Social Macroeconomics
Working Paper Series

Social norms and the evolution of conditional cooperation

SM-WP-2007-001
March 2007

Mathias Spichtig
Christian Traxler

Social Norms and the Evolution of Conditional Cooperation*

Mathias Spichtig† and Christian Traxler‡

March 15, 2007

Abstract

This paper develops a model of social norms and cooperation in large societies. Within this framework we use an indirect evolutionary approach to study the endogenous formation of preferences and the coevolution of norm compliance. Thereby we link the multiplicity of equilibria, which emerges in the presence of social norms, to the evolutionary analysis: Individuals face situations where many others cooperate as well as situations where a majority free-rides. The evolutionary adaptation to such heterogenous environments will favor conditional cooperators, who condition their pro-social behavior on the others' cooperation. As conditional cooperators react flexibly to their social environment, they dominate free-riders as well as unconditional cooperators.

JEL classification: C70; Z13

Keywords: Conditional Cooperation; Indirect Evolution; Social Norms; Heterogenous Environments.

*We would like to thank Florian Herold, Friederike Mengel, Klaus Schmidt, Aljaž Ule and seminar participants at CREED, University of Amsterdam, for helpful comments. Christian Traxler acknowledges financial support by the German Research Foundation (DFG, Project SPP 1142), the European Network for the Advancement of Behavioural Economics (ENABLE) as well as the kind hospitality of CREED, University of Amsterdam, where most of the work to this project was done. All errors and mistakes remain our own.

†Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam. Kruislaan 320, 1098 SM Amsterdam, The Netherlands. E-mail: spichtig@science.uva.nl

‡*Corresponding Author.* Seminar for Economic Policy, University of Munich. Akademiestrasse 1/II, 80799 Munich, Germany. Phone/Fax: +49 89 2180 -3303/-6296; E-mail: christian.traxler@lrz.uni-muenchen.de

1 Introduction

Starting with Keser and van Winden (2000) and Fischbacher et al. (2001), economic research has pointed out the role of *conditional cooperation*. Agents who follow this behavioral pattern condition their cooperation on the cooperativeness of others respectively on their beliefs about others' behavior – they *"are willing to contribute the more to a public good, the more others contribute"* (Fischbacher et al., 2001, p.397). There is now a solid body of empirical evidence from lab experiments as well as from field studies, which documents the prevalence of conditional cooperation.[1] Motivated by this evidence, several theoretical models have emerged, which are capable to explain conditional cooperative behavior. The question under which circumstances conditional cooperation – respectively the preferences inducing it – survives evolution, has gained little attention. The main concern of this paper is to address this question.

One possible way to capture conditional cooperation is based upon *social norms*.[2] Social norms are rules of conduct, which are enforced by internal or external sanctions (Coleman, 1990). As the sanctions for a norm deviation are harsher the more people adhere to the norm, a social norm for cooperation can trigger conditional cooperative behavior (Rege, 2004). The present analysis incorporates such a concept of social norms into a model of voluntary public good provision in a large society. Within this framework we study the evolution of a cooperation norm and the coevolution of behavior. This allows us to discuss the prerequisites for the emergence of conditional cooperation. Our analysis thereby provides several novel elements.

First of all, the strength of the social norm, respectively the impact of norm-enforcing sanctions, depends – next to the level of norm compliance in the society – on an individual specific level of norm sensitivity: Some agents suffer more from sanctions than others do. For a given distribution of norm sensitivity in the population, we can then endogenously derive the equilibrium level of cooperation. Similar as in other models of social norms (e.g. Lindbeck et al., 1999), there is scope for a multiplicity of equilibria: Society could either coordinate on equilibrium states with a strong social norm and far-reaching cooperation or on states with weak norm-enforcement and widespread free-riding.

In a next step, we study the evolutionary process of norm adaptation. So far, the literature has mainly focused on actual behavior as the determinant of an endogenous norm strength (see e.g. Azar, 2004; Rege, 2004). In addition to this channel, we also consider the individual norm sensitivity as an endogenous factor accounting for the power of a norm. We model the evolution of the norm sensitivity as an indirect evolutionary process.[3] Individuals learn about the social status of agents with heterogenous preferences, i.e. different levels of norm sensitivity. Status is determined by the economic payoff from free-riding and cooperation as well as from the norm-

[1] For recent field experiments compare e.g. Frey and Meier (2004), Martin and Randal (2005), Shang and Croson (2005). Gächter (2006) provides a survey of further evidence.

[2] Other theoretical approaches which account for conditional cooperation are theories of fairness, conformity, inequity aversion and reciprocity, surveyed in Fehr and Schmidt (2006).

[3] Compare e.g. Güth (1995), Bester and Güth (1998), Fershtman and Weiss (1998) for further indirect evolutionary studies. A comprehensive overview of standard evolutionary game theory is provided in Weibull (1995).

enforcing sanctions. Depending on whether these sanctions are strong enough to outbalance the cost of cooperation, either the pro-social or the selfish behavior dominates in terms of social status. Accordingly, either agents with higher norm sensitivities (who tend to cooperate) or agents with lower norm sensitivities (who will free-ride) get more frequently imitated. In this vein, adaptation endogenously forms the distribution of the norm sensitivity in the society. Individual behavior, the level of cooperation within the population and the associated strength of sanctions evolves indirectly, along with the endogenous change in preferences. In an evolutionary equilibrium, the social outcome is shaped by preferences and – at the same time – the social outcome shapes these preferences.

We first discuss the evolutionary process for the case where agents adapt to a *homogenous environment* associated with one particular equilibrium state of the public good game. Under certain conditions, there exits an evolutionary equilibrium with a distribution of norm-sensitivities such that free-riders and cooperators coexist. This equilibrium, however, turns out to be instable. Typically, adaptation will induce a decline in the norm sensitivity and cooperation would break down. In the evolutionary equilibrium the social norm has eroded and nobody contributes to the public good.

This result changes, once we incorporate the multiplicity of equilibria from our basic model into the analysis. We focus on the case of a *heterogenous environment*, in the sense that the population faces an equilibrium state with strong norm-compliance as well as a state with widespread norm violations, where both states are supported by one given distribution of preferences. Agents then interact in 'cooperative' and 'non-cooperative' situations, with a strong status-impact of sanctions in the former and a weak norm in the latter environment. In this setup, we observe three different types of behavior: Free-riders, who violate against the norm in both situations, (unconditional) cooperators, who always comply with the social norm, and finally conditional cooperators. These agents cooperate in the 'good' state, where many others follow the norm, but defect in the 'bad' state, where a majority free-rides. In the environment with a strong social norm, conditional cooperators avoid harsh sanctions, making them more successful than free-riders. In the environment where the norm is weak they free-ride and earn a higher status payoff than unconditional cooperators. Hence, the conditional strategy dominate both unconditional strategies in terms of social status. Evolutionary adaptation will favor conditional cooperators, since they react flexibly to their social environment. We can characterize conditions, under which this dominance of conditional cooperation forms a stable evolutionary equilibrium.

While there are several indirect evolutionary approaches explaining the emergence of pro-social behavior (e.g. Bester and Güth 1998; Fershtman and Weiss 1998), only Mengel (2006) discusses conditional cooperation. Her paper studies the impact of migration on an internalized norm for cooperation. For some degrees of population viscosity – which can be neatly linked to the level of integration in a society – she finds a stable evolutionary equilibrium, where norm-sensitive and norm-insensitive agents coexist. Similar as in our study, norm-sensitive individuals behave conditionally cooperative: they start to defect, if norm-insensitive agents become more

frequent in the population. This protects conditional cooperators from getting exploited and supports their evolutionary success. The result as well as it's intuition is quite similar to our findings in the case of heterogenous environments. In Mengel's analysis, conditional cooperation is a response to the heterogeneity in selfish respectively norm-guided interaction partners. In our model, it is the heterogeneity in social environments related to different equilibrium states, which supports the conditional behavior. This structural similarity in the results suggests, that the role of heterogenous environments as a driving force in the evolution of conditional cooperation provides a robust finding which generalizes to different model frameworks.

Finally, our paper also contributes to the literature by introducing a technique from quantitative genetics, which – to the best of the author' knowledge – is novel in evolutionary economics. The method, originally developed in Lande (1976), provides a simple tool to analyze the evolution of a *continuously* distributed trait – in our case, the norm sensitivity. We discuss the crucial assumptions of Lande's approach and show that our main findings are qualitatively robust to the application of standard replicator dynamics (see e.g. Weibull, 1995). The fact that we study the evolution of a continuous distribution of preferences instead of a discrete number of types, also distinguishes our model from Mengel (2006) and other contributions in the field.

The remaining paper is structured as follows. We first study a model of social norms and cooperation in a large population. In section 3 we introduce an evolutionary approach from quantitative genetics. We then apply this method on our model and discuss the evolution of social norms and cooperative behavior in a homogenous respectively in a heterogeneous environment. Section 5 provides a critical discussion of our findings and section 6 concludes.

2 Social Norms and Cooperation

Consider a large society represented by a continuum of individuals $[0, 1]$. Each agents i chooses $x^i \in \{0, 1\}$, to contribute to the public good ($x^i = 1$, 'cooperate') or not to contribute ($x^i = 0$, 'free-ride'). The payoff $y(x^i)$ for strategy x^i is given by

$$y(x^i) = -x^i c \tag{1}$$

where $c > 0$ depicts the costs of the public good contribution. The action x^i also determines a payoff $z(x^i, n)$, where n depicts the share of free-riders in the society. This payoff is defined as

$$z(x^i, n) = \left(x^i - 1\right) s(n) \tag{2}$$

where $s(n)$ relates to the sanctions an agent incurs if she violates against the social norm for cooperation. The origin of these sanctions could in principle be internal, external or a mixture of both (Coleman, 1990). In the context of internalized social norms, emotions represent an internal sanctioning mechanism.[4] If an agent has internalized a cooperation norm, free-riding would be

[4] A review on emotions in economic theory is provided by Elster (1998).

associated with emotions like guilt, remorse or the loss of self-esteem. External sanctions could be monetary or non-monetary, e.g. related to social disapproval.[5] This paper does not study the origin of these sanctions – i.e. why people engage in (costly) norm-enforcement activities. We simply assume that there exists a mechanism which induces a certain punishment of free-riders.

Throughout our analysis we employ the following assumption:

Assumption A1: The finite-valued function $s(n)$ is continuously differentiable in n. For $n \in [0,1]$ there holds $s'(n) \leq 0$. Moreover $s(0) > 0$ and $s(n) \to 0$ for $n \to 1$.

Allowing the sanctions to depend on other agents' behavior captures the idea that the degree of norm compliance (co)determines the strength of norm-enforcement and thereby the strength of the social norm. Following the literature (e.g. Lindbeck et al. 1999, Mengel 2006), we assume $s(n)$ to be non-increasing in n. A deviant agent is supposed to suffer from weaker internal sanctions, as free-riding becomes more widespread: one feels less guilty about violating a norm, the more others do the same. The equivalent is supposed to hold for external sanctions.[6] For the case of perfect norm compliance ($n = 0$), sanctions are strictly positive. In a society where everybody free-rides, however, the cooperation norm has eroded. The norm-based moral connotation of 'wrong' (free-riding) and 'right' (contributing) have vanished and sanctions are infinitesimal.

2.1 Preferences

Let the preferences of agent i, defined over $y(.)$, $z(.)$ and the public good payoff $v(g)$, be given by an additive separable utility function

$$u^i(x^i, n) = y(x^i) + \theta^i z(x^i, n) + v(g(n)), \tag{3}$$

with the individual specific parameter $\theta^i \in [-\infty, \infty]$. The public good is defined by $g = g(n)$, $g' < 0$, and $v' > 0$. We can interpret the parameter θ^i as the degree of norm sensitivity. While an agent with $\theta^i = 0$ is solely concerned about the material payoff from the game, those with $\theta^i > 0$ also consider the norm-based payoff in their decisions.[7]

In a large population, a single decision maker takes n as well as g as given. Hence, agent i will cooperate iff $u^i(1, n) > u^i(0, n)$, which holds for $\theta^i s(n) > c$. An individual contributes to the public good, if the utility loss from the sanction dominates the costs of cooperation. This implies the threshold

$$\hat{\theta}(n) \equiv \frac{c}{s(n)}, \tag{4}$$

which divides society into norm-adhering and norm-breaking individuals. Those with $\theta^i >$

[5] For evidence on the role of non-monetary sanctions compare e.g. Masclet et al. (2003). For a theoretical analysis of social sanctions compare e.g. Fershtman and Weiss (1998).

[6] Falk et al. (2005) and Masclet et al. (2003) discuss experimental evidence which supports this assumption.

[7] Agents with $\theta^i < 0$ hold anti-social preferences, as they derive benefits from a norm-violation. As will become clear in the following, we only include this latter group for technical convenience. Excluding negative values of θ would not change any of our results.

$\hat{\theta}(n)$ cooperate, while those with $\theta^i \leq \hat{\theta}(n)$ free-ride. The action x^i is then determined by an individual's norm sensitivity θ^i and the share of free-riders n,

$$x^i = x(\theta^i, n) = \begin{cases} 0 & \text{for } \theta^i \leq \hat{\theta}(n) \\ 1 & \text{for } \theta^i > \hat{\theta}(n) \end{cases} \tag{5}$$

Note that the threshold $\hat{\theta}(n)$ is non-decreasing in n,

$$\frac{\partial \hat{\theta}(n)}{\partial n} \geq 0, \tag{6}$$

since $s'(.) \leq 0$. As more agents deviate from the norm, the sanctions associated with a norm violation become smaller. Hence, an agent who cooperates for low levels of n may turn into a free-rider for higher levels of n. Those individuals with $\theta^i \in (\hat{\theta}(0), \hat{\theta}(1))$ condition their cooperation on the behavior of others. They act as *conditional cooperators*. Agents with $\theta^i \leq \hat{\theta}(0)$, however, would always free-ride, irrespectively of other subjects behavior. Allowing for a heterogeneity in θ, the model therefore captures the two main patterns of behavior typically found in experimental studies (e.g. Fischbacher et al. 2001).

2.2 Equilibrium

Let the cumulative distribution function of the parameter θ be given by $\Phi(\theta)$. We assume that $\Phi(\theta)$ is continuously differentiable on the interval $[-\infty, \infty]$. The corresponding density function $\phi(\theta)$ has full support, this is $\phi(\theta) > 0$ for $\theta \in (-\infty, \infty)$.

Assumption A2: (i) The inverse function of the cumulative distribution is given by $\Phi^{-1}(n)$ for $n \in [0, 1]$, with $\Phi^{-1}(0) = -\infty$ and $\Phi^{-1}(1) = \infty$. (ii) $\exists\ n \in (0, 1) : \Phi^{-1}(n) > \hat{\theta}(n)$.

A social equilibrium state in such a society is given by a share of free-riders n^*, characterized by the fixed point equation

$$n^* = \Phi(\hat{\theta}(n^*)). \tag{7}$$

Lemma 1 *For any $s(n)$ and $\Phi(\theta)$ as characterized in A1 and A2(i) there always exists an equilibrium with $n^* = 1$. If A2(ii) holds, there always exists at least one further equilibrium with $0 < n^* < 1$.*

Proof. We can rewrite condition (7) as $\Phi^{-1}(n^*) = \hat{\theta}(n^*)$. From A2(i) we know that $\Phi^{-1}(1) = \infty$ and from A1 follows $\hat{\theta}(n) \to \infty$ for $n \to 1$. Hence, there always exits an equilibrium with $n^* = 1$. From A1 we know $s(0) > 0 \Rightarrow \hat{\theta}(0) > 0$ which implies $\hat{\theta}(0) > \Phi^{-1}(0)$. From this follows that there must exist at least one $n^* \in (0, 1)$ where $\Phi^{-1}(n^*) = \hat{\theta}(n^*)$ holds as long as A2(ii) is fulfilled, since both $\hat{\theta}(n)$ and $\Phi^{-1}(n)$ are continuously increasing functions defined over the unit interval. ■

An equilibrium constitutes a self-supporting share of norm-violators: The threshold $\hat{\theta}(n^*)$ is such that the share of agents with $\theta^i \leq \hat{\theta}(n^*)$ is exactly n^*. There always exists one equilibrium

where nobody contributes, $n^* = 1$. The cooperation norm has eroded and everybody free-rides. Given that assumption A2(ii) holds, the strength of the norm sensitivity is distributed such that there exists a level of free-riding n, where the maximum level of norm sensitivity among free-riders, $\Phi^{-1}(n)$, is above the cooperation threshold $\hat{\theta}(n)$. In this case, the system is characterized by a multiplicity of equilibria. In addition to the equilibrium with $n^* = 1$, there is at least one equilibrium with a positive share of contributors. A graphical representation of two possible scenarios is provided in figure 1. While assumption A2(ii) is fulfilled for the example depicted in panel (a) of the figure, it does not hold for the example in panel (b). In the first case, there are multiple equilibria, in the latter there is a unique equilibrium at $n^* = 1$.

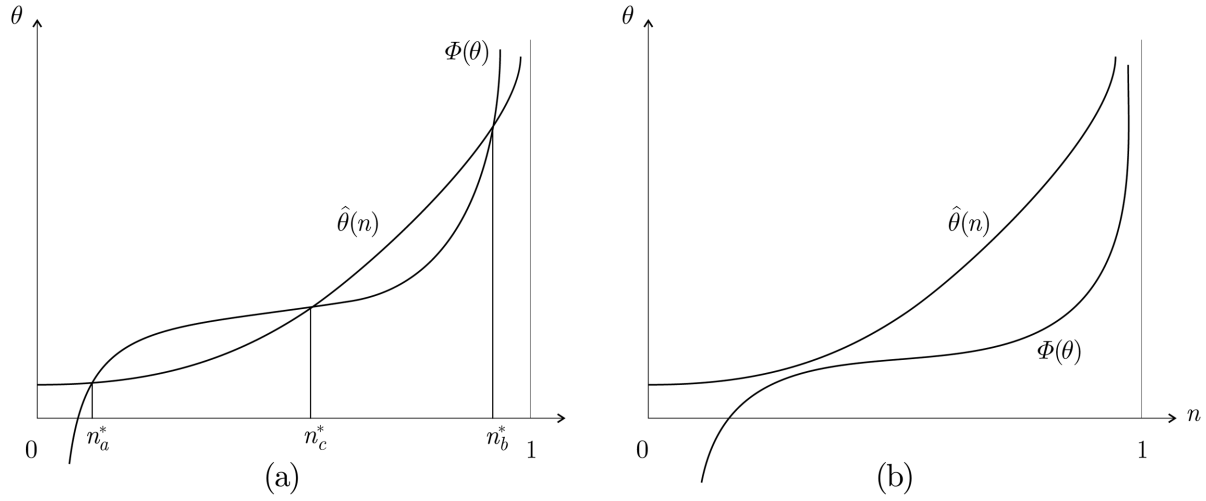

Figure 1: Equilibrium Share of Free-Riders

If the distribution $\Phi(\theta)$ is common knowledge, society immediately coordinates into one of the possible equilibria. Alternatively one could consider $\Phi(\theta)$ to be unknown, but assume that agents can induce the behavior of other members in society from the public good level. Agents could then learn about the share of free-riders. As long as players base their next periods' decision on this share – i.e. cooperate in the next period if θ^i is above the current period's threshold $\hat{\theta}(n)$ and free-ride otherwise – society would converge into an asymptotically stable equilibrium, characterized by

$$\frac{\partial \Phi^{-1}(n^*)}{\partial n} \geq \frac{\partial \hat{\theta}(n^*)}{\partial n}. \tag{8}$$

In the following we call an equilibrium n^* a *stable equilibrium state*, if (8) holds for n^*. In the scenario depicted in panel (a) in figure 1, there are two instable (the one with n_c^* and another one at $n^* = 1$) and two stable equilibrium states: one with a low level of free-riding n_a^* and another one where free-riding is widespread, n_b^*. In panel (b) the only equilibrium, $n^* = 1$, is also stable, since the cumulative distribution approaches the $\hat{\theta}(n)$-curve 'from below' (and therefore condition (8) holds).

3 Evolutionary Quantitative Genetics

In the following we will study the evolution of the distribution $\Phi(\theta)$. For this purpose, we introduce a technique from evolutionary quantitative genetics, first analyzed by Lande (1976).[8] The approach offers a tractable method to study an evolutionary process within a continuously heterogenous population. In particular, it will provide us with one easy to interpret parameter – the mean value of θ – which characterizes the distribution $\Phi(\theta)$ in an evolutionary equilibrium. In section 5 we will discuss the applicability of this technique to our problem as well as the differences to standard replicator dynamics (see e.g. Weibull, 1995).

Consider a large population which is heterogeneous along one trait α. The trait value is normally distributed with mean $\bar{\alpha}$ and variance σ^2, $\alpha \sim N(\bar{\alpha}, \sigma^2) \equiv F(\alpha, \bar{\alpha}, \sigma^2)$. To simplify notation, we denote the distribution function by $F(\alpha)$ and the corresponding density function by $f(\alpha)$. Let the fitness of an α-type, i.e. an individual with a trait value α, for a given distribution with mean $\bar{\alpha}$ be given by $w(\alpha, \bar{\alpha})$. Allowing individual fitness to depend on the distribution accounts for *frequency dependent* fitness. Fitness is called frequency dependent, if the fitness of an α-individual does also depend on the composition of the population.[9] In economic terms, frequency dependence is given if one group of agents – respectively the strategy played by these individuals – creates an externality on other agents' fitness.[10]

Within one generation, the change in the mean trait value in response to selection is defined as

$$\Delta\bar{\alpha} = \bar{\alpha}_s - \bar{\alpha}, \tag{9}$$

where $\bar{\alpha}_s$, the mean trait value after selection, is given by

$$\bar{\alpha}_s = \frac{1}{\bar{w}} \int \alpha w(\alpha, \bar{\alpha})\, dF(\alpha) \tag{10}$$

and $\bar{w}$, the mean fitness of the population, is

$$\bar{w} = \int w(\alpha, \bar{\alpha})\, dF(\alpha). \tag{11}$$

The selection described in (10) follows a standard replicator dynamic. While the initial frequency of a type was $f(\alpha)$, the post-selection frequency of this type, $\frac{w(\alpha,\bar{\alpha})}{\bar{w}} f(\alpha)$, will be higher for types with above-average fitness. Hence, in the computation of $\bar{\alpha}_s$, more successful types will get more weight than less successful types.

The analysis so far describes selection within one generation. In order to address the inter-generational evolution of the trait α, Lande (1976) introduces the following structure of re-

[8] Compare Falconer and Mackay (1995) and Roff (1997) for an introduction to quantitative genetics.

[9] As we will consider the variance to be fixed, we have suppressed this variable in $w(.)$ to ease notation.

[10] Consider for example the decision to commit a crime where the likelihood of a criminal act to be 'successful' depends on the crime rate in the society. (E.g. the detection probability might be lower, the more other agents become criminals.) If decisions depend on individual risk-preferences, the distribution of these preferences clearly influences the success of a criminal.

production: First, only selected individuals produce the next generation of offspring. Second, partner selection and genetic recombination transforms the post-selection distribution into an offspring distribution which is again normal: it is characterized by the initial variance σ^2 but a different mean.[11] According to this structure, selection will then first lead to a distribution which deviates from the initial one. Starting from a norm distribution with mean $\bar{\alpha}$, the mean of the (non-normal) distribution after selection is given by $\bar{\alpha}_s$ from (10). After mating and reproduction, however, the distribution of α in the new generation is again normal with $F(\alpha, \bar{\alpha}_s, \sigma^2)$. While the variance is preserved, the mean of the distribution changes from $\bar{\alpha}$ to $\bar{\alpha}_s$. The direction of evolution is therefore determined by selection, characterized in (9) and (10). This allows us to analyze the evolutionary process in more detail.

From (11) we can derive the change in mean fitness from a marginal change in $\bar{\alpha}$,

$$\frac{\partial \bar{w}}{\partial \bar{\alpha}} = \int w(\alpha, \bar{\alpha}) \, \frac{\partial f(\alpha)}{\partial \bar{\alpha}} \, d\alpha + \int \frac{\partial w(\alpha, \bar{\alpha})}{\partial \bar{\alpha}} \, dF(\alpha). \tag{12}$$

While the first term characterizes the direct change in the mean fitness due to a change in the composition of the population, the second term depicts the indirect, frequency dependent fitness impact. From the density of the normal distribution we can easily compute $\partial f(\alpha)/\partial \bar{\alpha}$. Substituting in (12) and rearranging yields

$$\Delta \bar{\alpha} = \frac{1}{\bar{w}} \int w(\alpha, \bar{\alpha}) \, (\alpha - \bar{\alpha}) \; dF(\alpha). \tag{13}$$

(For the derivation of (13) see Appendix A.) The right hand side in equation (13) characterizes pace and direction of the evolutionary process. As $\bar{w} > 0$ (per assumption), the direction of the evolutionary change in the mean trait value $\bar{\alpha}$ is determined by the sign of the integral in (13). Note that the integral term represents only the direct change in mean fitness (the first term in equation (12)). From (13) therefore follows that the evolution of $\bar{\alpha}$ is independent of the frequency dependent fitness change associated with a change in $\bar{\alpha}$. If the direct fitness impact is positive (negative), the distribution will evolve towards a higher (lower) mean $\bar{\alpha}$. An *evolutionary equilibrium* is reached if $\Delta \bar{\alpha} = 0$. Such an equilibrium is characterized by

$$\int w(\alpha, \bar{\alpha}^e) \, (\alpha - \bar{\alpha}^e) \; dF(\alpha) = 0, \tag{14}$$

where $\bar{\alpha}^e$ denotes the mean trait value in equilibrium.

[11] The assumptions underlying this structure are justified by the observation that most metric traits have a normal distribution, or that the distribution can be transformed to normal by a change in the scale of measurement (e.g. by log transformation). Similar arguments are incurred to account for the independence of the variance in respect to the mean, and for that the variance is assumed constant over evolutionary time. For a detailed discussion see Lande (1976). Compare also Falconer and Mackay (1995), Roff (1997).

4 Indirect Evolution of Conditional Cooperation

We now apply the method introduced in the previous section in order to study the evolution of the distribution $\Phi(\theta)$ and the associated coevolution of cooperation in the model from section 2. As we do not believe that the norm sensitivity θ is genetically determined, we interpret evolution as a cultural process, related to social transmission and learning mechanisms. Fitness describes the success of a certain θ-type, i.e. an individual with norm sensitivity θ, in terms of social status. In the course of evolution, individuals learn about the social status of different θ-types and accordingly adapt their θ values. In this way, the adaptation process endogenously shapes preferences. Individual behavior and thereby the level of cooperation within society evolves indirectly with the change in preferences from one generation to the next.[12] The term generation thereby describes a population with a given distribution of preferences $\Phi(\theta)$, rather than a parent and offspring-population in the biological sense.

We deviate from the typical approach in evolutionary economics, which only considers the economic payoffs as determinant of evolutionary fitness (see e.g. Fershtman and Weiss, 1998; Mengel, 2006). Apart from the economic payoff $y(x^i)$, fitness is also determined by the norm-based sanctions imposed on free-riders, $z(x^i, n)$. If, for example, norm-violators get stigmatized and are excluded from some social interactions, this results in a decrease of social status.[13] The fitness impact of norm-enforcing sanctions, is thereby assumed to be non-increasing in the share of norm-violators n. It is less 'costly' (in terms of fitness) to free-ride in a population where norm violations are widespread and social sanctions are less severe.[14]

Let the fitness for an action x^i be given by

$$w(x^i) = y(x^i) + z(x^i, n) + \upsilon. \tag{15}$$

The payoff from the public good and some constant, exogenous fitness component is subsumed in υ. In the following we will neglect υ in our analysis, since including this additional payoff would not alter our results. Note, however, that we implicitly assume υ to be sufficiently large to guarantee $\bar{w} > 0$.

The basic structure of the adaptation process is the following: An initial generation with a given distribution $\Phi(\theta)$ faces the public good game described in section 2. After (finitely) many repetitions of the game, agents have learned about the social status of different θ-types and adapt their own θ^i. The resulting change in the $\Phi(\theta)$ is assumed to be characterized by the process from (13). In section 5 we discuss the crucial differences of this approach from quantitative genetics to an adaptation process according to standard replicator dynamics.

[12] For other indirect evolutionary approaches, compare e.g. Güth (1995), Bester and Güth (1998), Fershtman and Weiss (1998).

[13] Compare e.g. Riedl and Ule (2002) for experimental evidence on social exclusion.

[14] Note that ostracism, e.g. in form of exclusion from the public good consumption, follows a similar pattern. Loosing the benefits from the public good in a society with a high level of cooperation represents a more severe punishment than exclusion in a society white less cooperation. Compare Hirshleifer and Rasmusen (1989).

We will now study this structure for two scenarios. First, we consider the case, where each generation coordinates (always) on one social equilibrium state n^*. Then we turn to the case, where – in the context of multiple equilibria – one generation will face different equilibrium states. We will call the first scenario a *homogenous* and the latter a *heterogenous environment.*

4.1 Adaptation to a Homogenous Environment

Let θ be normally distributed according to $\theta \sim \phi(\theta, \bar{\theta}, \sigma^2)$, and the cumulative distribution is given by $\Phi(\theta, \bar{\theta}, \sigma^2)$. Substituting for $y(x^i)$, $z(x^i, n)$ and $x^i = x(\theta^i, n)$ from (1), (2) and (5), we can express individual fitness as

$$w(\theta, \bar{\theta}) = \begin{cases} -c & \text{for } \theta > \hat{\theta}(n^*) \\ -s(n^*) & \text{for } \theta \leq \hat{\theta}(n^*) \end{cases} \tag{16}$$

where $n^* = \Phi(\hat{\theta}(n^*), \bar{\theta}, \sigma^2)$ is a stable equilibrium, analogous to (7), for a normal distribution with mean $\bar{\theta}$ and σ^2 is exogenously given.

It is important to note three points here. First, it is only the heterogeneity in actions – determined by different levels of θ – which results in fitness differences. Within the group of cooperators respectively free-riders, the heterogeneity in θ does not result in different levels of fitness. Second, individual fitness as described by (16) is frequency dependent. As the distribution of θ changes, the share of free-riders n^* and thereby the fitness costs of a norm deviation will change. Remember, that the method introduced in section 3 accounts for such spillovers. Third, we assume that a generation always coordinates on one equilibrium state n^*. In this sense, we study the adaptation to a homogenous environment. After the adaptation process, the next generation (with a new distribution of θ) is assumed to coordinate on an equilibrium state in the close neighborhood of the previous one – even if there exist different possible equilibrium states.[15]

The mean fitness is defined by $\bar{w} = \int w(\theta, \bar{\theta})\, \phi(\theta)$. Using (16), we can express $\bar{w}$ as

$$\bar{w} = -c + (c - s(n^*)) \int_{-\infty}^{\hat{\theta}(n^*)} d\Phi(\theta) \tag{17}$$

with the integral expression being equal to $n^* = \Phi(\hat{\theta}(n^*), \bar{\theta}, \sigma^2)$. Following (13), the intergenerational change in $\bar{\theta}$ is determined by

$$\Delta\bar{\theta} = \frac{1}{\bar{w}} \left(s(n^*) - c\right) \left(\bar{\theta} n^* - \bar{\theta}^*\right) \tag{18}$$

[15]This assumption on equilibrium selection can be justified by the fact that after a small change in the distribution (i.e. in $\bar{\theta}$) there always exists a new stable equilibrium state in the close neighborhood of the previous one. This 'close by' equilibrium may be more salient than more distant equilibrium states and hence becomes a focal point equilibrium.

(compare Appendix A) where $\bar{\theta}^*$ represents the mean level of θ among the n^* agents who free-ride in an equilibrium,

$$\bar{\theta}^* \equiv \int_{-\infty}^{\hat{\theta}(n^*)} \theta d\Phi(\theta). \tag{19}$$

As long as $0 < n^* < 1$, there holds $\bar{\theta} n^* > \bar{\theta}^*$. Remember also that $\bar{w} > 0$ per assumption (compare (15)). Hence,

$$\text{sign}\left\{\Delta\bar{\theta}\right\} = \text{sign}\left\{s(n^*) - c\right\} \quad \text{for } 0 < n^* < 1. \tag{20}$$

From (18) and (20) we can derive:

Proposition 1 ***(i)*** *An evolutionary equilibrium where cooperators and free-riders coexist is characterized by $s(n^e) = c$, where $0 < n^e = \Phi(\hat{\theta}(n^e), \bar{\theta}^e, \sigma^2) < 1$ constitutes a stable equilibrium state, supported by a normal distribution with mean $\bar{\theta}^e$.* ***(ii)*** *In such an equilibrium, $\hat{\theta}(n^e) = 1$ and all agents have the same fitness $w(\theta, \bar{\theta}^e)$.* ***(iii)*** *An evolutionary equilibrium where cooperation fails, $n^{e1} = 1$, is characterized by a stable equilibrium state $n^{e1} = \Phi(\hat{\theta}(n^{e1}), \bar{\theta}^{e1}, \sigma^2)$, supported by a normal distribution with mean $\bar{\theta}^{e1}$.*

Proof. The proof of **(i)** follows immediately from (18). From (4) we know that $c = \hat{\theta}(n^*)s(n^*)$ must hold for any equilibrium state. $s(n^e) = c$ then implies $\hat{\theta}(n^e) = 1$. Using this in (16) and substituting for (4) proofs **(ii)**. Part **(iii)** derives from $n^* = 1 \Rightarrow \bar{\theta} n^* = \bar{\theta}^*$. Hence, for $n^{e1} = 1$ the term in the last brackets in (18) is zero and $\Delta\bar{\theta} = 0$. ■

The evolutionary equilibrium described in part (i) of the proposition is characterized by a positive share of cooperators such that there is no fitness differential between free-riders and cooperators. In equilibrium, the preferences of agents with $\theta^i = \hat{\theta}(n^e)$, who are indifferent between defection and cooperation, coincide with the fitness function from (15) since $\hat{\theta}(n^e) = 1$. In other words, these θ-types are 'perfectly adapted' – the norm sensitivity in their preferences coincides with the fitness impact of sanctions. In addition, there is also an evolutionary equilibrium where everybody free-rides. While we know from Lemma 1 that $n^* = 1$ constitutes a possible equilibrium state for *any* distribution, condition (8) has to hold to guarantee the stability of the equilibrium state. Therefore, any level $\bar{\theta}$ for which (8) holds at $n^* = 1$ could be the mean of the distribution in an evolutionary equilibrium with zero cooperation. By the time the whole society free-rides, the evolutionary pressure on $\bar{\theta}$ to decline vanishes and the system reaches a rest point. Note, that we could also describe an evolutionary equilibrium with $n^e = 0$. For this case, $n^* = 0 \Rightarrow \bar{\theta} n^* = \bar{\theta}^*$. Hence, the last bracket term in (18) would equal zero and $\Delta\bar{\theta} = 0$. However, an equilibrium state with $n^* = 0$ would only be supported by a distribution with $\bar{\theta} \to \infty$. We do not include this case in our further analysis, as such a distribution would violate our assumption A2(i).

Let us now turn to the existence of these different types of equilibria.

Proposition 2 ***(i)*** *Iff $s(0) > c$, there exists an evolutionary equilibrium with $0 < n^e < 1$.* ***(ii)*** *For all distributions where (8) holds at $n^* = 1$, there exists an evolutionary equilibrium with $n^{e1} = 1$. If $c > s(0)$, this is the only equilibrium.*

Proof. **(i)** Since $c > s(n)$ for $n \to 1$ and $s(.)$ is continuously non-increasing in n, $s(0) > c$ assures that there exists a level of n where $s(n) = c$ holds. Moreover, we can always find a distribution $\phi(\theta, \bar{\theta}, \sigma^2)$, a function $s(n)$ and a level c, which supports such an equilibrium share of free-riders n^e. **(ii)** From Lemma 1 we know that $n^* = 1$ is supported by any distribution as long as A1 and A2(i) hold. Proposition 1(iii) implies that any equilibrium with $n^* = 1$ where (8) holds, constitutes an evolutionary equilibrium n^{e1}. From A1 follows $c > s(0) \Rightarrow c > s(n)$ for all $n \in [0, 1]$. It therefore follows from $c > s(0)$ that there cannot exist an equilibrium with $n^e < 1$, as $\nexists\, n$ with $s(n) = c$. ■

The result from proposition 2 is straightforward. If the fitness costs of cooperation are higher than the fitness damage of sanctions even for the state where $n^* = 0$, free-riding yields a higher social status than cooperation for any level of n. Starting from any $n^* < 1$, the adaptation process induces $\bar{\theta}$ to fall and society moves towards an equilibrium with $n^{e1} = 1$. If, however, sanctions are sufficiently strong such that cooperators get a higher fitness than free-riders for the full-cooperation state $n^* = 0$, there must exist an equilibrium state $0 < n^e < 1$ where both actions result in the same level of fitness.[16]

Finally, we address the evolutionary stability of the system. An evolutionary equilibrium is locally stable if $d\Delta\bar{\theta}/d\bar{\theta} < 0$ holds in the close neighborhood of $\bar{\theta}^e$ (respectively $\bar{\theta}^{e1}$).[17] If this is the case, small mistakes in the adaptation process would not affect the evolutionary equilibrium. Remember, that the stability of an equilibrium state within one generation is given by (8). In addition, evolutionary stability demands that also the preferences remain stable between generations. Consider for example a positive shock on $\bar{\theta}$. One can derive from (7) that an increase in the mean norm sensitivity would result in a drop in the share of free-riders below n^e. The stability condition would then demand that $\Delta\bar{\theta} < 0$, which would provide a pressure on $\bar{\theta}$ to fall and consequently on n^* to increase, thereby adapting 'back' towards the initial equilibrium $\bar{\theta}^e$ respectively n^e. In our case, however, an evolutionary equilibrium where cooperators and free-riders coexist can never be stable.

Proposition 3 *An evolutionary equilibrium with $0 < n^e < 1$ is never stable. In contrast, an evolutionary equilibrium with $n^{e1} = 1$ is locally stable.*

Proof. See Appendix B.

[16] Note, that for the distribution in this evolutionary equilibrium A2(ii) will hold, such that there exists a (stable) equilibrium state $n^* < 1$. (Compare Lemma 1.)

[17] One could also consider the stability with respect to shocks on n. Note, however, that the fitness payoff can be interpreted as the average from (finitely) many repetitions of the one-shot game from section 2 *within one generation*. As the equilibrium states n^* within an evolutionary equilibrium must be stable according to (8), we neglect deviations from n^*. Moreover, in our case $d\Delta\bar{\theta}/d\bar{\theta} \leq 0$ implies that the equilibrium would be also stable after shocks in n.

Due to assumption A1, $s'(n) \leq 0$. Hence, any small deviation from n^e would tip the balance in fitness-payoffs between the two strategies. After a positive shock on $\bar{\theta}^e$, the share of free-riders falls short of n^e and we get $s(n) \geq c$. Cooperators would be more successful than free-riders, $\bar{\theta}$ would increase and n^* would decline further. If, on the other hand, the level of free-riding exceeds n^e, the norm-based sanctions would become less effective and we get $c \geq s(n)$. Free-riders, i.e. individuals with low values of θ, have a higher fitness than cooperators; consequently $\bar{\theta}$ decreases and the system moves into an equilibrium with $n^{e1} = 1$. Note that the system would return to such an equilibrium n^{e1} after small shocks in $\bar{\theta}$, as in the neighborhood of $n^{e1} = 1$ there holds $c > s(n^{e1})$ since $s(n) \to 0$ for $n \to 1$. Hence, an evolutionary equilibrium with $\bar{\theta}^{e1}$ and n^{e1} would be stable.

The analysis provided so far yields an unsatisfactory result. While there can exist an evolutionary equilibrium where free-riders and cooperators coexist, such an equilibrium turns out to be instable. The system either evolves towards an equilibrium where the norm has eroded and everybody free-rides, or the society would evolve towards full cooperation. As we will discuss in section 5, this result also carries over if we apply standard replicator dynamics.

4.2 Adaptation to a Heterogeneous Environment

So far, we have studied the adaptation to a homogenous environment. Agents encounter one particular situation – one equilibrium state – and adaptation shapes their preferences according to the strength of the social norm in this equilibrium. In reality, however, we typically face heterogeneous environments, as social interaction can result in quite diverse outcomes. The level of cooperation varies for different collective action problems, along time and space. We now discuss a way to capture such heterogeneous environments within our framework. In contrast to the case of a homogenous environment, we find (presumably) stable evolutionary equilibria where cooperators and free-riders coexist.

Let us consider an initial distribution such that assumption A2(ii) is fulfilled. In this case, there exists a multiplicity of equilibria (compare Lemma 1). Within each generation, the population sometimes coordinates on a stable equilibrium state n_a^*, sometimes on n_b^* with $n_j^* = \Phi(\hat{\theta}(n_j^*), \bar{\theta}, \sigma^2)$ for $j \in \{a, b\}$. Without loss of generality, we assume $n_a^* < n_b^*$. The likelihood at which a generation coordinates on equilibrium state n_j^* is exogenously given by $0 < \pi_j < 1$. The actions an agent i with θ^i chooses according to (5) in the equilibrium states n_a^* respectively n_b^* is denoted by $\left(x_a^i, x_b^i\right)$. The corresponding fitness for $\left(x_a^i, x_b^i\right)$ is then given by

$$w\left(x_a^i, x_b^i\right) = \sum_{j=a,b} \pi_j \left(y(x_j^i) + z(x_j^i, n_j^*)\right). \tag{21}$$

From $n_a^* < n_b^*$ and (6) follows $\hat{\theta}(n_a^*) < \hat{\theta}(n_b^*)$. Hence, we will observe three different strategies: On the one hand, agents with $\theta^i \leq \hat{\theta}(n_a^*)$ will free-ride in both equilibrium states. Agents with $\theta^i > \hat{\theta}(n_b^*)$ on the other hand, will cooperate in both states. A third group of individuals, those with $\hat{\theta}(n_a^*) < \theta^i \leq \hat{\theta}(n_b^*)$, behaves conditionally cooperative. They cooperate in equilibrium state

a, where many others cooperate as well, but defect in state b, as more others' are free-riding. Making use of (1), (2) and (5), we can express individual fitness in the following way:

$$w(\theta,\bar{\theta}) = \begin{cases} -c & \text{for } \theta > \hat{\theta}(n_b^*) \\ -\pi_a c - \pi_b s(n_b^*) & \text{for } \hat{\theta}(n_a^*) < \theta \leq \hat{\theta}(n_b^*) \\ -\pi_a s(n_a^*) - \pi_b s(n_b^*) & \text{for } \theta \leq \hat{\theta}(n_a^*) \end{cases} \tag{22}$$

The crucial difference to the case of a homogenous environment is the fact that agents with intermediate levels of θ obtain a fitness-payoff from two different actions. The success of the conditional cooperative strategy consists of the cooperation payoff for equilibrium state a plus the payoff from free-riding in state b.

From (22) we can compute the mean fitness of the population for a given π_a and $\pi_b = 1-\pi_a$,

$$\bar{w} = -c + \pi_a \left(c - s(n_a^*)\right) \int_{-\infty}^{\hat{\theta}(n_a^*)} d\Phi(\theta) + (1-\pi_a)\left(c - s(n_b^*)\right) \int_{-\infty}^{\hat{\theta}(n_b^*)} d\Phi(\theta). \tag{23}$$

According to (13), the evolution of $\bar{\theta}$ is then determined by $\Delta\bar{\theta} = \frac{1}{\bar{w}}\Psi$ with

$$\Psi \equiv \pi_a \left(s(n_a^*) - c\right)\left(\bar{\theta} n_a^* - \bar{\theta}_a^*\right) + (1-\pi_a)\left(s(n_b^*) - c\right)\left(\bar{\theta} n_b^* - \bar{\theta}_b^*\right), \tag{24}$$

and $\bar{\theta}_j^*$ captures the mean level of θ among the free-riders for equilibrium state n_j^*, analogous to (19).[18] The evolutionary dynamics on $\bar{\theta}$ are given by

$$\text{sign}\left\{\Delta\bar{\theta}\right\} = \text{sign}\left\{\Psi\right\} \tag{25}$$

This leads to the following proposition:

Proposition 4 ***(i)*** *An evolutionary equilibrium in a heterogenous environment is characterized by $\Psi = 0$, where the stable equilibrium states $n_a^e = \Phi(\hat{\theta}(n_a^e), \bar{\theta}^e, \sigma^2)$ and $n_b^e = \Phi(\hat{\theta}(n_b^e), \bar{\theta}^e, \sigma^2)$ are supported by a normal distribution with mean $\bar{\theta}^e$.* ***(ii)*** *If $n_b^e < 1$, there holds $s(n_a^e) > c > s(n_b^e)$.*

Proof. Part **(i)** follows immediately from (25). Part **(ii)** derives from (24): Note that $\bar{\theta} n_j^* > \bar{\theta}_j^*$ as long as $n_j^* < 1$. Hence, the first term in (24) would be negative if $c > s(n_a^e)$. Since $n_a^e < n_b^e$, (6) implies that the second term would be negative as well. We would get $\Psi < 0$. Therefore $c > s(n_a^e)$ cannot hold in an equilibrium with $n_b^e < 1$. Iff $s(n_a^e) > c$, the first term in (24) is positive. In order to get $\Psi = 0$ for $n_b^e < 1$, the second term in (24) must be negative, which holds for $c > s(n_b^e)$. ■

[18] The derivation of $\Delta\bar{\theta}$ respectively Ψ is analogous to the one of (18). Compare Appendix A.

The Proposition characterizes an evolutionary equilibrium for a heterogenous environment. As long as $n_b^e < 1$, the distribution in the evolutionary equilibrium supports two equilibrium states such that $s(n_a^e) > c > s(n_b^e)$.[19] In terms of fitness, cooperation dominates free-riding in equilibrium state a. For state b, however, the opposite holds: Free-riding is more widespread, and the fitness costs from the norm-enforcing sanctions are lower than the costs of cooperation. From this follows

Corollary 1 *In an evolutionary equilibrium in a heterogeneous environment with $n_b^e < 1$ conditional cooperators have a strictly higher fitness than both, free-riders and cooperators.*

Proof. From Proposition 4(ii) we know that $s(n_a^e) > c > s(n_b^e)$. Using this in (22) proofs the Corollary. ■

Figure 2 graphically illustrates an example of such an evolutionary equilibrium. The graph on the left hand side captures a system with a distribution $\Phi(\theta)$ and a function $\hat{\theta}(n)$ supporting two stable equilibrium states $n_a^* < n_b^* < 1$. The graph on the right hand side depicts the fitness difference between the strategies for the two equilibria.

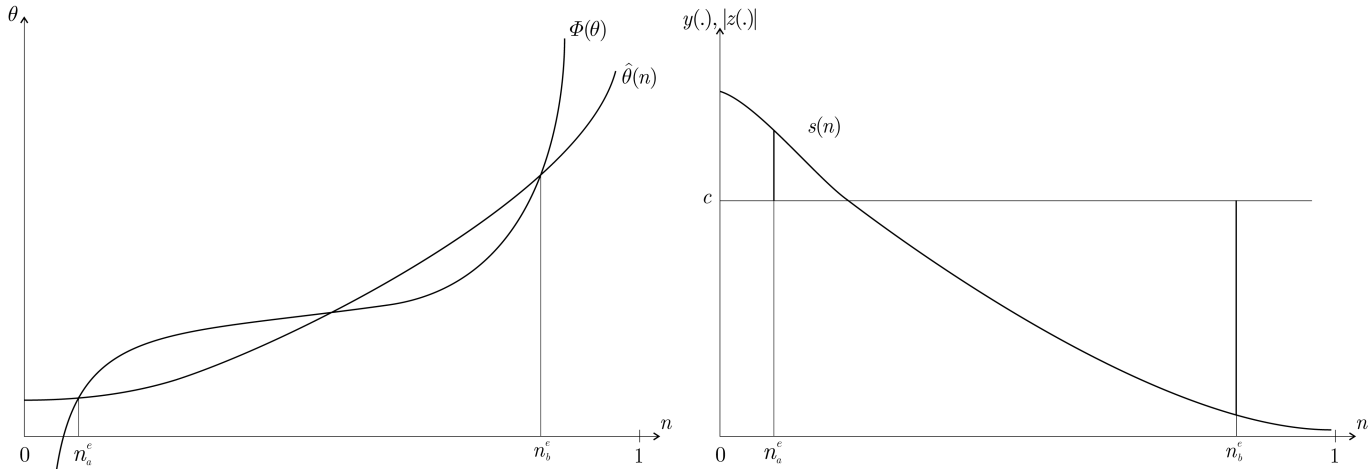

Figure 2: Evolutionary Equilibrium in a Heterogenous Environment

From figure 2 as well as from the analysis above (compare Proposition 2) it is clear that $s(0) > c$ is a necessary condition for an evolutionary equilibrium to exist. In addition, assumption A2(ii) has to hold in order to guarantee a multiplicity of equilibria. Analogous to before, the necessary conditions for the local stability of an evolutionary equilibrium is $d\Delta\bar{\theta}/d\bar{\theta} < 0$. A formal analysis yields the following result:[20]

[19] Another possible equilibrium would be $n_b^e = 1$ and $s(n_a^e) = c$. As this type of equilibrium has very similar properties as the one discussed in the previous section, we do not discuss this case. Moreover, the equilibrium condition, $\Psi = 0$, would also be fulfilled for $n_a^e = 0$ and $n_b^e = 1$ respectively $n_a^e = 0$ and $n_b^e < 1$ with $s(n_b^e) = c$. Note, however, that assumption A1 implies $\hat{\theta}(0) > 0$. Unless $\bar{\theta} \to \infty$, there is always a positive mass of individuals with $\theta \leq \hat{\theta}(0)$, which makes an equilibrium state $n_a^e = 0$ impossible.

[20] In the Appendix we discuss the conditions in the proposition and show that they can be both fulfilled.

Proposition 5 *Sufficient conditions for the stability of an evolutionary equilibrium with* $n_b^e < 1$ *are given by* $n_a^e \leq \min\{\gamma_a; \delta_a\}$ *and* $\gamma_b \leq n_b^e \leq \delta_b$*, with*

$$\gamma_j \equiv \int_{-\infty}^{\hat{\theta}(n_j^e)} \phi(\theta) \frac{\left(\theta - \bar{\theta}^e\right)^2}{\sigma^2} d\theta,$$

$$\delta_j \equiv \frac{\bar{\theta}_j^*}{\bar{\theta}^e} + \phi(\hat{\theta}(n_j^e))\ \hat{\theta}(n_j^e)\left(1 - \hat{\theta}(n_j^e)\right)\left(1 - \frac{\hat{\theta}(n_j^e)}{\bar{\theta}^e}\right).$$

Proof. See Appendix B ■

Since the stability of an evolutionary equilibrium is in general ambiguous, we conducted a series of numerical simulations. Typically, we found two levels of $\bar{\theta}$ which supported an evolutionary equilibrium.[21] The one with the higher mean norm-sensitivity was *always* stable, even for cases where the (sufficient) condition $n_a^e \leq \min\{\gamma_a; \delta_a\}$ from Proposition 5 was violated. We are therefore confident, that stable evolutionary equilibria within a heterogeneous environment exist for a wide range of parameters. This position is also backed by a straightforward intuition: Small shocks in the adaptation would not change the result from Corollary 1 – conditional cooperation would still perform more successful than the two unconditional strategies. Since conditional cooperators have intermediate values of θ, preferences in the 'middle' of the θ distribution are more successful and dominate against those with more extreme – either low or high – values of θ.

The evolutionary dominance of conditional cooperators is the main result of our analysis. Individuals who lack pro-social preferences – those with low θ values – as well as individuals with 'overly' pro-social preferences – i.e. very high values of θ – play one particular strategy, irrespectively of the other agents' behavior. In a stable evolutionary equilibrium within a homogenous environment, one of these two strategies will dominate the other. In a heterogeneous environment, however, when individuals face a 'good' state with rather high levels of cooperation as well as a 'bad' state with many free-riders, a third strategy appears: conditional cooperation. In the adaptation to such a heterogeneous environment, the two unconditional strategies prove less successful that the conditional strategy. Agents who cooperate in the good but free-ride in the bad state dominate the free-riders in the former and the cooperators in the latter environment. The evolutionary pressure to adapt to heterogenous environments provides a simple explanation for the success of conditional cooperative behavior.

[21] We focused on the functional form $s(n) = \lambda\left(1 - r\left(n^a/a - n^b/b\right)\right)$ and parameters in the range $c = 1$, $\lambda \in (1, 2]$, $r \in [1.5, 2.5]$, $a \in [1, 2]$, $b \in [2, 4]$, a standard deviation $\sigma \in [1.5, 2.5]$ and $\pi_a \in (0, 1)$. The program code is available from the authors upon request.

5 Discussion

5.1 Replicator Dynamics

Would our results still hold if evolution follows a conventional replicator dynamic rather than the quantitative genetic process? Consider any initial distribution of θ and let the frequency of a type, $\phi(\theta)$, evolve according to

$$\dot{\phi}(\theta) = \phi(\theta)\left(w(\theta) - \bar{w}\right). \tag{26}$$

From the analysis in section 4.1 immediately follows that any distribution which supports an equilibrium share n^e with $s(n^e) = c$ also constitutes an evolutionary equilibrium according to (26). If $s(n^e) = c$ holds, there are no fitness-differences between free-riders and cooperators (compare Proposition 1) and we would get $w(\theta) = \bar{w} \Rightarrow \dot{\phi}(\theta) = 0$ for all θ. Similarly, the stability properties of such an equilibrium with $0 < n^e < 1$ carries over: any small deviation from n^e would either lead to a break down in cooperation or a move towards full cooperation.

The analysis of section 4.2 suggests that conditional cooperation will always dominate the two unconditional strategies in a heterogenous environment. This result holds for *any* evolutionary dynamics. Adaptation according to (26), however, would eliminate all preferences which induce an unconditional strategy. In an evolutionary equilibrium according to (26), the whole population would consist of conditional cooperators. All agents would cooperate in one equilibrium state ($n_a^* = 0$) and free-rider in the other state ($n_b^* = 1$). Any distribution of θ with $\phi(\theta) \geq 0$ for $\hat{\theta}(0) \leq \theta \leq \hat{\theta}(1)$ and $\phi(\theta) = 0$ otherwise, which supports these equilibrium states, would constitute an evolutionary equilibrium. Hence, the dynamics from (26) do (in general) not lead to a society with one homogenous level of norm sensitivity θ. Once there are only conditional cooperators (such that the two supported equilibrium states are $n_a^* = 0$ respectively $n_b^* = 1$), the adaptation process stops.

5.2 Quantitative Genetics

In section 4 we have applied a method from quantitative genetics to a cultural, social learning process. According to this approach, originally studied by Lande (1976), the trait θ follows a normal distribution and the frequency of a trait changes according to the fitness-differential $w(\theta)/\bar{w}$. If the fitness of a θ-type is above the mean population fitness, the frequency of these types will increase (and shrink otherwise). The resulting (non-normal) distribution is then transformed back to a normal distribution with a new mean. According to this approach, adaptation will result in a change in the mean trait value, $\bar{\theta}$, while the other two characteristics of the distribution – its normal character and the variance – are preserved.

Our motivation to apply this method is technical. The methodology provides a tractable tool to study the adaptation of a continuous distribution within the model from section 2. A formal analysis based upon the replicator process from (26) would cause sever technical problems,

related to the possibility of degenerate distributions and discontinuities in $\Phi(\theta)$. This would make the analysis of existence and stability of both, equilibrium states as well as evolutionary equilibrium distributions quite cumbersome.

Admittedly, the quantitative genetic method has also several limitations.[22] Most important, it implies an imperfect learning process, as the initial variance in θ is maintained during the course of evolution.[23] Hence, by using this method we neglect the case where all agents adapt one unique θ value (e.g. $\theta = 1$). Note, however, that such a perfectly homogenous population does in general not constitute a stable evolutionary equilibrium according to the replicator dynamic from (26) discussed above. In contrast to the quantitative genetic approach, however, the dynamic process from (26) does *not* allow for a co-existence of different strategies, i.e. free-riding, cooperation and conditional cooperation, in an evolutionary equilibrium within a heterogenous environment. The heterogeneity in behavioral patterns which emerges in the equilibrium characterized in Proposition 4 is only an artefact of the method which implies a constant variance. For the case of a normal distribution with infinitesimal small variance, however, the evolutionary equilibrium according to Proposition 4 would be a population of conditional cooperators (such that $n_a^* \to 0$ and $n_b^* \to 1$). For this special case, behavior – but not necessarily the distribution of θ – in the evolutionary equilibrium would be equivalent for replicator dynamics as well as the quantitative genetic approach.

5.3 Heterogeneous Environments

This paper introduces a concept of heterogeneous environments, where – in the context of multiple equilibria – society coordinates with fixed probabilities on one or another equilibrium state. One could extend and generalize the approach in several directions. First, we could study heterogenous environments with more than two equilibrium states (in scenarios with a higher number of stable equilibrium states n^*). Such an extension would somewhat complicate our analysis, since there would be more than 3 behavioral patterns. In particular, there would be different forms of conditional cooperation. E.g. for the case of three equilibria, $n_a^* < n_b^* < n_c^*$, we would observe conditional strategies $\left(x_a^i, x_b^i, x_c^i\right)$ with $(1, 0, 0)$ as well as $(1, 1, 0)$. Our main result – the fitness dominance of conditional cooperation over unconditional behavior – would not be effected. Which of the two conditional cooperative strategies yields a higher fitness, only depends on the comparison of a free-riders' fitness costs with the costs for cooperation in the three different equilibrium states.

Another possible extension is the endogenous formation of the likelihood π_j. We could relate the probability to face one particular equilibrium state to the size of the basin of attraction for this equilibrium n_j^*. From the discussion in section 2 it is clear, that a stable equilibriums' basin

[22] One crucial limitation of the method would be the case with evolutionary pressure on low *and* high θ-types to grow. This would suggest an evolution towards a bimodal distribution, which is excluded by assumption in Lande's approach. However, such a disruptive evolution cannot occur in our framework.

[23] One could justify this implication by a systematic noise embedded in the social learning process. If the errors in the adaptation process are normally distributed and remain constant during evolution, these deviations from perfect adaptation in θ would maintain a normal distribution $\Phi(\theta)$.

of attraction is defined by the position of the surrounding, instable equilibria (fixed points). For the case of two stable equilibria depicted in the example from panel (a) of figure 1, it is the location of the instable equilibrium n_c^* which separates the distinct basins of attraction. As an increase in $\bar{\theta}$ would shift the $\Phi(\theta)$-curve upwards, the level of free-riding for the instable fixed point would increase. Hence, with an increase in the mean norm sensitivity, the basin of attraction for the equilibrium with a low level of free-riding, n_a^*, becomes larger and the one of the other equilibrium n_b^* shrinks. Accordingly, the probability π_a (π_b) would increase (decrease) in $\bar{\theta}$. This effect would only quantitatively alter the properties of an evolutionary equilibrium in a heterogeneous environment. Endogenous probabilities π_j, however, could add further restrictions for the stability of an evolutionary equilibrium.

6 Conclusion

While the impact of heterogenous 'habitats' on evolutionary processes is well studied by biologists,[24] this idea has been so far neglected in evolutionary economics. In this paper we take a first step to close this gap in the literature. We develop a model of voluntary public good provisions in the context of a social norm for cooperation. As the power of the norm-enforcement depends on the level of cooperation, there is scope for multiple equilibria. Society may coordinate on an equilibrium with a high level of cooperation, where norm deviations would result in severe sanctions, or on a state with widespread free-riding and weak norm-enforcement. We link this multiplicity of equilibria to the idea of heterogenous habitats, in the sense that the evolutionary success of a certain norm-sensitivity, respectively the behavior induced by it, is evaluated for different equilibria of the game. Following an indirect evolutionary approach, preferences – i.e. individual norm-sensitivities – are then endogenously shaped according to their performance in both, equilibrium states with a strong norm as well as states with a weak norm. In such heterogenous environments, conditional cooperation is more successful than any unconditional strategy. In the 'cooperative' environment, conditional cooperators follow the norm and avoid the punishment free-riders incur. In the environment where the norm is weak and sanctions do hardly play a role, conditional cooperators reap the same payoff as free-riders, which dominates that of an (unconditional) cooperator. Hence, the preferences underlying conditional cooperation are well adapted to heterogeneous environments. An intermediate level of norm sensitivity allows individuals to react flexibly to different social situation. Thereby, they dominate unconditional strategies, which are specialized on one particular condition.

Members of modern human societies typically interact in various cooperation problems where cooperation fails sometimes but works quite well in other situations. Our analysis suggests that exactly this heterogeneity in our social environments is a driving force in the evolution of conditional cooperation.

[24] Among many others, see e.g. Levins (1968), Maynard Smith and Hoekstra (1980).

Appendix A – Section 3

For the density of the normal distribution, $f(\alpha)$, one can easily derive

$$\frac{\partial f(\alpha)}{\partial \bar{\alpha}} = f(\alpha)\frac{\alpha - \bar{\alpha}}{\sigma^2}. \tag{A.1}$$

Making use of this term in (12) and rearranging, we get

$$\frac{\partial \bar{w}}{\partial \bar{\alpha}} = \frac{1}{\sigma^2}\int \left[\alpha\, w(\alpha,\bar{\alpha})\, f(\alpha) - \bar{\alpha}\, w(\alpha,\bar{\alpha})\, f(\alpha)\right] d\alpha + \int \frac{\partial w(\alpha,\bar{\alpha})}{\partial \bar{\alpha}} dF(\alpha). \tag{A.2}$$

From (11) respectively (10) follows that the first expression in the first integral equals $\bar{\alpha}_s \bar{w}$, and the second expression is $\bar{\alpha}\bar{w}$. We arrive at

$$\frac{\partial \bar{w}}{\partial \bar{\alpha}} = \frac{\bar{w}}{\sigma^2}\left(\bar{\alpha}_s - \bar{\alpha}\right) + \int \frac{\partial w(\alpha,\bar{\alpha})}{\partial \bar{\alpha}} dF(\alpha). \tag{A.3}$$

Rearranging and substituting for (9) yields

$$\Delta\bar{\alpha} = \frac{\sigma^2}{\bar{w}}\left(\frac{\partial \bar{w}}{\partial \bar{\alpha}} - \int \frac{\partial w(\alpha,\bar{\alpha})}{\partial \bar{\alpha}} dF(\alpha)\right) \tag{A.4}$$

which is equivalent to

$$\Delta\bar{\alpha} = \frac{\sigma^2}{\bar{w}}\int w(\alpha,\bar{\alpha})\,\frac{\partial f(\alpha)}{\partial \bar{\alpha}}\, d\alpha. \tag{A.5}$$

Making use of (A.1) we finally get

$$\Delta\bar{\alpha} = \frac{1}{\bar{w}}\int w(\alpha,\bar{\alpha})\,(\alpha - \bar{\alpha})\, dF(\alpha). \tag{A.6}$$

Appendix A – Section 4

The mean fitness is given by

$$\bar{w} = -s(n^*)\int_{-\infty}^{\hat{\theta}(n^*)} d\Phi(\theta) - c\int_{\hat{\theta}(n^*)}^{\infty} d\Phi(\theta). \tag{A.7}$$

As $\Phi(\hat{\theta}(n^*)) = n^*$, we can rearrange $\bar{w}$ and get

$$\bar{w} = -\left(1 - n^*\right)c - n^* s(n^*). \tag{A.8}$$

From this follows (17).

As we have demonstrated in the section 3, only the direct fitness impact of a change in $\bar{\theta}$ is important for the evolution of this variable. The indirect effect – related to the frequency

dependent fitness from $s(n)$ – is irrelevant. Hence, we follow (13) and derive

$$\Delta\bar{\theta} = \frac{\sigma^2}{\bar{w}}\left(c - s(n^*)\right) \int_{-\infty}^{\hat{\theta}(n^*)} \frac{\partial\phi(\theta,\bar{\theta},\sigma^2)}{\partial\bar{\theta}} d\theta. \tag{A.9}$$

For the density of the normal distribution we get analogously to (A.1)

$$\frac{\partial\phi(\theta,\bar{\theta},\sigma^2)}{\partial\bar{\theta}} = \phi(\theta)\frac{\theta-\bar{\theta}}{\sigma^2}. \tag{A.10}$$

With this, we can rewrite $\Delta\bar{\theta}$ as

$$\Delta\bar{\theta} = \frac{1}{\bar{w}}\left(s(n^*) - c\right) \int_{-\infty}^{\hat{\theta}(n^*)} \phi(\theta)\left(\bar{\theta} - \theta\right) d\theta, \tag{A.11}$$

where the first term in the integral is equal to $n^*\bar{\theta}$. The second expression in the integral depicts the mean level of θ for agents with $\theta^i \leq \hat{\theta}(n^*)$. Using (19) we finally arrive at (18).

Appendix B

Proof of Proposition 3. From (A.11) we get

$$\begin{aligned}
\frac{d\Delta\bar{\theta}}{d\bar{\theta}} &= \frac{1}{\bar{w}}\left(s(n^*) - c\right) \int_{-\infty}^{\hat{\theta}(n^*)} \phi(\theta) - \phi(\theta)\frac{\left(\theta-\bar{\theta}\right)^2}{\sigma^2} d\theta \\
&\quad -\frac{1}{\bar{w}^2}\left[\frac{\partial\bar{w}}{\partial\bar{\theta}} + \frac{\partial\bar{w}}{\partial n^*}\frac{\partial n^*}{\partial\bar{\theta}}\right]\left(s(n^*) - c\right) \int_{-\infty}^{\hat{\theta}(n^*)} \phi(\theta)\left(\bar{\theta} - \theta\right) d\theta \\
&\quad +\frac{1}{\bar{w}}\left[s'(n^*) \int_{-\infty}^{\hat{\theta}(n^*)} \phi(\theta)\left(\bar{\theta} - \theta\right) d\theta + \left(s(n^*) - c\right)\frac{\partial\hat{\theta}(n^*)}{\partial n^*}\phi(\hat{\theta})\left(\bar{\theta} - \hat{\theta}(n^*)\right)\right]\frac{\partial n^*}{\partial\bar{\theta}}
\end{aligned} \tag{A.12}$$

where we made use of the Leibnitz Rule of integral differentiation to derive the last term in the third line's squared brackets. Rearranging and making use of (4), (7) and (19) we arrive at

$$\begin{aligned}
\frac{d\Delta\bar{\theta}}{d\bar{\theta}} &= \frac{1}{\bar{w}}\left(s(n^*) - c\right)\left[n^* - \int_{-\infty}^{\hat{\theta}(n^*)} \phi(\theta)\frac{\left(\theta-\bar{\theta}\right)^2}{\sigma^2} d\theta\right] \\
&\quad -\frac{1}{\bar{w}^2}\left[\frac{\partial\bar{w}}{\partial\bar{\theta}} + \frac{\partial\bar{w}}{\partial n^*}\frac{\partial n^*}{\partial\bar{\theta}}\right]\left(s(n^*) - c\right)\left(\bar{\theta}n^* - \bar{\theta}^*\right) \\
&\quad +\frac{1}{\bar{w}}\left[\left(\bar{\theta}n^* - \bar{\theta}^*\right) + \left(s(n^*) - c\right)\frac{\hat{\theta}(n^*)}{s(n^*)}\phi(\hat{\theta})\left(\hat{\theta}(n^*) - \bar{\theta}\right)\right] s'(n^*)\frac{\partial n^*}{\partial\bar{\theta}}
\end{aligned} \tag{A.13}$$

From Proposition 1 we know that an evolutionary equilibrium with $0 < n^e < 1$ is characterized by $s(n^e) = c$. Therefore, the expressions in the first and the second line of (A.13) equal zero for such an equilibrium n^e. Using (7), one can easily show that $\partial n^*/\partial\bar{\theta} \leq 0$ for any stable equilibrium state n^*. As $s'(n^*) \leq 0$ and $\bar{\theta}n^* > \bar{\theta}^*$ for $0 < n^* < 1$ it follows that the expression in the third line of (A.13) must be non-negative and we get $d\Delta\bar{\theta}/d\bar{\theta} \geq 0$ for any evolutionary equilibrium with $0 < n^e < 1$. Such an evolutionary equilibrium is never stable.
Let us now consider an evolutionary equilibrium with $n^{e1} = 1$. Since $\hat{\theta}(n^{e1}) \to \infty$ for $n^{e1} = 1$, the integral term in the first line of (A.13) equals the variance σ^2 and the term in the squared brackets becomes zero. For $n^{e1} = 1$ there also holds $\bar{\theta}n^* = \bar{\theta}^*$ and the expression in the second line of (A.13) also equals zero. From $s(n^{e1}) \to 0$, $\hat{\theta}(n^{e1}) \to \infty$ and $\bar{\theta}n^* = \bar{\theta}^*$ follows that the term in the third line's squared brackets is strictly negative. Together with $\partial n^*/\partial\bar{\theta} \leq 0$ and $s'(n^*) \leq 0$ this implies that $d\Delta\bar{\theta}/d\bar{\theta} < 0$ holds for $n^{e1} = 1$. ■

Proof of Proposition 5. Analogously to (A.13) we can derive from (23) and (24)

$$\begin{aligned}\frac{d\Delta\bar{\theta}}{d\bar{\theta}} &= \frac{1}{\bar{w}}\sum_j \pi_j \left(s(n_j^*) - c\right)\left[n_j^* - \int_{-\infty}^{\hat{\theta}(n_j^*)} \phi(\theta)\frac{\left(\theta - \bar{\theta}\right)^2}{\sigma^2}d\theta\right] \\ &\quad -\frac{1}{\bar{w}^2}\left[\frac{\partial\bar{w}}{\partial\bar{\theta}} + \sum_j \pi_j \frac{\partial\bar{w}}{\partial n_j^*}\frac{\partial n_j^*}{\partial\bar{\theta}}\right]\Psi \\ &\quad +\frac{1}{\bar{w}}\sum_j \pi_j \left[\bar{\theta}n_j^* - \bar{\theta}_j^* + \left(s(n_j^*) - c\right)\frac{\hat{\theta}(n_j^*)}{s(n_j^*)}\phi(\hat{\theta}(n_j^*))\left(\hat{\theta}(n_j^*) - \bar{\theta}\right)\right] s'(n_j^*)\frac{\partial n_j^*}{\partial\bar{\theta}}\end{aligned} \tag{A.14}$$

Since in an evolutionary equilibrium $\Psi = 0$ (Proposition 4), the second line of (A.14) equals zero. In an equilibrium as characterized in Proposition 4(ii), i.e. where $n_b^e < 1$, there holds $s(n_a^e) > c > s(n_b^e)$. If the squared bracket term in the first line is positive for equilibrium state n_b^e and negative for n_a^e, the expression in the first line of (A.14) is unambiguously negative. The two corresponding conditions are

$$n_a^e \leq \int_{-\infty}^{\hat{\theta}(n_a^e)} \phi(\theta)\frac{\left(\theta - \bar{\theta}^e\right)^2}{\sigma^2}d\theta, \qquad \text{and} \qquad n_b^e \geq \int_{-\infty}^{\hat{\theta}(n_b^e)} \phi(\theta)\frac{\left(\theta - \bar{\theta}^e\right)^2}{\sigma^2}d\theta. \tag{A.15}$$

(Note, that the integral term in (A.15) takes values in the range $(0, 0.5]$ for $0 < n_a^e \leq 0.5$ and $[0.5, 1)$ for $0.5 \leq n_a^e < 1$.)

Let us now turn to the third line of (A.14). Remember that $s'(n_j^*) \leq 0$ and $\partial n_j^*/\partial\bar{\theta} \leq 0$ since both equilibrium states n_j^* are stable as characterized by (8). It is therefore sufficient for the expression in the third line to be negative, if the term in the squared brackets is negative

for both equilibrium states. Rearranging, we get the following condition

$$n_j^e \leq \frac{\bar{\theta}_j^*}{\bar{\theta}} + \phi(\hat{\theta}(n_j^e))\ \hat{\theta}(n_j^e)\left(1-\hat{\theta}(n_j^e)\right)\left(1-\frac{\hat{\theta}(n_j^e)}{\bar{\theta}}\right), \tag{A.16}$$

where we have substituted for (4). The first term on the right hand side of (A.16) is positive for any $n^* > 0$. Moreover, for $n_a^* \leq 0.5$ there holds $\hat{\theta}(n_a^*) \leq \bar{\theta}$. Since $1-\hat{\theta}(n_j^*) = (s(n_j^*) - c)/\ s(n_j^*)$, $s(n_a^*) > c$ implies that the second term on the right hand side is also positive for $n_a^* \leq 0.5$. For an equilibrium state $n_b^* \geq 0.5$ we know that $\hat{\theta}(n_b^*) \geq \bar{\theta}$. From $s(n_b^*) < c$ then follows that the right hand side is again strictly positive. (As the first term approaches unity for $n_b^* \to 1$ and since the second term is strictly positive, the right hand side of (A.16) could be strictly larger than unity for high levels of n_b^*. For $n_a^* \to 0$, the second term will be positive, as $\hat{\theta}(0) > 0$ holds due to assumption A1. Hence, condition (A.16) should hold for extreme equilibrium levels of n_j^*.) ■

Literature

Azar, Ofer H. (2004), What sustains social norms and how they evolve? The case of tipping, *Journal of Economic Behavior and Organization* 54(1), 49-64.

Bester, Helmut and Werner Güth (1998), Is altruism evolutionary stable? *Journal of Economic Behavior and Organization* 34(2), 193-209.

Coleman, James S. (1990), *Foundations of Social Theory*, Harvard University Press, Cambridge (MA).

Elster, Jon (1998), Emotions and Economic Theory, *Journal of Economic Literature* 36(1), 47-74.

Falconer, Douglas S. and Trudy F.C. Mackay (1995), *Introduction to Quantitative Genetics.* 4th Edition. Addison Wesley Longman, New York.

Falk, Armin, Ernst Fehr and Urs Fischbacher (2005), Driving Forces Behind Informal Sanctions, *Econometrica* 73(6), 2017-2030.

Fehr, Ernst and Klaus Schmidt (2006), The Economics of Fairness, Reciprocity and Altruism – Experimental Evidence and New Theories, in: Serge-Christophe Kolm and Jean Mercier Ythier (Eds.), *Handbook on the Economics of Giving, Reciprocity and Altruism*, Vol.1, North Holland, Amsterdam.

Fershtman, Chaim and Yoram Weiss (1998), Social Rewards, Externalities and Stable Preferences, *Journal of Public Economics* 70(1), 53-73.

Fischbacher, Urs, Simon Gächter and Ernst Fehr (2001), Are People Conditionally Cooperative? Evidence from a Public Goods Experiment, *Economics Letters* 71(3), 397-404.

Frey, Bruno and Stephan Meier (2004), Social Comparisons and Pro-social Behavior: Testing 'Conditional Cooperation' in a Field Experiment, *American Economic Review* 94(5), 1717-1722.

Gächter, Simon (2006), Conditional Cooperation: Behavioral Regularities from the Lab and the Field and their Policy Implications, CeDEx Discussion Paper No. 2006-03, University of Nottingham.

Güth, Werner (1995), An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives, *International Journal of Game Theory* 24(4), 323-344.

Hirshleifer, David and Eric Rasmussen (1989), Cooperation in a Repeated Prisoner's Dilemma with Ostracism, *Journal of Economic Behavior and Organization* 12(1), 87-106.

Keser, Claudia and Frans van Winden (2000), Conditional Cooperation and Voluntary Contributions to Public Goods, *Scandinavian Journal of Economics* 102, 23-39.

Lande, Russel (1976), Natural selection and random genetic drift in phenotypic evolution, *Evolution* 30(2), 314-334.

Levins, Richard (1968), *Evolution in changing environments.* Princeton University Press, Princeton.

Lindbeck, Assar, Sten Nyberg and Jörgen W. Weibull (1999), Social Norms and Economic Incentives in the Welfare State, *The Quarterly Journal of Economics* 114(1), 1-35.

Masclet, David, Charles Noussair, Steven Tucker and Marie-Claire Villeval (2003), Monetary and Non-Monetary Punishment in the Voluntary Contributions Mechanism, *American Economic Review* 93(1), 366-380.

Martin, Richard and John Randal (2005), Voluntary Contributions to a Public Good: A Natural Field Experiment. Mimeo, Victoria University, New Zealand.

Maynard Smith, John and Rolf Hoekstra (1980), Polymorphism in a varied environment: How robust are the models? *Genetical Research* 35, 45-57.

Mengel, Friederike (2006), A Model of Immigration, Integration and the Cultural Transmission of Social Norms, Working Paper AD 2006-08, University of Alicante.

Rege, Mari (2004), Social Norms and Private Provision of Public Goods, *Journal of Public Economic Theory* 6(1), 65-77.

Riedl, Arno and Aljaž Ule (2002), Exclusion and Cooperation in Social Network Experiments, Mimeo, University of Amsterdam.

Roff, Derek A. (1997), *Evolutionary Quantitative Genetics.* Chapman and Hall, New York.

Shang, Jen and Rachel Croson (2005), Field Experiments in Charitable Contribution: The Impact of Social Influence on Voluntary Contributions to Public Goods, Mimeo, University of Pennsylvania.

Weibull, Jorgen W. (1995), *Evolutionary Game Theory.* The MIT Press, Cambridge (MA).