit less formally, how it satisfactorily can be accounted for in the agent's utility considerations if the social environment allows for more complex patterns in the interaction (3.1). Building on these general considerations, we finally leave the realm of stylised models of social interaction completely and discuss the

general implications of our considerations regarding observable behavioural patterns as well as related experimental 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 cost c. This was sufficient for our argument as we presumed the individual matching process to be random. However, if we allow for patterns in the matching process, for example because the agents of a society occasionally encounter long episodes of interaction with members of different societies (e.g. as salesmen or envoys), 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 to that agent as he may be exploited and, thus, be disadvantaged in his evolutionary fitness.[11] Thus, increasing correlation in the matching process, e.g. through increasing levels of inter-group interaction which certainly accompanied the development of human societies through the centuries, 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 inconsistent behaviour $c(.)$ to depend on the context (do the 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

[11] 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 individually detrimental.

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[12]) 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.[13] 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.[14] Again, in terms of our model, the agents of a society may, for example, agree to relegate only the offspring of repeatedly disobedient agents. In that case, if the readjustment to an again cooperative environment is sufficiently fast, a more flexible cost from norm disobedient behaviour appears to preferable.

In the following, we propose a stylised model of utility which still accounts for the main parts of our prior argument and nonetheless offers the desired flexibility.

Let us assume that the utility relevant for decision making 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 player's behaviour with the internalised norms,[15] that relative weights are fixed (idiosyncratic) constituents

[12]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.

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

[14]Similarly, if the agent once was overwhelmed by the economic prospect of a defection, this may increase the cost of later inconsistencies, e.g. through a guilty conscience, and may reinforce the norm once subscribed to. See also the discussion of cognitive dissonance in Section 4.3.

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

of the players' identity, and that the cost from inconsistent behaviour is omnipresent and equal for all agents.[16] The clear distinction between identity dependent relative weights and a generally valid cost from norm disobedience is made for expositional purposes only. Yet, it is very intuitive as it allows us to think of $c(.)$ as some commonly agreed upon standard to which agents can subscribe to 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.[17]

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, 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 is evocative of (in our case) cooperative social norms.[18] Putting things together,

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.

[16]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.

[17]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.

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

we obtain a utility function of the following form:

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

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.[19]

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, note that 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 as each agent, having met the defector, afterwards would follow the purely economic self-interest, i.e. would defect himself.

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.

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 do not act as uncontingent cooperators. Instead, as we have argued,

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

they adjust behaviour according to their past experience and, more generally, to how evocative the respective context is of cooperative social norms. The remainder of this section aims to make the behavioural consequences of our argument more visible by discussing some of its general implications in the context of 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. $s_i \in [0, 1]$.[20] 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 and that past experience h_i also is measured on a zero-one scale, i.e. $h_i \in [0, 1]$, where larger values of h_i indicate a more cooperative past

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

experience. More specifically, we assume that $h_i(1) = 1$ if the agent has no prior experience with the respective context but is otherwise determined by the average cooperativeness of all actions (but the agent's own one) observed in the previous round. Then, for given G, the psychologic cost $c(s_i, h_i \mid G)$ can be written as a function

$$c(.|G) : [0,1] \times [0,1] \rightarrow \subset \mathbb{R}_- \ .$$

Assuming $c(.)$ to be twice differentiable, our previous arguments can be translated into the following requirements:

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

i.e. the cost is increasingly negative if past experience or own actual behaviour are less cooperative, and the less cooperative past experience is the weaker is the latter effect.

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 sice of the material rewards from non-cooperative behaviour.*

Intuitively, Implication 1.1 follows from the fact that we assumed the distribution of ρ_i in the population to have full support. 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" is inappropriate as it cannot be enforced. Hence, in the next round 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

In fact, experimental evidence from the lab, 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 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).[21] 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

[21] That defectors make use of the punishment option if it is cheap but effective, i.e. if the resulting cost to the punished agent are higher than those that accrue to the punisher, may be due to increased relative fitness which can be achieved that way.

cooperation rates.[22] 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). On the other hand, increased anonymity in dictator games is found to decrease the amount of money left (Hoffman et al., 1994). And, in fact, tennis professionals in Wimbledon are found to almost perfectly match game theoretic predictions with their variation in services (Walker and Wooders, 2001). All this clearly indicates the importance of the context for the relative influence of social norms on decision making.

Discussion

We want to emphasise at this point that, if social norms do influence behaviour the way outlined above, both the rational paradigm and the more recent models of fairness and reciprocity (cf. Section 4.1) can be accommodated within our approach. According to our argument, the pure rational agents model will prove most valuable in neutral, purely competitive circumstances (e.g. auctions).[23] However, as soon as social norms (or stereotypes) become more pronounced, we have to account for the additional incentives related to identity (cf. Akerlof and Kranton, 2000). As regards cooperative social norms, the models of fairness and reciprocity, which are briefly discussed in the next section, appear to be an intuitive and tractable extension of the rational paradigm, although they leave the realm of pure self-interest (which is not done in the present treatise). Yet, once the importance of cooperative social norms for a certain context is undoubted, the application of these models on an *as if* basis seems justified. The present discussion, however, may help to clarify which model is most appropriate under which circumstances or whether even new models accounting for different social norms are necessary.

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

[23] Complexity considerations regarding decision making are deliberately neglected at this point.

4 Related Literature

Finally, we want to put our analysis into the context of the existing literature. In particular, we want to discuss how it relates to other approaches to 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), and how it is connected to the research on cognitive dissonance in psychology (4.3).[24]

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, 2000; 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, though, are different and can be roughly 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 which a general preference of the agent for the

[24] 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).

equal (or fair) split is assumed (which has to be balanced with pure self-interest). Thus, in this type of model, the agents are assumed to have a kind of uncontingent preference to give to or cooperate with their opponents.

The models in the second category (to which all other papers cited above belong) essentially try to circumvent the assumption of uncontingent goodwill by assuming the agent's utility as derived from the other player's payoff to be aligned with the intentions the agent believes the other player to have towards him.[25] 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 derive a benefit from being mean to B. Thus, also these models allow for a very flexible adjustment of behaviour to the context or past experience; it just has to be accounted for in the beliefs about intentions. Consequently, predictions may be very much in line with our approach.

Yet, the approaches differ in more than intuition. The point is that the models of fairness and reciprocity discussed above aim to capture observed patterns in behaviour without discussing in greater detail their origins or general value for the individual or society. The current approach, however, focuses on the question why and how such patterns may have developed and what this should imply for their particularities. As a result of this, the present framework allows us to ex ante assess the importance of reciprocity or fairness (or other non-material) concerns for a certain type of interaction by explicitly asking about the significance of these (and other) social conventions in the respective context.

In fact, in our view, the degree to which an agent acts *as if* he cares about the utility of his opponent only reflects the agent's assessment of the salience of cooperative norms in the respective context (as this will influence

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

the cost of disobedience). And the agent's seeming concerns about intentions of his opponent, in turn, are part of this assessment (as is the agent's past experience which may serve as a proxy of intentions). Yet, it is only through the reference to the underlying social norms and the agent's (selfish) general desire to act in accordance with his identity that we are able to ex ante assess the direction and strength of these considerations. Thus, in a sense, our approach might serve as a tool to calibrate the other models to a particular context emphasising, for example, the type of intentions most likely to dominate. In spirit, though, it remains quite different from these models but close to the discussion of economics and identity.

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 may arise if the individual distaste for identity threatening acts, both by him-/herself and by others, is taken into account. For example, they emphasise the importance of gender associations with certain jobs for the supply of labour (e.g. if a certain type of work is considered a man's job, women will be inefficiently hesitant to apply for such a job).

In a later paper, Akerlof and Kranton (2004) extend their discussion and emphasise the positive effects of a strong association with a group on the degree of cooperative behaviour with/within that group.[26] In essence, they claim that if an agent (e.g. a worker) sufficiently internalises the objective function of the group (e.g. a firm), the resulting (internal) motivation will suffice to sustain a high effort level, and conclude that there is a strong incen-

[26]The discussion of identity is also taken up, for example, by Benabou and Tirole (2004). In their 2004 paper, they demonstrate nicely how an individual concern for prosocial behaviour, which may be interpreted in terms of identity, will affect contributions to social goods in a very intuitive way.

tive for any organisation to attempt to achieve a high degree of identification with the organisation among its members.[27] Concerning utility they write, very much in the spirit of our analysis, 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, 2004, p. 4).

Very much in support of the ideas expressed Akerlof and Kranton, this paper adds an (indirect) evolutionary argument for the emergence of a preference for identity consistent behaviour to this discussion, and discusses in more detail and from a somewhat different perspective how exactly identity concerns will influence behaviour and how they can be accounted for in utility.

4.3 Cognitive Dissonance

Cognitive Dissonance is a psychologic phenomenon the discussion of which can be traced back to Festinger (1957).[28] 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.

Clearly, the concept of cognitive dissonance is closely related to the Akerlof and Kranton discussion about economics and identity as well as to our

[27] 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.

[28] See also Aronson (1966), Aronson et. al (1999), Nail et al. (2004), Steele (1988), and Steele et al. (1993). See Akerlof and Dickens (1982), for a discussion of the economic relevance of this phenomenon.

analysis. In particular, the apparent evidence that inconsistent behaviour indeed causes mental distress very much supports the assumption that this is taken into account in the agent's economic decision making. And, of course, it is very reassuring that psychologists are able to identify today what we just claimed to be evolutionary advantageous.

Moreover, psychologists, to some extent, also have addressed the effects of inconsistent behaviour over time (though we are not aware of any evolutionary approach to this issue). For example, psychologists have studied the question what happens if additional incentives to overcome the dissonance are strong enough[29] or if the dissonance arousing behaviour simply cannot be avoided[30]. The rough answer to this is that, in order to alleviate the (mental) consequences from his behaviour, the individual will attempt to either create an internal justification such as an adjustment of subjective beliefs aimed to resolve the dissonance or reinforce the positive self-image through pronounced consistent behaviour elsewhere.[31] Apparently, both reaction tie in well with the argument in favour of a history dependent cost from norm-disobedience (cf. Setion 3); the individual may learn the inappropriateness of the norm in a certain context (belief adjustment) or may use later cooperation, initiated by increased costs from defection, to redeem the rule breaking (reinforcement)[32].

5 Concluding Remarks

In this paper, we have outlined an argument in favour of an individual preference for identity-consistent behaviour, where identity refers to the social

[29]Strong economic incentives are but one possibility; cf. Akerlof and Dickens (1982) or Aronson (1966).

[30]See, e.g. Brehm, 1956.

[31]In fact, there seems to be some disagreement in psychology as to which effect dominates (or is the right one); see Nail et al. (2004). See also Steele (1988).

[32]Cf. Footnote 14.

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, e.g. decreasing cooperation rates over time if punishment is unavailable, indeed, are very much in line with the evidence.

In view of the current increase in models of utility highlighting immaterial aspects of economic decision making, we have tried to emphasise that we do not view our approach as a substitute to any of these but rather as a complement indicating why (e.g.) fairness considerations may be of great importance in some circumstances but not in others. We are convinced that there is a whole bunch of features beyond material rewards which influence individual decision making and which give rise to many of the phenomena that are so puzzling against the backdrop of the standard rational agent model. However, psychologic incentives are much more fickle and more difficult to grasp than material ones which (obviously) either are present or not. Consequently, prior assessments as to the relative strength of these effects are usually connected with considerable uncertainty. To (slightly) reduce this uncertainty through a discussion of the underlying psychological mechanisms at work as well as their general benefit for the respective agent, and thus to contribute to a better understanding of idiosyncracies of individual decision making and their dependence on the context is what this paper has aimed to do.

Appendix

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.

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 put 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 $\alpha_\tau = \overline{\alpha} \cdot \frac{\lambda^*_\tau}{\lambda^*_\tau+\hat{\lambda}_\tau+\lambda'_\tau}$ 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

$$1-\alpha_\tau \leq 1-\alpha_0 = 1-\overline{\alpha}\cdot\frac{\lambda_0^*}{\lambda_0^*+\hat{\lambda}_0+\lambda_0'} < 1.$$

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 $q, \overline{\alpha}$ 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, we 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 65, pp. 715-753.

Akerlof, G. and R. Kranton, 2004, "Identity and the Economics of Organizations," forthcoming, Journal of Economic Perspectives.

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

Aronson, E., 1966, "The Psychology of Insufficient Justification: An Analysis of some Conflicting Data," pp. 109-133 in S. Feldman (ed.), *Cognitive Consistency*, Academic Press, New York.

Aronson, J., G. Cohen and P. Nail, 1999, "Self-Affirmation Theory: An Update and Appraisal," pp. 127-147 in E. Harmon-Jones and J. Mills (eds.), *Cognitive Dissonance: Progress on a Pivotal Theory in Social Psychology*, American Psychological Association, Washington DC.

Benabou, R., and J. Tirole, 2004, "Incentives and Prosocial Behavior," mimeo.

Brehm, J., 1956, "Postdecision Changes in the Desirability of Alternatives," Journal of Abnormal and Social Psychology 52, pp.384-389.

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.

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, 2000, "A Theory of Reciprocity," IEW working paper no. 6, University of Zürich.

Fehr, E., and S. G"achter, 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.

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.

Nail, P., J. Misak and R. Davis, 2004, "Self-Affirmation versus Self-Consistency: A Comparison of two Competing Self-Theories of Dissonance Phenomena," Personality and Individual Differences 36, pp. 1893-1905.

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.

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.

Smith, A. 1776, *The Wealth of Nations*, edited by E. Cannan (1904), reprinted 2002, Bantam Dell, New York.

Steele, C., 1988, "The Psychology of Self-Affirmation: Sustaining the Integrity of the Self," Advances in Experimental Social Psychology 21, pp. 261-302.

Steele, C., S. Spencer and M. Lynch, 1993, "Self-Image Resilience and Dissonance: The Role of Affirmational Resources," Journal of Personality and Social Psychology 64, pp. 885-896.

Walker, M. and J. Wooders, 2001, "Minmax Play at Wimbledon," American Economic Review 91, pp. 1521-1538.

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.