A MODEL OF IMMIGRATION, INTEGRATION AND CULTURAL TRANSMISSION OF SOCIAL NORMS*

Friederike Mengel

WP-AD 2006-08

Correspondence to: Friederike Mengel. Departamento de Fundamentos del Análisis Economico, Universidad de Alicante. Campus de San Vicente, 03080 Alicante, Spain. Teléfono: 965903614. Fax: 965903898. E-mail: friederike@merlin.fae.ua.es

Editor: Instituto Valenciano de Investigaciones Económicas, S.A.
Primera Edición Junio, 2006
Depósito Legal: V-2385-2006

* This work has benefitted enormously from discussions with my supervisor Fernando Vega Redondo. Special thanks go to Christian Traxler for his valuable suggestions. I also wish to thank participants at seminars and conferences in A Coru¡na, Jena, Karlsruhe, K¨oln and Valencia for their comments. Part of this research was conducted while I was staying at the Max-Planck Institute of Economics in Jena. I thank the Institute for its hospitality.

A MODEL OF IMMIGRATION, INTEGRATION AND CULTURAL TRANSMISSION OF SOCIAL NORMS.

Friederike Mengel

ABSTRACT

I present and study an evolutionary model of immigration and cultural transmission of social norms in a set-up where agents are repeatedly matched to play a one-shot interaction prisoner´s dilemma. Matching can be non-random due to limited integration (or population viscosity). The latter refers to a tendency of individuals to have a higher rate of interaction with individuals of their type than with similar numbers of other agents. I derive a cultural transmission mechanism in order to examine the influence of viscosity and of other institutional characteristics of society on the evolutionary selection of pro-social norms. The main findings are that strict norms, sustained by strong internal punishment, need either viscosity or strong institutional pressures to persist, while norms of intermediate strength persist under a variety of institutional characteristics. Endogenizing norm strength allows to identify two scenarios in which pro-social norms survive: One of rigidity in which separation (high viscosity) leads to monomorphic equilibria with strict norms for cooperation. And one of integration (low viscosity) where intermediate norms persist in polymorphic equilibria. Furthermore, with endogenous norms, viscosity and cooperation are not linked in a monotone way.

JEL classification: C70, C73, Z13.

1 Introduction

In the last decades Europe has experienced a marked rise in immigration and migration policy is one of the most fiercely debated issues in many european countries. Economists have been quick to analyze some of the consequences of migration on the host countries' economies. Topics studied include the effect of migration on wages, employment or the welfare state.[1] One issue, though, that has been mostly neglected is the impact of immigration on culture or more precisely on social norms prevailing in a society, in spite of the fact that this is one of the most widely discussed topics in european societies nowadays.[2] Maybe people worry just as much about the erosion of social norms and cultural values as a consequence of migration as they do worry about wages and jobs. The extent to which migration impacts culture and social norms is clearly linked to the issue of integration. On the one hand, social norms might be more easily transmitted and cultural clashes avoided if there is a high level of integration of immigrants. On the other hand, isolation can help to protect groups adhering to pro-social norms.[3]

This study is an attempt to provide some insights into these and other important issues within an evolutionary game-theoretic framework. I consider a set-up where agents are matched to play a one-shot interaction prisoner's dilemma in a society where there is a social norm for cooperation. And I address the following more general questions:

- Can pro-social norms survive immigration of agents that do not adhere to this norm ?

- How does the answer to the previous question depend on the institutions of a society and in particular on the degree of integration ?

From the standard perspective of a *direct evolutionary approach* the first question has a clear-cut answer: If evolutionary selective forces apply directly to strategies, if the population dynamics is payoff monotonic, and if matching takes places randomly within the whole population, cooperation is never evolutionary stable.[4]

The idea of integration appears in a somewhat different disguise in contributions to the biological literature - termed population viscosity there - and relates to matching probabilities.[5] In fact a population is called viscous if agents have an increased probability of interacting with agents of their own type. Viscosity is maybe the most narrow measure of the degree of integration of a society. The

[1] See Borjas (1999) for a survey on this literature.

[2] Most of the research on the relation between migration and culture is empirical and asks the reverse question. How does culture determine migration and trade ? There are few theoretical contributions related to this topic. One example is Kónya (2002, 2001) who presents macroeconomic models of cultural assimilation. An interesting empirical study of assimilation is DeLeire et al. (2004).

[3] Pro-social norms are norms that induce agents to act in a way conferring benefits to others at a cost to themselves. A norm for cooperation in the prisoner's dilemma is a well-known example. Used in this sense the concept of pro-sociality can be seen as more or less equivalent to the concept of altruism in evolutionary biology.

[4] Weibull (1995), Vega-Redondo (1996).

[5] The concept of population viscosity is due to Hamilton (1964). See also Price (1970).

second question is answered in this literature as follows: Cooperative behaviour can only survive if the society displays a high enough degree of viscosity. The intuition is clear: Cooperation in the Prisoner's dilemma promotes the fitness of defectors at the expense of the cooperators themselves. If cooperation is to survive as a trait it has to be that the benefits of this altruistic behaviour fall disproportionately onto other cooperators. This is the case whenever the population is sufficiently viscous.[6]

In contrast to direct evolution the *indirect evolutionary approach* has selection work on preferences instead of strategies. An (internalized) social norm for cooperation in the prisoner's dilemma affects an agent's preferences in the sense that deviations from the norm cause him to suffer feelings of guilt, shame, embarrasment or anxiety. Consequently norm-adherence in these approaches will be determined by the (material) fitness implications of the strategies induced by the norm. Bester and Güth (1998) or also Guttman (2003) study such mechanisms.

Cultural evolutionary models try to go beyond the pure fitness implications of preferences and (induced) strategies and consider explicitly the process of transmission of traits through either the family (vertical transmission), peer-groups (horizontal transmission) or socializing institutions of society (oblique transmission).[7] Gintis (2003) presents a model with exogenous vertical and oblique transmission and an (also exogenous) fitness-disadvantage for agents that have a preference for altruism. His main finding is that in order for the altruistic preference to survive the level of oblique transmission has to be sufficiently high. Henrich and Boyd (2001) consider a model in which norms are transmitted through social learning. In their model pro-social norms are stable because the horizontal transmission process stabilizes punishment of non-adherers.

The *rational socialization* approach to preference formation assumes that altruistic and forward-looking parents deliberately pass on preferences to their children trying to maximize what they, as parents, see to be the children's future well-being. Bisin, Topa and Verdier (2004) present a model of endogenous vertical transmission in which altruistic preferences survive, because minorities have higher incentives to socialize their offspring to their own preferences than majorities do. [8]

In contrast to the previous literature that focuses on individual preference traits, in this study I concentrate on social norms taking into account the role of society for the evolution of preferences. In many cases it is (internalized) social norms that shape preferences by determining what are socially and morally acceptable behaviours.[9] On the other hand what constitutes a social norm is to a large extent determined by what are common behaviours in a society. I derive

[6] Mitteldorf and Wilson (2000), Hamilton (1964), Bowles and Gintis (1997), Axelrod, Hammond and Grafen (2004).

[7] Cavalli-Sforza and Feldman (1981), Henrich and Boyd (2001), Henrich and Boyd (1998) Henrich and Gil-White (2000), Boyd and Richerson (2005), Richerson, Boyd and Henrich (2003), Henrich (2003).

[8] See also Guttman (2001a, 2001b).

[9] Azar (2001), Cialdini et al (1990), Grasmick and Green (1980), Liu (2003), Reno et al. (1993), Young (1998).

a cultural evolutionary model to analyze the interplay between economic incentives, the formation of social norms, the evolutionary selection of preferences and rational behaviour of agents given these preferences.

While cultural evolution in my model puts selection pressures on preferences it does not deny that holding their preferences fixed agents act rationally. In this aspect my model relates very much to the indirect evolutionary approach. It differs from this approach in conceiving the evolution of preferences as an essentially social and cultural phenomenon, placing a focus explicitly on social norms. This is the first contribution of the paper.

The second main contribution of this paper is to introduce the important question of population viscosity as a special institutional characteristic into the study of norm-transmission and the evolution of pro-social behaviour. Viscosity has been little studied in Economics and there are almost no formal models. An exception are Myerson, Pollock and Swinkels (1991) who extend Nash-equilibrium to viscous populations. Related studies in evolutionary biology are Henrich (2003), Boyd and Richerson (2002) or Mitteldorf and Wilson (2000) among others. I add to these studies by rigorously introducing population viscosity into a model of cultural evolution. Viscosity in my approach has two kinds of effects: Short-run effects by changing the incentives of rational players and long-run effects by affecting norm strength and the evolution of preferences.

Finally my paper is also related to studies of norm-guided behaviour in other fields of Economics. Lindbeck, Nyberg and Weibull (1999) use a model with endogenous social norms to examine the interaction of monetary incentives and social norms in the welfare state. In Benabou and Tirole (2005) norms with endogenous strength are part of a theory of pro-social behaviour. Unlike the analysis in my paper their models are essentially static. To my knowledge my study is unique in examining the consequence of endogenous social norms for the evolutionary selection of preference traits.[10]

For exogenous norms the main results of the paper are: If norms are very strict, in the sense that they induce strong feelings of guilt if violated, a high level of viscosity or other forms of institutional pressures are needed to have cooperation survive. Norms of intermediate strength on the contrary can survive under a variety of institutional settings.

Endogenizing social norms delivers the following results: Pro-social norms persist in two polar scenarios: One of rigidity in which separation (high viscosity) leads to monomorphic equilibria with strict norms for cooperation (sustained by high levels of internal punishment). Cooperation in this scenario is achieved through rigid population structures (viscosity) which in turn lead to strict norms. In this sense rigidity is self-reinforcing. The second scenario is one of an integrated society with intermediate norms sustained by lower internal punishment and displaying heterogeneity of types in equilibrium. Here integration stabilizes a polymorphic equilibrium with norms that are less strict. Thus

[10] Obviously my study also ties in with other studies of norm-guided behaviour in Economics such as Azar (2001), Elster (1989), Guttman (2001a), Nyborg and Rege (2003), Traxler (2005), Young (1998) among many others

in contrast to standard direct and indirect evolutionary approaches, my mechanism based on endogenous social norms always produces polymorphic equilibria in fully integrated societies (where matching is random).

Furthermore I show that - contrary to what is often taken for granted in the literature - viscosity and cooperation are not linked in a monotone way. Pro-sociality that is sustained through culturally transmitted social norms differs from genetically transmitted pro-sociality in this important aspect.

The exogenous institutional characteristics I consider - while having a quite straightforward and monotonic impact on behaviour whenever preferences are fixed - turn out to have interesting non-monotonic effects when one allows for changing social norms and the evolution preferences. The differing implications different institutional designs have together with the relatively fast speed of cultural (as opposed to biological) evolution make the results an issue for policy design. At the end of the paper I shortly discuss the welfare implications of different institutional designs.

The paper is organized as follows: In section 2 the model is described. In section 3 I study the equilibria of the basic model (with exogenous norm-strength) and in section 4 norm-strength is endogenized as described above. Section 5 concludes.

2 The Model

2.1 The Social Norm

Consider a society consisting of a (unit-mass) continuum of individuals I. Individuals are randomly and repeatedly matched in pairs to interact in Prisoner's dilemma type of situations.[11]

In the bilateral game each player has two actions available: X and Y. The *action set* $Z = \{X, Y\}$ is the same for all players $i \in I$. Payoffs from the (symmetric) Prisoner's dilemma interaction can be summarized by the following payoff matrix $A \in \mathbb{R}^{2\times 2}$ (payoffs for the row player):

	X	Y
X	a	0
Y	1	d

(1)

where $1 > a > d > 0$. It is well known that in this game Y is a dominant strategy for both players and consequently the unique equilibrium prediction leads to a payoff of d for both players.

Assume now that there is a social norm for cooperation ("for playing X in the Prisoner's dilemma") in the society. Individuals have internalized this norm and deviating from it thus causes them feelings of guilt, shame or embarrasment.[12]

[11] As explained in the introduction matching will not always be perfectly random. The exact matching technology is specified in section 2.2.

[12] In principle social norms can also be sustained by external mechanisms such as social disapproval. This is distinct from the internal, "self-imposed" sanctions I consider here. See

This psychological cost w is reflected in the following payoff matrix $A^w \in \mathbb{R}^{2\times 2}$.

	X	Y
X	a	0
Y	$1-w$	$d-w$

(2)

I will distinguish between three different strengths of the norm. In particular I will call the social norm *weak* if $w < \min\{1-a, d\}$, i.e. if violation of the norm causes feelings of guilt so weak that they are always outweighed by the material payoff-advantage of defecting (playing Y). In this case the two game forms (1) and (2) represent the same strategic context, namely that of a Prisoner's dilemma. I will call the norm *intermediate* in the following two cases: If $w \in [1-a, d]$ (2) represents a stag-hunt game, having two symmetric Nash-equilibria in pure strategies where both agents play the same strategy (either X or Y). If $w \in [d, 1-a]$ then (2) represents a chicken game, with two asymmetric Nash-equilibria in pure strategies where one players X and the other player Y. The unique symmetric Nash-equilibrium in this case is in mixed strategies where each player plays $\frac{w-d}{1-a-d}X \oplus \frac{1-a-w}{1-a-d}Y$. Finally I will call a norm *strict* if $w > \max\{1-a, d\}$, i.e. if the internal punishment caused by a norm-violation is so strong that cooperation is a dominant strategy for an agent having internalized this norm.[13]

Suppose now that there is migration of agents from a different cultural background, who adhere to different social norms. In particular let us assume that they do not have internalized the social norm w, so their payoffs in the Prisoner´s dilemma are given by matrix (1).[14] Then there are two different types in the economy. To model strategic interaction between these two types of agents we describe the following population game:

2.2 The Population Game

Let the *type space* be $T = \{0, w\}$ with typical element τ, where a w-type's payoffs are given by matrix A^w as defined in (2) and a 0-type's payoffs by matrix A as defined in (1). Agents have incomplete information about each other's type. When choosing an action $z_i \in Z$ in the bilateral game they estimate the type of their match from the distribution of types in the economy and from their knowledge about the matching technology described below.

The set of *population states* (or distributions of types) is $P = \{p : p \in [0, 1]\}$ where p denotes the share of w-types in the population. Obviously then the

Elster (1989) or Gintis (2003) for a discussion. An alternative modeling approach would be to assume that the psychological payoff-loss w depends on the opponent's action. In this case the analysis becomes slightly more complicated but results do not change qualitatively.

[13] I assume that agents in choosing their strategy rationally trade off material benefits and psychological incentives. Empirical support for this assumption can be found in Bosman and van Winden ((2001), (2002)). Theoretical papers employing the same or similar social norms are Benabou and Tirole (2005) or Lindbeck, Nyberg and Weibull (1999) among others.

[14] There is evidence of huge differences in the domains of cooperative behaviour among different cultural groups, that are independent of differences in environment or local regularities. (Henrich and Boyd 1998)

share of 0-types is $1-p$. A complete description of the population is given by the *population profile* (σ_0, σ_w, p) where $\sigma_\tau = (\sigma_X^\tau, \sigma_Y^\tau)' \in \mathbb{R}^{2\times 1}$ denotes the distribution of actions among τ-types.[15] σ_X^τ being the share of τ-types that use action X. Obviously $\sigma_z^\tau \in [0,1] \forall z \in Z, \forall \tau \in T$ and $\sigma_X^\tau + \sigma_Y^\tau = 1 \forall \tau \in T$ have to hold. $\sigma = (\sigma_0, \sigma_w)$ is the collection of these measures or the *distribution of actions* in the population. Furthermore distributions of actions where $\sigma_z^\tau \in \{0,1\} \forall z \in Z, \forall \tau \in T$, i.e. where all agents of the same type choose the same action are denoted (z_0, z_w) where z_0 indicates the action chosen by all 0-types and z_w the action chosen by all $w-$types.[16]

Matching takes place randomly in a viscous population. The latter meaning that individuals have a tendency to interact more often with individuals that are of the same type than agents of another type do. I measure the degree of integration of a society with the paramter $x \in [0,1]$, where $x = 1$ means that the society is fully integrated and $x = 0$ means that the society is fully viscous, implying that types interact with probability 1 among themselves and never with agents of another type.

For any fixed distribution of types p and degree of viscosity x *material payoffs* in the population game are given by the collection of bilinear vector fields

$$F(\sigma) = (\Pi^w(\cdot), \Pi^0(\cdot))$$

where $\Pi^w(\sigma) = (\Pi^w(X,\sigma), \Pi^w(Y,\sigma))' \in \mathbb{R}^{2\times 1}$ is the vector field that for a w-type associates expected material payoffs (corresponding to each of the possible actions $z_i \in Z$) to every distribution of actions in the population $\sigma = (\sigma_0, \sigma_w)$. Analogously $\Pi^0(\sigma)$ describes expected material payoff of a 0-type. The matching technology implies that

$$\Pi^0(\sigma) = \begin{pmatrix} \Pi^0(X,\sigma) \\ \Pi^0(Y,\sigma) \end{pmatrix} = \left[\begin{pmatrix} A & \mathbf{0} \\ \mathbf{0} & A \end{pmatrix} \begin{pmatrix} \sigma_0 \\ \sigma_w \end{pmatrix} \right]' \begin{pmatrix} 1-px \\ px \end{pmatrix} \tag{3}$$

and

$$\Pi^w(\sigma) = \begin{pmatrix} \Pi^w(X,\sigma) \\ \Pi^w(Y,\sigma) \end{pmatrix} = \left[\begin{pmatrix} A & \mathbf{0} \\ \mathbf{0} & A \end{pmatrix} \begin{pmatrix} \sigma_0 \\ \sigma_w \end{pmatrix} \right]' \begin{pmatrix} (1-p)x \\ 1-(1-p)x \end{pmatrix} \tag{4}$$

As a consequence of population viscosity a w-type is matched with probability $(1-p)x$ with a 0-type and with probability $1-(1-p)x$ with another w-type. While a 0-type is matched with probability px with a w-type and with probability $(1-px)$ with another 0-type.

The vectors $\begin{pmatrix} (1-p)x \\ 1-(1-p)x \end{pmatrix}$ and $\begin{pmatrix} 1-px \\ px \end{pmatrix}$ could thus be called the *matching vectors* of a w-type and a 0-type respectively. They are known to all agents at all times.

[15] Here $'$ indicates the transpose of the vector/matrix in question.

[16] Note that $z = (z_0, z_w)$ can be seen as formally equivalent to a pure strategy in the (bilateral) Bayesian game where z_0 denotes the action a player chooses conditional on being a 0-type and z_w the action a player chooses conditional on being a w-type.

If the society is fully viscous (i.e. if $x = 0$) the matching vectors are given by $(0,1)'$ for a w-type and by $(1,0)'$ for a 0-type. In this case (3) and (4) reduce to $\Pi^\tau(\sigma) = A\sigma_\tau \ \forall \tau \in T$. As in fully viscous societies both types interact exclusively among each other material payoffs for any agent depend neither on the distribution of types in the population p nor on the distribution of actions among agents of a distinct type.

If the society is fully integrated (if $x = 1$) matching is random and the matching vector given by $(1-p,p)'$ for both types. In this case (3) and (4) reduce to $\Pi^\tau(\sigma) = (1-p)A\sigma_0 + pA\sigma_w \ \forall \tau \in T$. With random matching material incentives are thus the same for both types.

Note that bilinearity of $\Pi^\tau(\sigma)$ implies that expected (material) payoffs of a τ-type i when using the "mixed action" $(\sigma_i X \oplus (1-\sigma_i)Y)$ are given by

$$\Pi^\tau(\sigma_i, \sigma) = \sigma_i \Pi^\tau(X, \sigma) + (1-\sigma_i)\Pi^\tau(Y, \sigma) \tag{5}$$

Individuals are von Neumann-Morgenstern expected utility maximizers and for any fixed distribution of types p *utility* (or total payoff) is given by

$$F(\sigma) = (\pi^w(\cdot), \pi^0(\cdot))$$

where $\pi^w(\sigma) = (\pi^w(X,\sigma), \pi^w(Y,\sigma))'$ is the expected utility of a w-type for each of his actions when he faces a distribution of actions in the population of σ. Utility relates to material payoffs as follows: For a 0-type where there are no psychological payoffs obviously $\Pi^0(\sigma) = (\Pi^0(X,\sigma), \Pi^0(Y,\sigma))' = \pi^0(\sigma)$. By contrast a w-type suffers a utility-loss everytime he plays Y and π^w is obtained from (4) by replacing A with A^w. The utility functions extend to mixed actions in an analogous way as given by (5).

Having specified payoffs we have a complete description of the population game $\Gamma = (I, T, Z, P, F(\cdot))$. To describe optimal behaviour I rely on the concept of Nash-equilibrium:[17]

Definition 1 *A Nash-equilibrium of the population game* Γ *is any population profile* (σ, p) *s.th.* $\sigma_z^\tau > 0 \Rightarrow z \in \arg\max_Z \pi^\tau(z, \sigma) \forall \tau \in T$.

We are interested in how rational behaviour in this population game affects the dynamics of norm-adherence and the evolutionary selection of preferences. To answer these questions we have to specify the process of cultural transmission of social norms:

2.3 The cultural transmission process

Social norms are adopted via 2 mechanisms: First they are transmitted horizontally via peer interaction. There is a huge amount of evidence that humans aquire much of their behaviour through social learning. However both theory

[17] Again there is a formal equivalence between the Nash-equilibria of Γ and the symmetric Bayes-Nash equilibria of the bilateral game with incomplete information.

and empirical research indicate that humans do not simply copy other individuals at random, but they seem to rely on rules that make copying of succesful agents more likely.[18] Success is identified here with material payoffs as described by Π^0 and Π^w. Norm transmission is typically seen to be biased in this sense because individuals with higher material payoffs are likely to enjoy higher status in society. This makes the norms they adhere to more appealing and gives them a higher cultural impact.[19] I will refer to this process as *(payoff-biased) horizontal transmission.*

Secondly the adoption of the pro-social norm is enhanced because institutions of society promote this norm. By structuring interactions institutions lead to framing and other situation construal effects that favor the spread of some social norms.[20] Also legal norms or public policies can induce social norms by stigmatizing some behaviours while promoting others.[21] The pro-social norm can be (more or less) explicitly transmitted through socialization institutions such as schools, universities or churches. And finally communication media can shift reference points and in this way affect norm-transmission. In my analysis one particular institutional characteristic - namely the degree of integration - is highlighted. I will subsume all other effects under the term *institutional transmission.*[22]

[18] Henrich and Boyd (2001,1998), Henrich and Gil-White (2000), Boyd and Richerson (2005).

[19] The relevant payoffs here are material payoffs. While this seems clearly the right approach in a biological context with genetic evolution, it is maybe not as natural in a set-up where evolution of preferences is a cultural phenomenon. Nevertheless psychological/emotional payoffs should not affect an agents aptitude as a cultural model for the following reasons: There is evidence suggesting that a) material wealth and happiness are separated by individuals (Bosman and van Winden (2001),(2002)) and b) that we are more likely to adopt the preferences of agents who are materially rich rather than of those individuals who are "happy" (Huck (1998)).

[20] Many experimental studies show that ideals and norms are not absolute but influenced by the institutional structure in which an agent is placed. (Hoffmann et al. (1994) Schotter, Weiss and Zapater (1996)). Alesina and Fuchs-Schündeln (2005) use data from separated and reunified Germany to test whether there exists a feedback effect from the economic regime on individual preferences. They find strong and significant evidence of the impact of institutions on preferences. See also Bowles (1998), Huck (1998) or Gintis (2003).

[21] Hirschmann (1984) recognized this fact when he remarked that raising the cost of antisocial behaviour might not be the appropriate policy measure whenever it is mainly values instead of tastes that drive behaviour. In this case he suggested that legal measures (such as prohibition) are more effective as they can shift norms.

[22] Related to this is the process of oblique transmission as considered by for example Gintis (2003), Gintis (2002). In this paper I am interested in the formation and transmission of social norms rather than in socialization through intergenerational (vertical or oblique) transmission of exogenous preference traits.

The cultural transmission process is illustrated in Figure 1.

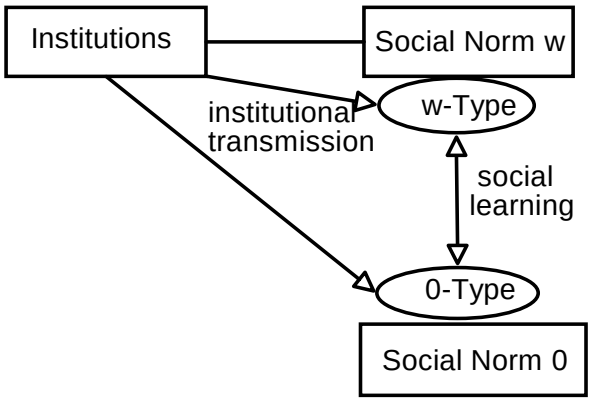

Figure 1: The Cultural Transmission Process

Horizontal transmission is modelled as follows: Suppose that at some point in time t an individual of type 0 meets an individual of type w with probability $p_t x$ and observes that individual's type and average material payoff $\Pi_t^w(\sigma_w, \sigma) =: \Pi_t^w$ in that period.[23] With complementary probability a 0-type meets another 0-type. And analogously with probability $(1-p_t)x$ a w-type meets a 0-type observing her type and average payoff and with complementary probability someone of her own type.

Denote the type of an individual's *cultural model* by $m \in \{0, w\}$. After such a random encounter individuals might adapt the norm of their model. More specifically if an agent's cultural model is of his own type he will stick to his norm with probability 1. If the cultural model is of another type he might adapt her norm with a probability that depends linearly on (positive) payoff-differences. The probability that an individual of type 0 adapts the w-norm at time t is then given by:

$$\Pr(w|0)_t = \begin{cases} (1-\alpha) + \alpha(\Pi_t^w - \Pi_t^0)1_+ & \text{if } m = w \\ 0 & \text{if } m = 0 \end{cases} \tag{6}$$

$\alpha \in [0,1]$ is a parameter that measures the importance of payoff for (horizontal) norm-transmission. 1_+ is the indicator function taking the value 1 if the preceding term is positive and 0 otherwise. Note that as $\Pi_t^w - \Pi_t^0 \in [-1, 1)$ this probability is always between 0 and 1. Analogously we have

$$\Pr(0|w)_t = \begin{cases} (1-\alpha) + \alpha(\Pi_t^0 - \Pi_t^w)1_+ & \text{if } m = 0 \\ 0 & \text{if } m = w \end{cases} \tag{7}$$

Independently of the parameter α this rule is neutral in the sense that if both types get the same payoff in expectation their population shares stay unchanged.

[23] In the following I will omit the argument of the payoff function when it can be done without ambiguity and denote average material payoffs of a w-type at time t by Π_t^w and average material payoffs of a 0-type by Π_t^0.

A high value of α implies that the norm transmission process is quite inert in the absence of payoff-differences. Social Learning in this case is highly adaptive. On the other hand if α is small the norm is transmitted with high probability even if the cultural model has lower or equal payoff. In this case agents can be seen as displaying a high degree of conformity in the sense that they easily (and without particular reason) adopt the norm of their model.[24]

The total share of w-types in the population after horizontal transmission can be computed easily using (6) and (7) as:

$$\begin{aligned} \Pr(w)_t &= p_t(1-\Pr(0|w)_t) + (1-p_t)\Pr(w|0)_t \\ &= p_t + (1-p_t)p_t x\alpha(\Pi_t^w - \Pi_t^0) \end{aligned} \tag{8}$$

Accordingly

$$\Pr(0)_t = 1 - \Pr(w)_t \tag{9}$$

denotes the total share of 0-types after horizontal transmission.

Let us now consider the impact of institutions: Under the influence of institutional pressures some of the 0-types from equation (9) will switch to the w-norm. I assume that institutional transmission is proportional to the "effective" number of w-types in the society, i.e. to the number of w-types a 0-type perceives in his environment $p_t x$. Having the parameter $\Delta \in [0,1]$ measuring the strength of institutional pressures the share of 0-types that adapt the w-norm because of institutional transmission is given by $p_t x\Delta$. While it is clear that institutional transmission cannot be independent of the effective number of w-types, the assumption of exact proportionality is maybe the most conservative in an attempt to keep the number of parameters in the model to a limit.[25]

Adding institutional transmission to horizontal transmission we get the following population dynamics:

$$\begin{aligned} p_{t+1} &= \Pr(w)_t + p_t x\Delta(1-\Pr(w)_t) \\ &= p_t + p_t(1-p_t)x[\alpha(\Pi_t^w - \Pi_t^0)(1-p_t x\Delta) + \Delta] \end{aligned}$$

or in continuous time

$$\dot{p} = p(1-p)x[\alpha(\Pi^w - \Pi^0)(1-px\Delta) + \Delta] =: f(p) \tag{10}$$

[24] It is important though to note that this is distinct from *conformist transmission,* as usually used in the literature. Conformist transmission refers to a tendency to copy the most frequent behaviour in a population. There is quite some evidence though - both theoretical and empirical - that apart from payoff-biases people display a tendency to copy frequent behaviour. (Henrich and Boyd (2001), Boyd and Richerson (2005)). See also Ellison and Fudenberg (1993) or Bernheim (1994). I will give a short discussion of conformist biases in Appendix 0.

[25] One could argue that immigrants create their own institutions promoting different norms. Letting Δ_w denote the strenght of established institutions and Δ_0 the strenght of new institutions (promoting the 0-norm) with $\Delta_w > \Delta_0$ the state equation is as follows:

$$\dot{p} = p(1-p)x[\alpha(\Pi^w - \Pi^0)(1-px(\Delta_w - \Delta_0) - x\Delta_0) + (\Delta_w - \Delta_0)]$$

It can be seen that focusing on $(\Delta_w - \Delta_0)$ and normalizing $\Delta_0 = 0$ does not impact the behaviour of this dynamics qualitatively. The basic assumption is that $\Delta_w > \Delta_0$.

Note that if $\Delta = 0$ this equals the familiar Replicator Dynamics (up to a change of time scale).

3 Cultural Equilibrium

I call a cultural equilibrium in this model a situation where - given equilibrium play in the population game - the share of norm-adherers in the population remains constant. Or more precisely:

Definition 2 *A cultural equilibrium is a population state p that satisfies $\dot{p} = 0$ in equation (10).*

Typically though we will be interested in cultural equilibria that are locally stable in the sense that the state trajectory can be kept arbitrarily close to the equilibrium state given that the initial state is sufficiently close.[26]

The set of locally stable cultural equilibria obviously depends on the strength of the norm. I will describe the different cases in turn.[27]

3.1 Weak Norm

If the norm is weak, defection is a dominant strategy in the bilateral game for both the w- and the 0-types. In this case the population dynamics is trivial: As payoffs for both types are the same horizontal transmission is neutral and the dynamics is governed by institutional transmission only. Full adherence to the w-norm ($p = 1$) is globally stable whenever $(\Delta, x) >> 0$. Note though that with weak norms norm-adherence leads to behaviour that is "phenotypically" indistinguishable from behaviour without the norm. Any population equilibrium will be characterized by full defection. The more interesting cases are consequently those in which the norm is of a strength to induce (at least sometimes) a different behaviour of the two types. These are the cases of intermediate and strict norms.

3.2 Strict Norm

If the norm is strict, cooperation is a dominant strategy for the w-type. Consequently all the Nash-equilibria of the population game are of the form (Y, X, p) i.e. population profiles where all 0-types play Y and all w-types play X. The

[26] Or more formally I will call p^* locally stable if $\forall R > 0, \exists r > 0, s.th. p(t_0) \in B_r \Rightarrow p(t) \in B_R, \forall t \geq t_0$, where B_R is an open ball around p^* with radius R, $B_R : \|p\| < R$. In fact most of the equilibria I will call locally stable below satisfy a stronger criterium of asymptotic local stability, i.e. they are not only stable in the above sense but also local attractors of the system.

[27] I will use the terms population equilibrium and cultural (population) equilibrium interchangeably to denote an equilibrium state p^* of equation (10) . Strategy profiles that constitute Nash-equilibria for any given population state p will be called (Nash-) equilibria of the population game.

cultural equilibria in this case are both monomorphic states as well as the polymorphic states p_1 and p_2.[28] Which of these will be locally stable depends on the vector of institutional characteristics (Δ, x). It is clear that very high institutional pressures Δ always lead to the spread of the w-norm. Let us then focus first on the more interesting case where Δ is arbitrarily small (but strictly positive). Integration impacts the set of stable cultural equilibria as follows:

If the degree of integration is very small (if $0 < x < \min\{\frac{a-d}{1-d}, 1-\frac{d}{a}\}$) the monomorphic equilibrium $p = 1$ is globally stable. The reason is that for low x both types mainly interact among each other. As a consequence w-types will get the high payoff for joint cooperation relatively often while 0-types will often get the lower payoff associated with mutual defection. This biases the social learning process in favor of the w-norm.

If integration takes on intermediate values two mutually exclusive cases arise depending on the payoff parameters. Cooperation survives in both: If $x \in (\frac{a-d}{1-d}, 1-\frac{d}{a})$ (implying that $a + d < 1$, i.e. that the material gains from unilateral defection are higher then the losses from unilateral cooperation in the Prisoner´s dilemma), the globally stable equilibrium is the polymorphic state p_1. The reason is that for low levels of norm-adherence p 0-types will obtain lower material payoffs in expectation, what biases social learning in favor of the w-norm and has p rise. As p rises this payoff bias shrinks and reverts at p_1. If on the other hand $x \in (1-\frac{d}{a}, \frac{a-d}{1-d})$ (implying that $a + d > 1$) this reasoning goes the other way round. The polymorphic equilibrium will be unstable and both monomorphic states will be locally stable with their basins of attraction separated by the interior equilibrium p_2.

Finally if the degree of integration is very high ($x > \max\{1-\frac{d}{a}, \frac{a-d}{1-d}\}$) 0-types will be able to benefit from the cooperative behaviour of the w-types and thus obtain a higher material payoff. Payoff-biased social learning then always works against the w-norm.

We have the following proposition:

Proposition 1 *If $w > \max\{1-a, d\}$, $\Delta > 0$ arbitrarily small and*
(i) if $0 < x < \min\{\frac{a-d}{1-d}, 1-\frac{d}{a}\}$ the globally stable equilibrium is $p^* = 1$.
(ii) if $x \in (\frac{a-d}{1-d}, 1-\frac{d}{a})$ the globally stable equilibrium is $p^* = p_1$.
(iii)if $x \in (1-\frac{d}{a}, \frac{a-d}{1-d})$ the locally stable equilibria are $p^* = \{0, 1\}$.
(iv) if $x > \max\{1-\frac{d}{a}, \frac{a-d}{1-d}\}$ the globally stable equilibrium is $p^* = 0$.

Proof. *see Appendix A* ■

The intuition for this result is clear: The strict social norm leads to strong internal sanctions for norm-violations. This induces norm-adherers to unconditionally cooperate in the Prisoner's dilemma thereby promoting the success of 0-types at the expense of the norm-adherers themselves. The strict norm survives as a preference trait if and only if the benefits of the altruistic behaviour it induces fall disproportionately onto other norm-adherers. This is the case

[28] The expressions for p_1 and p_2 are rather complicated and stated in Appendix A.

whenever the population is sufficiently viscous. The more integrated (less viscous) societies are the more institutional pressures are needed to sustain strict norms.

This point is made precise in the following corollaries: There are two critcial levels of institutional pressures that can ensure the persistence of the pro-social norm. These treshhold levels are given by $\Delta_1 =: \frac{[x(1-d)-(a-d)]\alpha}{1-x[(a-d)-x(1-d)]\alpha}$ and $\Delta_2 := \alpha[xa-(a-d)]$ for the two mutually exclusive paramter constellations where $a+d \lessgtr 1$. Note that both treshholds are strictly increasing with α and vanish if $\alpha = 0$. The reason is simply that for $\alpha = 0$ social learning displays no payoff-bias. But if material payoffs are irrelevant for the evolutionary selection of preferences any arbitrarily small level of institutional transmission will induce global convergence to $p = 1$. Note also that both treshholds rise with x. The intuition simply is that for strict norms more integration biases social learning against the pro-social norm. Consequently institutional pressures need to be higher to sustain it. Consider first the case where $a + d < 1$. This is the case where material gains of unilateral defection are higher than the opportunity costs of unilateral cooperation.

Corollary 1a *If $a+d < 1$ the monomorphic equilibrium state $p^* = 1$ is globally stable iff $\Delta > \Delta_1$.If $\Delta \in [\Delta_2, \Delta_1]$ cooperation survives in the polymorphic equilibrium $p = p_1$.*

Proof. Appendix A ■

In the second case where $a + d > 1$ we have:

Corollary 1b *If $a+d > 1$ the monomorphic equilibrium state $p^* = 1$ is globally stable iff $\Delta > \Delta_2$. If $\Delta \in [\Delta_1, \Delta_2]$ the monomorphism $p = 1$ is still locally stable.*

Proof. Appendix A ■

Figure 2 displays the state equation as a function of p and x for varying strengths of institutional pressures. [29] If Δ is small, as in Figure 2a, it is mainly the degree of integration of the society that acts as a selecting force to determine the set of locally stable equilibria of the system. It can be seen that for small x only $p^* = 1$ is locally stable, for intermediate x the globally stable equilibrium is polymorphic and for high levels of x $p^* = 0$ is globally stable. In Figure 2b the forces of institutional pressures outweigh the forces of integration, so $p^* = 1$ is globally stable, but for high x convergence is slow because of the weight of the induced payoff-bias against the strict norm. In Figure 2c institutional pressures dominate all other forces. Consequently $p^* = 1$ is selected, convergence being faster for higher levels of integration.

[29] The parameters used for the graphs are: $a = 1/2$, $d = 1/4$, and $\alpha = 1/2$.

We can sum up the findings of this subsection as follows:

Summary *Strict norms for cooperation do either need separation (high population viscosity) or sufficiently strong institutional pressures to persist in a cultural population equilibrium.*

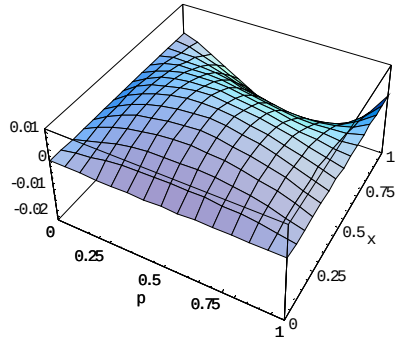

Figure 2a: $\Delta = 0.1$

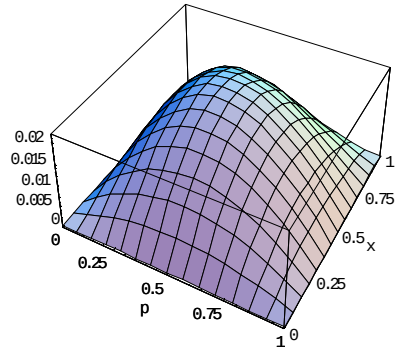

Figure 2b: $\Delta = 0.2$

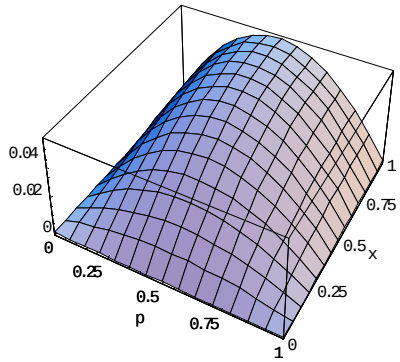

Figure 2c: $\Delta = 0.35$

As it should be clear by now that higher institutional pressures always enhance the evolutionary selection of the w-norm, I will focus in the following sections on the special, but most interesting case where Δ is arbitrarily small but strictly positive.

3.3 Intermediate Norm

If the norm is intermediate in strength, two mutually exclusive cases can arise depending on the payoff-parameters.

3.3.1 $a + d > 1$: Prisoners and Stag-Hunters

In this case the norm is intermediate whenever $\mathbf{w} \in [\mathbf{1 - a, d}]$. The payoff matrix A^w then describes a stag-hunt game. Remember that this (bilateral) game has two Nash-equilibria in pure strategies in which either both players cooperate or both free-ride. To see what are the Nash equilibria of the population game first note that Y is still a dominant strategy for a 0-type. Clearly then the profiles where all players defect ((Y, Y, p)) are Nash-equilibria $\forall p \in [0, 1]$. The profiles (Y, X, p) where w-types play X (and 0-types Y) will be equilibria if and only if the share of w-types in the population is sufficiently high. This can be seen by noting that it will be optimal for a w-type to choose X in such an equilibrium iff

$$\pi^w(X, z^*) \geq \pi^w(Y, z^*) \tag{11}$$

where $z^* = (z_0^*, z_w^*) = (Y, X)$. This is equivalent to

$$p \geq \frac{(1 - w - a) - x(1 - d - a)}{x(a + d - 1)} =: \widetilde{p} \tag{12}$$

We can state the following result:

Proposition 2 *If $w \in [1 - a, d]$ the Nash-equilibria of the population game Γ are given by $(Y, Y, p)\forall p \in [0, 1]$ and $(Y, X, p)\forall p \in [\widetilde{p}, 1]$.*

Obviously for $p \geq \widetilde{p}$ an issue of equilibrium selection arises. I will assume that in this case the equilibrium (Y, X, p) is selected. This choice is rationalized by noting that the equilibrium in which w-types cooperate pareto-dominates the equilibrium in which there is full defection.[30]

Note that $x < \frac{a+w-1}{a+d-1}$ implies $\widetilde{p} < 0$ what in turn implies that given our assumption on equilibrium selection a w-type will cooperate unconditionally for all population shares p. But then the analysis for this range of x corresponds to the case with the strict norm discussed above.

Focus on the case where $x \geq \frac{a+w-1}{a+d-1}$. Then there exists a non-empty range of population shares $[0, \widetilde{p})$ in which w-types will find it optimal to free-ride,

[30] Assuming that players coordinate on the inefficient equilibrium would also be plausible. But then the parameter region is indistinguishable from that of the weak norm and the implied dynamics is rather trivial. That is why I focus on this case.

in this way depriving the 0-types of their payoff advantage from unilateral defection. Consequently if $p < \widetilde{p}$ both types will earn the same material payoff in expectation and the dynamics of norm-adherence will be governed exclusively by institutional transmission. This leads to a steady growth in norm-adherence until the share of w-types reaches $\widetilde{p}$. Obviously $p = 0$ is always unstable in this region. One can further subdivide this region into two subregions: If $x \in (\frac{a+w-1}{a+d-1}, \frac{a-d}{1-d}]$ the globally stable equilibrium is $p = 1$ as w-types earn on average higher material payoffs than 0-types under this parameter constellation. This fact biases social learning in favor of the w-norm. Whereas for very high x ($x > \max\{\frac{a+w-1}{a+d-1}, \frac{a-d}{1-d}\}$) the payoff-bias works against the w-norm rendering $p = 1$ unstable and thus $p = \widetilde{p}$ globally stable.

As can be seen in the next proposition long-run cooperation is enhanced compared to the case of strict norms.

Proposition 3 *If* $w \in [1-a, d]$, $\Delta > 0$ *arbitrarily small and:*
(i) if $0 < x < \min\{\frac{a+w-1}{a+d-1}, \frac{a-d}{1-d}\}$ *the globally stable equilibrium is* $p^* = 1$.
(ii) if $x \in (\frac{a-d}{1-d}, \frac{a+w-1}{a+d-1})$ *the locally stable equilibria are* $p^* = \{0, 1\}$.
(iii) if $x \in [\frac{a+w-1}{a+d-1}, \frac{a-d}{1-d}]$ *the globally stable equilibrium is* $p^* = 1$.
(ii) if $x > \max\{\frac{a+w-1}{a+d-1}, \frac{a-d}{1-d}\}$ *the globally stable equilibrium is* $p^* = \widetilde{p}$.
Proof. *Appendix A* ■

Under the conditions of this proposition and if the degree of integration is sufficiently high ($x > \frac{a+w-1}{a+d-1}$) every stable equilibrium involves cooperation (even if $(\Delta, x) \to (0, 1)$). This is in stark contrast with the case of the strict norm. With strict norms as $(\Delta, x) \to (0, 1)$ the set of locally stable equilibria reduces to $p^* = 0$. Norms of intermediate strength though will always survive in fully integrated societies (where $x = 1$). The reason for this difference lies in the fact that for an intermediate strength of the norm, norm-adherers are conditional cooperators, cooperating only if the share of norm-adherers p is high enough. Consequently they cannot be as easily exploited by non-cooperators.

Note that the behaviour of an agent adhering to the intermediate norm in these equilibria cannot be distinguished from the behaviour of an agent adhering to the strict norm. "Phenotypically" thus the equilibrium $p = 1$ is identical for both strict norms and intermediate norms (with $a + d > 1$). Finally observe that as the degree of integration rises $\widetilde{p}$ also rises implying that there will be more cooperation in any stable polymorphic cultural equilibrium for higher degrees of integration.

Next turn to the case where $a + d < 1$

3.3.2 $a + d < 1$: Prisoners and Chickens

The intermediate norm corresponding to this case is $\mathbf{w} \in [\mathbf{d}, \mathbf{1} - \mathbf{a}]$. The game form A^w then represents a chicken game. This (bilateral) game has two asymmetric Nash-equilibria in pure strategies in which one player plays X and one player Y. This has as a consequence that in a population with "many" w-types, there is no Nash-equilibrium where all w-types choose the same action z. In any

population equilibrium in this region a w-type will randomize. If on the other hand the share of 0-types is sufficiently high, a w-type will find it optimal to play X. (As in this case he is matched with high probability with a 0-type who has as a dominant strategy to play Y). This case occurs whenever

$$\pi^w(X, z^*) \geq \pi^w(Y, z^*) \tag{13}$$

where $z^* = (Y, X)$ or equivalently iff

$$p \leq \frac{(1-w-a) - x(1-d-a)}{x(a+d-1)} = \widetilde{p}$$

We can state the following result:

Proposition 4 *If $w \in [d, 1-a]$ the Nash-equilibria of the population game Γ are given by: $(Y, X, p) \forall p \in [0, \widetilde{p}]$ and $(Y, (\sigma_X^{w*}, (1-\sigma_X^{w*}), p) \forall p \in (\widetilde{p}, 1]$, where $\sigma_X^{w*} = \frac{w-d}{(1-(1-p)x)[1-a-d]}$.*

Proof. Appendix A ■

Now w-types cooperate (play X) if there is a low level of norm-adherence and randomize if p is high. Again for high degrees of viscosity ($x < \min\{1 - \frac{d}{a}, \frac{1-d-w}{1-d}\}$) full norm-adherence to the w-norm ($p = 1$) is globally stable as in this case norm-adherers will be mainly matched with other norm-adherers. For intermediate degrees of integration $a+d < 1$ (meaning that the gain of unilateral defection is higher than the (opportunity) cost of unilateral cooperation) implies that if w-types are matched mainly with each other and if $p \geq \widetilde{p}$ s.th. they use the mixed action $\sigma_w^* = (\sigma_X^{w*}, (1-\sigma_X^{w*}))$ they will obtain a higher payoff on average than 0-types mainly matched with each other. As the degree of integration rises this material payoff advantage will diminish and finally reverse in favor of the 0-types. For $p < \widetilde{p}$ w-types will cooperate and obtain lower material payoffs than 0-types whenever integration is high. Consequently for very high degrees of integration ($x > \max\{\frac{1-d-w}{1-d}, 1 - \frac{d}{a}\}$) $p = 0$ is globally stable. We have the following proposition:

Proposition 5 *If $w \in [d, 1-a]$,$\Delta > 0$ arbitrarily small and*

(i) if $0 < x < \min\{1 - \frac{d}{a}, \frac{1-d-w}{1-d}\}$ the globally stable equilibrium is $p^* = 1$.

(ii) if $x \in [1 - \frac{d}{a}, \frac{1-d-w}{1-d})$ the locally stable equilibria are $p^* = \{0, 1\}$.

(iii) if $x \in [\frac{1-d-w}{1-d}, 1 - \frac{d}{a}]$ the globally stable equilibrium is $p^* = p_1$.

(iv) if $x > \max\{\frac{1-d-w}{1-d}, 1 - \frac{d}{a}\}$ the globally stable equilibrium is $p^* = 0$.

Proof. *Appendix A* ■

Now the pro-social norm does not survive as a preference trait in fully integrated societies (as $(\Delta, x) \to (0, 1)$), even though norm-adherers are conditional cooperators. The reason lies in the fact that now w-types find it optimal to cooperate whenever they are few. This maybe somewhat paradoxical result comes from the incentives the payoffs in the chicken game provide. Let us compare these incentives to those in the stag hunt game: In the stag-hunt game

establishing joint cooperation is difficult because of "fear". A w-type fears that whenever he plays X he could be matched with someone playing Y and in this way be exploited. On the contrary in the chicken game the problem is "greed" rather than "fear": A w-type matched with someone who cooperates wants to play Y because unilateral defection is still profitable in spite of the existence of the pro-social norm. This incentive structure has as a consequence that in the stag hunt game higher shares of norm-adherence enhance cooperation by w-types, because a high share of norm-adherers can reduce the fear of being exploited by making this more unlikely. In the chicken game context it is a high population share of 0-types that enhances cooperation by w-types because the probability of the match defecting is high. But this renders $p = 1$ unstable in integrated societies while making $p = 0$ a global attractor.

Summing up the results from this section we have:

Summary *If norms for cooperation are of intermediate strength they can survive the cultural evolutionary process in both scenarios: high viscosity and high integration. High institutional pressures are necessary for the persistence of the pro-social norm in integrated societies under some parameter constellations but not under others.*

4 Endogenous Norm-strength

4.1 Equilibria

The baseline case of exogenous norms illustrates that norm-strength matters when it comes to determining the equilibrium share of norm-adherers. Typically though norm strength will not be exogenous. Rather it will depend on the informational environment such as for example the distribution of preferences in an agent's sample. In this section I endogenize norm-strength by linking it to the share of norm-adherers in society.

In particular I will assume that the strength of internal punishment rises with the number of norm-adherers in the sample of a particular w-type.[31] " It's not right what I'm doing, but as everybody else does so, it's ok." is a revealing phrase that often accompanies norm-guided behaviour. Well-known examples where the mere fact that a behaviour is common reduces the strength of internal sanctions include not going to vote, minor tax evasion, welfare dependency, not going to church, divorce or free-riding on public transport.[32]

[31] It should be clear that the relevant number here is the number (or share) of norm-adherers in a particular w-types sample and not the number of norm-adherers in the society. For example the internal sanctions someone suffers because he did not go to vote might be quite low if noone else he knows went to vote - independently of whether overall participation in the election was high or low.

[32] Empirical support for such norms can be found in studies of norm-guided behaviour in economics (Azar (2001), Nyborg and Rege (2003)), in the law-literature (Grasmick and Green (1982), Liu (2003)) or in social psychology (Cialdini et al. (1990), Reno et al. (1993)). For models employing similar norms in a different contexts see Benabou and Tirole (2005), Traxler (2005) and Lindbeck,Nyberg and Weibull (1999). For a general discussion see Elster (1989).

To formalize this idea denote the proportion of w-types in a w-type's sample by

$$s := [1 - (1 - p)x]$$

and let the strength of the norm be given by some function

$$w(s) : [0, 1] \to [0, 1]$$

s.th. $w(1) = 1$, $w(0) = 0$, $w(s) \in C^2$ and $\frac{\partial w(s)}{\partial s} > 0$. The sign of the derivative expresses the fact that more norm-adherence tends to make a norm stronger. The cultural equilibrium determines thus norm-strength which in turn determines the equilibrium. This sort of feedback-effects between equilibrium and social norm are a characteristic pattern for norm-guided behaviour. In fact as Benabou and Tirole (2005) emphasize it is sometimes the very fact that "it is just not done" that makes a given behaviour socially or morally unacceptable. In this way norm-adherence determines the strength of the norm. On the other hand the strength of a social norm affects peoples' preferences, actions and the likelihood that the norm is internalized. In this way the strength of the social norm determines the cultural equilibrium. Focusing on only one of the two aspects - equilibrium or norm - misses an important part of the picture.

The change in the strength of the norm is linked to the evolution of norm-adherence as follows:

$$\dot{w} = \frac{\partial w(s)}{\partial s} x \dot{p} \tag{14}$$

It can be seen that population viscosity enhances norm strength($\frac{\partial s}{\partial x} < 0$). Higher population viscosity implies that norm-adherers mainly interact among each other, so in each norm-adherer's sample the share of norm-adherers will be very high and consequently the norm very strict.

This fact will strongly impact our previous results, as it breaks the monotone relationship between viscosity and cooperation. Consider first the case where the payoff matrix (1) is such that $\mathbf{a} + \mathbf{d} > \mathbf{1}$:

If the degree of integration is low norm- adherers almost exclusively interact with other norm-adherers. This implies that the share of norm-adherers in any w-type's sample is high, the social norm strict and thus (as we know from Section 3.2) only sustainable through very high degrees of viscosity. In this sense rigidity is self-reinforcing: Rigidity (viscosity) leads to strict norms which in turn need even more rigidity (either viscosity or strong institutional pressures) to persist.

Whenever $0 < x < \min\{1 - \frac{d}{a}, \frac{a+w(\tilde{s})-1}{a+d-1}\}$[33] the society is sufficiently viscous to sustain strict norms, as the benefits of pro-social behaviour fall disproportionately on norm-adherers. In this parameter range the globally stable equilibrium is $p^* = 1$.

Slightly higher degrees of integration will still lead to strict norms, but viscosity will not be high enough to ensure a material payoff-advantage to norm-adherers. Consequently the social norm will not be selected by the evolutionary

[33] $\tilde{s}$ denotes the solution to the following fixed point equation: $\frac{(1-w(s)-a)-x(1-d-a)}{x(a+d-1)} = p$.

dynamics. As the degree of integration further rises norm strength will fall. Finally high degrees of integration will lead to intermediate norms sustained in polymorphic equilibria.

Note that if $x \to 1$ both monomorphic equilibria are unstable: The reason is that if $p \to 1$ the norm will be strict and thus not sustainable with high integration. On the other hand if $p \to 0$ the norm becomes weak driving the dynamics away from $p = 0$. Fully integrated societies thus sustain a globally stable polymorphic equilibrium with intermediate norm strength.

We have the following proposition:

Proposition 6 *If* $a + d > 1, \Delta \to 0$ *and*

(i) $0 < x < \min\{1 - \frac{d}{a}, \frac{a+w(\tilde{s})-1}{a+d-1}\}$ *the globally stable equilibrium is* $p^* = 1$

(ii) $1 - \frac{d}{a} < x < \min\{\frac{a+w(\tilde{s})-1}{a+d-1}, \frac{a-d}{1-d}\}$ *the stable equilibria are* $p^* = \{0, 1\}$

(iii) $\frac{a-d}{1-d} < x < \frac{a+w(\tilde{s})-1}{a+d-1}$ *the globally stable equilibrium is* $p^* = 0$

(iv) $\frac{a+w(\tilde{s})-1}{a+d-1} < x < \frac{a-d}{1-d}$ *the locally stable equilibria are* $p^* = \{\widetilde{p}, 1\}$

(v) $x > \max\{\frac{a-d}{1-d}, \frac{a+w(\tilde{s})-1}{a+d-1}\}$ *the globally stable equilibrium is* $p^* = \widetilde{p}$

Proof. Appendix B ■

There are two scenarios in which cooperation survives in a globally stable equilibrium: In very viscous society sustained by strict norms and corresponding high levels of internal punishment and in very integrated societies sustained by intermediate norms and correspondingly lower levels of internal punishment. Note that only the latter equilibria are polymorphic. Maybe somewhat counterintuitively, integrated societies thus sustain heterogeneity while viscous societies imply monomorphic equilibria. Also note that the share of norm-adherers for any polymorphic equilibrium is maximized at $x = 1$.

With endogenous norm strengths the relation between viscosity and norm-adherence (and thus cooperation) is not monotone. This contrasts with a commonly held opinion in the literature.[34] In biological contexts where preferences for cooperation are passed on genetically, higher degrees of viscosity always enhance the fitness of cooperators and thus make cooperative outcomes more likely. In a context where cooperation is sustained through social norms and where preferences for cooperation are transmitted culturally this ceases to be true. The reason is that in this context viscosity affects preferences via two channels: It affects behaviour and thus norm-adherence but it also affects the strength of social norms. The strength of the social norm affects again behaviour and norm-adherence.

[34] Bowles and Gintis (1997), Mitteldorf and Wilson (2000), Boyd and Richerson (2005), Wilson and Sober (1994).

To illustrate the results look at the following example:

Example I *Consider the most simple case where norm strength depends linearly on norm-adherence, i.e. where* $w(s) = s$. *Assume that* $a = 3/4$ *and* $d = 1/2$. *In this case* $a + d > 1$, *i.e. for a w-type the loss of unilateral cooperation is higher than the gain of unilateral defection. As can be seen in Figure 3 for* $x < 1/3$ *the norm is strict in equilibrium* ($w = 1$) *and* $p = 1$ *is globally stable. For* $x \in [1/3, 1/2]$ *both monomorphic equilibria are locally stable with strict norms in both cases. In the equilibrium* $p = 0$ *norm-strength is linearly decreasing in* x ($w = 1 - x$). *For* $x \in (1/2, 3/5]$ *norm strength is intermediate but only* $p = 0$ *is locally stable. The reason is that for* $x < 3/5$ *norm-adherers are unconditional cooperators even for intermediate norm-strengths. Finally for* $x > 3/5$ *norm-strength is intermediate* ($w = 2/5$) *and the polymorphic equilibrium* $\widetilde{p} = 1 - \frac{3}{5x}$ *is globally stable.*

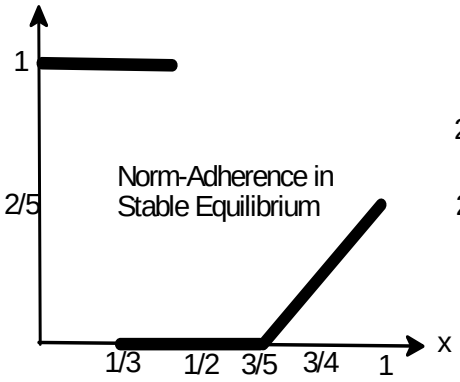

Fig. 3a: Locally stable equilibria

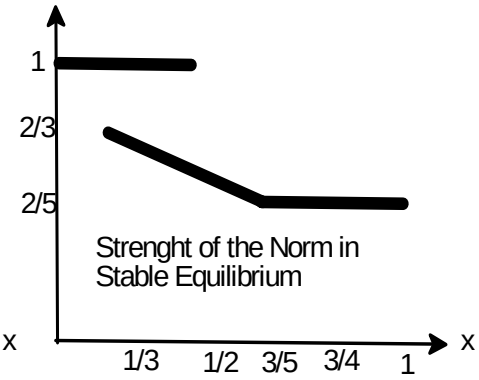

Figure 3b: Norm Strength

The case in which $\mathbf{a + d < 1}$ (where the material loss of unilateral cooperation is smaller than the gain of unilateral defection) delivers qualitatively the same result.[35] Cooperation survives in very viscous society sustained by strict norms in a monomorphic equilibrium and in very integrated societies sustained by intermediate norms in a polymorphic equilibrium.

[35]The case $a + d = 1$ is described at the end of Appendix B.

We can state the following proposition:

Proposition 7 *If* $a + d < 1, \Delta \to 0$ *and*
(i) $x < \frac{a-d}{1-d}$ *the globally stable equilibrium is* $p^* = 1$
(ii) $x \in [\frac{a-d}{1-d}, \min\{1-\frac{d}{a}, 1-w^{-1}(d)\}]$ *the globally stable equilibrium is* $p^* = p_1$
(iii) $x \in [1-\frac{d}{a}, 1-w^{-1}(d)]$ *the globally stable equilibrium is* $p^* = 0$
(iv) $x > \max\{1-w^{-1}(d), \frac{a-d}{1-d}\}$ *the globally stable equilibrium is* $p^* = \widehat{p} := 1 - \frac{1-w^{-1}(d)}{x}$

Proof. Appendix B ■

Note that for both cases $a + d \lesseqgtr 1$ the long-run equilibrium in fully integrated societies (where matching is random) is always polymorphic. This contrasts with what is obtained by relying on standard direct or indirect evolutionary mechanisms. Furthermore in all these polymorphic equilibria w-types are conditional cooperators. This is a behavioural pattern that is found also in many experimental studies on cooperation problems in western societies.[36]

We can sum up the results of this section as follows:

Summary *With endogenous norm strength and for vanishingly low levels of institutional pressures cooperation always survives in the following two scenarios: In very viscous societies sustained by strict norms in monomorphic equilibria and in very integrated societies sustained by intermediate norms in polymorphic equilibria.*

We have seen that with endogenous norm strengths the relation between viscosity and cooperation is non-monotonic (in contrast to for example what is obtained with direct evolutionary approaches). Exogenous institutional characteristics which drive behaviour in a very straightforward and monotonic way as long as preferences do not change can thus have interesting and non-monotonic effects when preferences are allowed to adjust through some selection dynamics. This has as a consequence that outcomes that seem (for policy makers) impossible to induce if preferences are assumed to be fixed can be induced by manipulating institutional characteristics if one carefully considers the evolution of preferences. The possibility of manipulating institutional characteristics in order to achieve a certain outcome is what motivates a welfare comparison of the different cases.

4.2 Welfare

Welfare analysis is quite problematic in our context because of two main problems: First preferences of individuals are not fixed over time. This problem can be tackled by comparing situations where preferences are stable, i.e. long-run equilibria. Note though that if the population state is polymorphic, individual

[36] Fischbacher, Gächter and Fehr (2001) find that roughly 50% of the participants in their public goods experiment are conditional cooperators, 30% always free-ride and only very few cooperate unconditionally. See also the references contained in their paper.

preferences are not fixed even in stable equilibria. What is fixed though in these equilibria is the distribution of preferences in the population. If one is willing to regard multiple "selves" as different individuals, the first problem can be regarded as essentially the same as the typical aggregation problem arising in welfare economics.

The second main problem with welfare analysis in our context is more severe and relates to the treatment of psychological payoffs. Sticking to revealed preference theory psychological payoffs constitute nothing else but an enlargement of the domain of preferences (in this way they are treated in this paper) and as such they should be included in any measure of welfare. The problem is that no assumption is (and maybe can be ?) made about the exact neural or psychological processes underlying these payoffs. In particular no point is made about the relation of positive emotions stemming from norm-conformity to negative feelings stemming from norm-violation. While this distinction is irrelevant for optimal behaviour and thus does not impact the previous analysis, it is clear that it has starkly differing welfare implications. This fact forces us to rely on material payoffs when making welfare comparisons. One possibility is to look at pareto-optimality.

We have the following result:

Proposition 8 *The stable cultural equilibrium $p = 1$ is always pareto-optimal. Furthermore $p = \widehat{p}$ as well as $p = \widetilde{p}$ and $p = p_1$ are pareto-optimal for some parameter constellations. $p = 0$ is never pareto-optimal.*

Proof. Appendix B ■

As it can be seen the pareto-criterion is not very discerning between equilibria. Another possibility then is to aggregate preference through some welfare function. I choose a classical utiliarianist criterion. With this criterion welfare in a polymorphic equilibrium depends only on the distribution of preferences. Using average material payoffs in long-run equilibrium as an indicator of welfare we state the following proposition:

Proposition 9 *(i) If $a + d > 1$ average material payoff is highest in any long-run equilibrium where $p^* = 1$ (independently of (Δ, x)).*

(ii) If $a + d < 1$ average material payoff is highest in a long-run equilibrium $p^* = \widehat{p} \in (0, 1)$.

Proof. Appendix B ■

The intuition is simply that in the first case where $a + d > 1$ the gains from unilateral defection $(1 - a)$ are smaller than the opportunity costs of unilateral cooperation (d). Consequently the state where everyone cooperates assures highest average payoffs independently of the value of (Δ, x). In the second case where $a + d < 1$ this relationship reverses. Average material payoffs are maximized in a polymorphic equilibrium with high integration and a non-vanishing level of institutional transmission.

The optimal choices of (Δ, x) for a policy-maker can also be guided by non-welfarist criteria or by aspects that are out of the scope of the present paper:

Obviously the degree of integration could constitute a policy goal per se. The reason is that high degrees of viscosity can be associated with high social costs as becomes clear if one reflects again about the introductory example of immigration. Another possible non-welfarist criterium could be the strength of the social norm. To the extent where strict social norms limit flexibility and impair the capacity of economic agents to adapt to varying environments, policy makers might be interested in bounding the strength of social norms. In short welfare analysis will depend very much on the particular context and finding suitable welfare-criteria is not straightforward in our context. It should have become clear though from the analysis that achieving a maximum level of adherence to the pro-social norm is not equivalent to maximizing welfare. In designing institutions policy makers have to account for the effect institutional characteristics have on social norms prevailing in a society and on the evolution of preferences.

5 Conclusions

In this paper I propose and study a cultural selection mechanism for preference traits. In particular I concentrate on social norms for cooperation and ask under which conditions norm-adherers can survive when matched in cooperative dilemmas with agents that do not adhere to these norms and thus do not cooperate. The main question examined is how the institutions of a society and in particular the degree of integration (as opposed to viscosity) impact norm adherence in the long run.

To these ends I present a cultural transmission process, based on two facts: 1) Agents adopt social norms from each other via processes of social learning. And 2) institutions affect the cultural learning process. One particular consequence of institutions is highlighted, namely the pattern of interaction they impose on the agents or more precisely the degree of integration. I find that strict norms for cooperation, inducing high levels of internal punishment, need either population viscosity or strong institutional pressures in order to survive. On the contrary intermediate norms can survive even in completely integrated societies and with vanishingly low levels of institutional pressures.

I endogenize the strength of the norm, assuming that it is positively correlated with the (subjectively felt) level of norm-adherence in the society. The results show that there are basically two scenarios under which cooperation can survive: The first scenario is that of a rigid society, displaying a high degree of viscosity and very strict norms sustained by strong internal punishment. Cooperation in this scenario is achieved through rigid population structures (viscosity) that in turn lead to strict norms. In this sense rigidity is self-reinforcing. The second scenario is one of an integrated society with intermediate norms sustained by lower internal punishment and displaying heterogeneity of types in equilibrium. Here integration stabilizes a polymorphic equilibrium with norms that are not as strict. In fact in fully integrated societies all stable equilibria are polymorphic. Furthermore there is always (conditional) cooperation in the long-run equilibrium. This contrasts with results obtained by relying on stan-

dard direct or indirect evolutionary mechanisms but is in line with experimental results.

Lastly my finding that population viscosity is not necessary for the evolutionary stability of pro-social norms also contrasts with results from the biological literature obtained by relying on replicator dynamics. Endogenizing the strength of social norms I find that, contrary to what kinship models or models of group selection suggest, it is not even clear that population viscosity locally enhances cooperation. If social groups are not completely isolated viscosity can be detrimental to cooperation. Given the recent revival of group-selection ideas it important to see how the rationale of these models depend on the process of cultural transmission assumed. With endogenous norm-strength viscosity and cooperation are not linked in a monotone way.

References

[1] Alesina, A. and N. Fuchs-Schündeln (2005), "Good bye Lenin (or not ?) - The effect of communism on people's preferences", NBER working paper.

[2] Axelrod, R., R.A. Hammond and A. Grafen (2004), "Altruism via Kin-Selection strategies that rely on arbitrary tags with which they co-evolve", *Evolution 58(8), 1833-1838.*

[3] Azar, Ofer H. (2001), "What sustains social norms and how they evolve ?", *Journal of Economic Behaviour and Organization 54(1),* 49-64.

[4] Benabou, R. and J.Tirole (2005), "Incentives and Prosocial Behavior", *NBER working paper 11535.*

[5] Bernheim, B.D. (1984), "Rationalizable Strategic Behaviour", *Econometrica 52(4).*

[6] Bernheim, B.D. (1994), "A Theory of Conformity", *Journal of Political Economy 102(5),* 841-877.

[7] Bester,H. and W.Güth (1998), "Is altruism evolutionary stable ?",*Journal of Economic Behaviour and Organization 34(2), 193-209.*

[8] Bisin,A., G.Topa and T.Verdier, (2004), "Cooperation as a Transmitted Cultural Trait", *working paper NYU.*

[9] Borjas, G.J. (1999), "The economic analysis of immigration", *Handbook of Labor Economics Vol 3A,* eds O.Ashenfelter and D.Card, North-Holland.

[10] Bosman, R. and F. van Winden (2001), "Anticipated and Experienced Emotions in an Investment Experiment", *Discussion paper 2001-058,* University of Amsterdam.

[11] Bosman, R. and F. van Winden (2002), "Emotional Hazard in a Power-to-Take Experiment", *Economic Journal 112,* 147-169.

[12] Boyd, R. and P.Richerson (1990), "Group Selection among alternative evolutionary stable strategies", *Journal of Theoretical Biology 145,* 331-342.

[13] Boyd, R. and P.Richerson (2002), "Group Beneficial Norms can spread rapidly in a Structured Population", *Journal of Theoretical Biology 215,* 287-296.

[14] Boyd,R. and P.Richerson (2005), "The Origin and Evolution of Cultures (Evolution and Cognition)", University of Chicago Press.

[15] Bowles,S. (1998). "Endogenous Preferences: The Cultural Consequences of Markets and other Economic Institutions", *Journal of Economic Literature 36, 75-111.*

[16] Bowles, S. and Gintis, H. (1998),"The Moral Economy of Communities: Structured Populations and the Evolution of Pro-Social Norms", *Evolution and Human Behaviour 19, 3-25.*

[17] Cavalli-Sforza,L. and M.Feldman (1981), "Cultural Transmission and Evolution", Princeton: Princeton University Press.

[18] Cialdini, R., B.Raymond, R.Reno and C.Kallgreen (1990), "A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places", *Journal of Personality and Social Psychology 58,* 1015-1026.

[19] DeLeire, T., B.Jacob, A.Lacuesta and R.LaLonde (2004), "The Americanization of Immigrant Children from Mexico", NBER working paper.

[20] Elison, G. and D.Fudenberg (1993), "Rules of Thumb for Social Learning", *Journal of Political Economy 101(4), 612-643.*

[21] Elster,J. (1989), "Social Norms and Economic Theory", *Journal of Economic Perspectives 3(4), 99-117.*

[22] Fischbacher, U., Gächter, S. and E.Fehr (2001), "Are people conditionally cooperative ? Evidence from a public goods experiment", *Economics Letters 71,* 397-404.

[23] Gintis, H. (2003), "The Hitchhiker's guide to Altruism: Gene-Culture Coevolution and the Internalization of Norms", *Journal of Theoretical Biology 220,* 407-418.

[24] Gintis, H.(2003b), "Solving the Puzzle of Prosociality", forthcoming *Rationality and Society.*

[25] Grasmick, H. and D. Green (1980), "Legal punishment, social disapproval and internalization as inhibitors of illegal behaviour", *Journal of Criminal Law and Criminology 71,* 325-335.

[26] Guttman, J. (2001a) "Self-Enforcing Reciprocity Norms and Intergenerational Transfers: Theory and Evidence",*Journal of Public Economics, 81,117-151.*

[27] Guttman, J. (2001b) "Families, Markets and Self-Enforcing Reciprocity Norms", *Annales d'Economie et de Statistique,(63/64) 89-110.*

[28] Guttman, J.(2003), "Repeated Interaction and the Evolution of Preferences for Reciprocity", *Economic Journal 113, 631-656.*

[29] Hamilton, W.D. (1964), "The Genetical Evolution of Social Behaviour", *Journal of Theoretical Biology 7,* 1-52.

[30] Henrich,J. and R.Boyd (1998), "The Evolution of Conformist Transmission and the Emergence of Between-Group Differences", *Evolution and Human Behaviour 19:215-241.*

[31] Henrich,J. and R.Boyd (2001), "Why people punish defectors", *Journal of Theoretical Biology 208,* 79-89.

[32] Henrich,J. and F.Gil-White (2000), "The evolution of prestige. Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission", *Evolution and Human Behavior 22, 165-196.*

[33] Henrich, J. (2003), "Cultural group selection, coevolutionary processes and large-scale cooperation", *Journal of Economic Behaviour and Organization.*

[34] Hirschmann, A.O. (1984), "Against parsimony: Three Easy Ways of Complicating Some Categories of Economic Discourse", *American Economic Review 74(2),* 89-96.

[35] Hoffman, E., K.McCabe, K.Shachat and V.Smith (1994), "Preferences, Property Rights, and Anonymity in Bargaining Games", *Games and Economic Behaviour 7,* 346-380.

[36] Huck, S. (1998), "Trust, Treason and Trials: An Example of How the Evolution of Preferences Can be Driven by Legal Institutions", *Journal of Law, Economics and Organization, V1411.*

[37] Kónya, I. (2001), "Optimal immigration, assimilation and trade", working paper Boston College.

[38] Kónya, I. (2002), "A dynamic model of cultural assimilation", working paper Boston College.

[39] Lindbeck,A., S.Nyberg and J.Weibull (1999), "Social Norms and Economic Incentives in the Welfare State", *Quarterly Journal of Economics 114,* 1-35.

[40] Liu, R.X. (2003), "The Moderating Effect of Internal and Perceived External Sanction Threats on the Relationship between Deviant Peer Associations and Criminal Offending", *Western Criminology Review 4(3),* 191-202.

[41] Mitteldorf,J. and D.S. Wilson (2000), "Population Viscosity and the Evolution of Altruism", *Journal of Theoretical Biology 204:* 481-496.

[42] Myerson R.B., G.B.Pollock and J.M.Swinkels (1991), "Viscous Population Equilibria", *Games and Economic Behaviour 3,* 101-109.

[43] Nyborg, K. and M. Rege (2003), "On social norms: the evolution of considerate smoking behaviour", *Journal of Economic Behaviour and Organization 52(3),* 323-340.

[44] Price, G. (1970), "Selection and Covariance", *Nature 227,* 520-521.

[45] Reno, R., R.Robert, B. Cialdini and C. Kallgreen (1993), "The Transsituational Influence of Social Norms", *Journal of Personality and Social Psychology 64(1),* 104-112.

[46] Richerson, P., R.Boyd and J.Henrich (2003), "Cultural Evolution of Human Cooperation" in: P.Hammerstein (ed.), *Genetic and Cultural Evolution of Cooperation,* MIT-Press.

[47] Schotter, A., A.Weiss and I.Zapater (1996), "Fairness and survival in ultimatum dictatorship games", *Journal of Economic Behaviour and Organization 31,* 37-56.

[48] Traxler, C. (2005), "Social Norms, Voting and the Provision of Public Goods, mimeo University of Munich.

[49] Vega-Redondo, F. (1996), "Evolution, Games and Economic Behaviour", Oxford University Press.

[50] Weibull, J. (1995), "Evolutionary Game Theory", Cambridge: MIT-Press.

[51] Wilson, D.S. and E.Sober (1994), "Re-Introducing Group Selection to the Human Behavioural Sciences", *Behavioral and Brain Science 17(4),* 585-654.

[52] Young, P. (1998), "Social norms and economic welfare", *European Economic Review 42* (821-830).

6 Appendix 0 (Conformist Transmission)

In this section I give an informal account of the effect of a conformist bias in horizontal transmission on the set of locally stable cultural equilibria.[37] Assume thus that now the switching probabilities (6)-(7) display a conformist bias as follows:

$$\Pr(w|0)_t = \begin{cases} (1-\alpha)p_t x + \alpha(\Pi_t^w - \Pi_t^0)1_+ & \text{if m=w} \\ 0 & \text{if m=0} \end{cases} \tag{15}$$

and

$$\Pr(0|w) = \begin{cases} (1-\alpha)(1-p_t)x + \alpha(\Pi_t^0 - \Pi_t^w)1_+ & \text{if m=0} \\ 0 & \text{if m=w} \end{cases} \tag{16}$$

Each individual weights the independent probability of adapting a norm with the popularity that a norm enjoys in the individuals sample. The parameter α now measures the relative importance of the conformist and the payoff-bias. Typically α should be larger than $1/2$, because if not even extremely beneficial norms could never spread. The state equation with popularity weighting is given by

$$\dot{p} = p(1-p)x[((1-\alpha)x(2p-1) + \alpha(\Pi^w - \Pi^0))(1-px\Delta) + \Delta] \tag{17}$$

The conformist bias can be nicely read from the first term in brackets.If $p > 1/2$,i.e. if there is a majority of norm-adherers this term is positive and - ceteris paribus - the share of norm-adherers will rise. If $p < 1/2$ this term is negative and thus norm-adherence will c.p. fall.

Obviously a conformist bias can locally enhance the stability of any monomorphic equilibrium. It is also clear though that this does not mean that polymorphic equilibria should dissappear. Whether they will depends on the relative strength of α and Δ as well as on payoffs.

7 Appendix A (Exogenous Norm strength)

Proof of Proposition 1:
Proof. Assume $a + d \neq 1$. (The case $a + d = 1$ is treated below). There are four zeros of (10): $p^* = 0, p^* = 1$ and

$$p^*_{1/2} = \frac{(1-a-d) + \Delta(a(1-x)-d)}{2(\Delta x(1-a-d))} \mp \frac{\sqrt{\begin{array}{c}[(a+d-1)\alpha x - \alpha x\Delta(a(1-x)-d)]^2 \\ -4(\alpha(a(1-x)-d)+\Delta)(\alpha\Delta x^2(1-a-d)\end{array}}}{2(\alpha\Delta x^2(1-a-d))}$$

[37] To state a formal proposition is rather complicated, as now the results will depend also qualitatively on the parameter α. Consequently many different cases and parameter constellations would have to be examined.

The derivative of the state equation evaluated at the two monomorphic equilibria is given by

$$f'(p)_{|p=0} = x(\Delta + \alpha((1-x)a - d)$$

$$f'(p)_{|p=1} = -x[\Delta + \alpha(1-\Delta x)((a-d) - x(1-d))]$$

$f'(p)_{|p_{1/2}}$ is a complicated expression. But we know that if $a + d \lesseqgtr 1$

$$\lim_{\Delta \to 0} p_{1/2} = \frac{a(1-x) - d}{(1-a-d)x} =: p_0$$

whereas the other zero diverges ($\lim_{\Delta \to 0} p_{2/1} = \infty$)

Furthermore we have that $f_\Delta(p)$ as given by (10) converges uniformly to $f(p) = p(1-p)x\alpha(\Pi^w - \Pi^0)$ as $\Delta \to 0$.[38] This can be seen by noting that

$$\begin{aligned} & |f_\Delta(p) - f(p)| \\ = \; & p(1-p)x\Delta(1 - px(\Pi^w - \Pi^0)) \\ \leq \; & \frac{\Delta}{4} \forall p \in [0,1] \end{aligned}$$

In addition $f'_\Delta(p) \overset{uniformly}{\to} f'(p)$. This allows us to write

$$\begin{aligned} \lim_{\Delta \to 0} f'_\Delta(p)_{|p_{1/2}} &= f'(p)_{|\lim_{\Delta \to 0} p_{1/2} = p_0} \\ &= -\alpha \frac{[(a-d) - xa][x(1-d) - (a-d)]}{1-d-a} \end{aligned}$$

Then it is easy to see that $p^* = 0$ is locally stable iff

$$0 < \Delta < \alpha[d - (1-x)a] := \Delta_2 \tag{18}$$

For $\Delta \to 0$ this condition reduces to $x > 1 - d/a$.

$p^* = 1$ is locally stable iff

$$((a-d) - x(1-d) > 0) \vee (\Delta > \frac{\alpha(x(1-d) - (a-d))}{1 - \alpha x((a-d) - x(1-d))} =: \Delta_1) \tag{19}$$

Again for $\Delta \to 0$ this condition reduces to $x < (a-d)/(1-d)$.

And (for $\Delta \to 0$) $p = p_{1/2}$ is locally stable iff

$$-\alpha \frac{[(a-d) - xa][x(1-d) - (a-d)]}{1-d-a} < 0 \tag{20}$$

[38] Actually as $f_\Delta(p)$ is a sequence of bounded functions mapping $[0,1]$ into $\mathbb{R}$ and $f(p)$ is bounded and also maps $[0,1]$ into $\mathbb{R}$ uniform convergence is equivalent to convergence in metric space $(F_{[0,1]}, d)$ where $F_{[0,1]}$ is the set of bounded functions from $[0,1] \to \mathbb{R}$ and d is the supremum metric.

Consider the four cases of Proposition 1 in turn:

(i) In this parameter range $f'(p)_{|p=0} > 0$ and $f'(p)_{|p=1} < 0$, so we have that $p^* = 0$ is unstable and $p^* = 1$ is locally stable. Continuity of $f(p)$ implies that the number of regular interior equilibria has to be even. As $\alpha(\Pi^w - \Pi^0)(1 - p_t x\Delta) + \Delta =: \Phi(p, \Delta)$ is a quadratic polynomial in p for any given Δ there are at most two regular interior equilibria. Two constellations of the payoff parameters have to be distinguished: If $a + d < 1$ we have that $p_2 > 1$. $a + d > 1 \Rightarrow p_1 < 0$. As there can neither be exactly two nor exactly one interior solution, there has to be none.

(ii) For the second part observe that in this parameter range $f'(p)_{|p=0} > 0 \ \forall \Delta \in [0, 1]$ while $f'(p)_{|p=1} < 0$ iff $\Delta \geq \Delta_1$. For Δ arbitrarily small both monomorphic equilibria are thus unstable. $a + d < 1 \Rightarrow p_1 \in (0, 1)$ while p_2 diverges. Note that the number of interior equilibria has to be odd. $f'(p)_{|p=p_0} < 0$ is implied by $a + d < 1$ ($\Leftrightarrow (a - d)/(1 - d) < 1 - d/a$).

(iii) Observe that in this parameter range $f'(p)_{|p=1} < 0$ whereas $f'(p)_{|p=0} < 0$ iff $\Delta \leq \Delta_2$. For $\Delta \to 0$, $p^* = 1$ and $p^* = 0$ are stable. The interior equilibrium p_2 is unstable (as $a + d > 1$) and separates the basins of attraction of the two locally stable equilibria.

(iv) In this region $f'(p)_{|p=1} < 0$ whenever $\Delta \geq \Delta_2$, while $f'(p)_{|p=0} < 0$ iff $\Delta \leq \Delta_1$. For arbitrarily small Δ it is clear that only $p = 0$ is stable. $f'(p)_{|p=p_0} > 0$ implies that interior equilibria are unstable. ■

Statement of result Case $\mathbf{a + d = 1}$:

With this parameter constellations cases (ii) and (iii) of Proposition 1 do not arise. It follows from straightforward calculation that whenever $x < 1 - \frac{d}{a} (= \frac{a-d}{1-d})$ the unique stable equilibrium is $p^* = 1$ and whenever $x > 1 - \frac{d}{a}$ the unique stable equilibrium is given by

$$p^* = \begin{cases} 1 & \text{if } \Delta > \Delta_1 \\ 0 & \text{if } \Delta < \Delta_1 \end{cases}$$

Proof of Corollary 1a:

Proof. "If": It follows from (19) that $\Delta > \Delta_1$ is sufficient for local stability of $p = 1$. $\Delta \in [\Delta_2, \Delta_1]$ implies that both monomorphic states are unstable. Exactly one regular interior zero thus exists. We know that if $a + d < 1$ this polymorphic equilibrium is locally stable.

"Only if": We know from (19) that local stability of $p = 1$ implies either $x < \frac{a-d}{1-d}$ or $\Delta > \Delta_1$. But note that $x < \frac{a-d}{1-d}$ implies $\Delta_1 < 0$. ■

Proof of Corollary 1b:

Proof. "If": It follows from (19) that $\Delta > \Delta_2 > \Delta_1$ is sufficient for local stability of $p = 1$. $\Delta \in [\Delta_1, \Delta_2]$ implies that both monomorphic states are locally stable.

"Only if": We know from (18) and (19) that global stability of $p = 1$ is sufficient for $x < 1 - \frac{d}{a} < \frac{a-d}{1-d}$. But note that $x < 1 - \frac{d}{a}$ implies $\Delta_2 < 0$. ■

Proof of Proposition 3:

Proof. First note that (Y, X, p) is a Nash-equilibrium iff $\pi_t^w(X, z^*) \geq \pi_t^w(Y, z^*)$

where $z^* = (Y, X)$. Substituting from (2) into

$$\begin{pmatrix} \pi_t^w(X, z^*) \\ \pi_t^w(Y, z^*) \end{pmatrix} = [1 - (1-p)x]A^w \begin{pmatrix} 1 \\ 0 \end{pmatrix} + (1-p)xA^w \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

it can be easily seen that $\pi_t^w(X, z^*) \geq \pi_t^w(Y, z^*)$

$$\begin{aligned} &\Leftrightarrow \quad [1-(1-p)x]a \geq [1-(1-p)x](1-w) + (1-p)x(d-w) \\ &\Leftrightarrow \quad p \geq \frac{(1-w-a) - x(1-d-a)}{x(a+d-1)} =: \widetilde{p} \leq 1 \end{aligned}$$

$\widetilde{p} > 0$ iff $x > \frac{a+w-1}{a+d-1} \in [0,1]$. Note also that if $p < \widetilde{p}$ the unique Nash-equilibrium of the population game is (Y, Y, p).

Given our equilibrium selection the population dynamics is then given by

$$\dot{p} = \begin{cases} p(1-p)x\Delta & \text{if } p < \widetilde{p} \\ p(1-p)x[\alpha(\Pi^w - \Pi^0)(1-px\Delta) + \Delta] & \text{if } p \geq \widetilde{p} \end{cases}$$

In the case of arbitrarily small Δ there are two zeros of this dynamics: $p^* = 0$ and $p^* = 1$. Note that $\lim_{\Delta \to 0} p_2 = p_0 < \widetilde{p}$ and $\lim_{\Delta \to 0} p_1 = \infty$.

The derivative of the state equation is

$$f'(p) = \begin{cases} (1-2p)x\Delta & \text{if } p < \widetilde{p} \\ (1-2p)x[\alpha(\Pi^w - \Pi^0)(1-px\Delta) + \Delta] & \\ \quad +p(1-p)x(-x\Delta\alpha(\Pi^w - \Pi^0)) & \text{if } p \geq \widetilde{p} \end{cases}$$

Note that $p = 0$ is unstable whenever $\widetilde{p} > 0$ and $x > 0$ as in this case

$$f'(p)_{|p=0} = x\Delta > 0$$

Furthermore we know that $p = 1$ is locally stable iff

$$\begin{aligned} f'(p)_{|p=1} \quad &= \quad -x[\Delta + \alpha(1-x\Delta)(a - d(1-x) - x)] < 0 \\ &\Longleftrightarrow \quad x < \frac{a-d}{1-d} \vee [x > \frac{a-d}{1-d} \wedge \Delta > \Delta_1] \end{aligned}$$

Remember that $x \leq \frac{a+w-1}{a+d-1} \Leftrightarrow \widetilde{p} < 0$. If this is the case w-types are unconditional cooperators and the proof of case (i) and case (ii) can be read directly from the Proof of Proposition 1.

Case (iii): $x \in (\frac{a+w-1}{a+d-1}, \frac{a-d}{1-d}]$. We have that

$$\begin{aligned} x \quad &> \quad \frac{a+w-1}{a+d-1} \Rightarrow \widetilde{p} > 0 \Rightarrow f'(p)_{|p=0} > 0 \\ x \quad &\leq \quad \frac{a-d}{1-d} \Rightarrow f'(p)_{|p=1} < 0 \end{aligned}$$

Consequently $p = 0$ is unstable ($\dot{p} > 0 \ \forall p < \widetilde{p}$) and $p = 1$ globally stable.

Case (iv): $x > \max\{\frac{a+w-1}{a+d-1}, \frac{a-d}{1-d}\}$: We have that

$$\begin{aligned} x &> \frac{a+w-1}{a+d-1} \Rightarrow \widetilde{p} > 0 \Rightarrow f'(p)_{|p=0} > 0 \\ x &> \frac{a-d}{1-d} \Rightarrow f'(p)_{|p=1} < 0 \end{aligned}$$

Consequently both $p = 0$ and $p = 1$ are unstable. As furthermore there is no interior regular equilibrium, $\widetilde{p}$ is stable with basin of attraction $[0, 1]$. ■

Proof of Proposition 4 and 5:

Proof. First note that (Y, X, p) is a Nash-equilibrium iff $\pi_t^w(X, z^*) \geq \pi_t^w(Y, z^*)$ where $z^* = (Y, X)$

$$\begin{aligned} &\Leftrightarrow \quad [1-(1-p)x]a \geq [1-(1-p)x](1-w) + (1-p)x(d-w) \\ &\Leftrightarrow \quad p \leq \frac{(1-w-a) - x(1-d-a)}{x(a+d-1)} =: \widetilde{p} \leq 1 \end{aligned}$$

If $p \geq \widetilde{p}$ w-types will randomize using action $\sigma_w^* = (\sigma_X^{w*}, (1-\sigma_X^{w*}))$. $\pi_t^w(X, \sigma) = \pi_t^w(Y, \sigma)$ implies

$$\begin{aligned} [1-(1-p)x]\sigma_X^w a \quad &= \quad [1-(1-p)x][\sigma_X^w(1-w) + (1-\sigma_X^w)(d-w)] \\ &\quad\quad +(1-p)x(d-w) \\ &\Longleftrightarrow \quad \sigma_X^{w*} = \frac{w-d}{[1-(1-p)x](1-a-d)} \end{aligned}$$

Exspected material payoffs of a w-type are thus given by

$$\Pi^w = \begin{cases} [1-(1-p_t)x]a & \text{if } p_t \leq \widetilde{p} \\ \frac{ad-w(1-w)-(1-p)(ad-(1-d)w)x}{(a+d-1)(1-(1-p)x)} & \text{if } p_t > \widetilde{p} \end{cases} \tag{21}$$

The exspected material payoff of a 0-type is

$$\Pi^0 = \begin{cases} p_t x + (1-p_t x)d & \text{if } p_t \leq \widetilde{p} \\ d + \frac{(1-d)(w-d)px}{(1-a-d)(1-(1-p)x)} & \text{if } p_t > \widetilde{p} \end{cases} \tag{22}$$

Remember the population dynamics

$$f_\Delta(p) = p(1-p)x[\alpha(\Pi^w - \Pi^0)(1-px\Delta) + \Delta] \tag{23}$$

where the payoffs are given by (21) - (22). We have

$$f'_\Delta(p)_{|p=0} = x(\Delta + \alpha((1-x)a - d)$$

and

$$f'_\Delta(p)_{|p=1} = -\frac{\alpha(w-d)x(1-d-w-x(1-d))}{1-a-d}$$

Then $p = 0$ is locally stable iff

$$0 < \Delta < \alpha(d - (1-x)a) = \Delta_1$$

and $p = 1$ is locally stable iff

$$x < \frac{1-d-w}{1-d}$$

Inserting the payoff-difference into (23) it can be easily seen that in the limit where $\Delta \to 0$ no interior regular equilibrium exists for the region where $p \geq \widetilde{p}$. In the region where $p < \widetilde{p}$ the unique regular interior equilibrium is given by p_1. Remember that $\lim_{\Delta \to 0} p_1 = p_0 < \widetilde{p}, \lim_{\Delta \to 0} p_2 = \infty$ and that $p_0 > 0$ is equivalent to $x < 1 - \frac{d}{a}$ in this parameter region. Furthermore given that $x < 1 - \frac{d}{a}$ stability of $p = p_1$ requires

$$\begin{aligned} f'(p)_{|p=p_1} &= -\alpha \frac{[(a-d) - xa][x(1-d) - (a-d)]}{1-d-a} < 0 \\ &\Leftrightarrow \quad x > \frac{a-d}{1-d} \end{aligned}$$

By noting that $1 - \frac{d}{a} > \frac{a-d}{1-d}$ and $\frac{1-w-d}{1-d} \geq \frac{a-d}{1-d} \forall w \in [d, 1-a]$ it can be easily seen that in

Case (i) $p = 0$ is unstable just as $p = p_1$ and thus $p = 1$ is globally stable
Case (ii) $p = 0$ and $p = 1$ are locally stable and $p_1 < 0$
Case (iii) $p = 0$ and $p = 1$ are unstable and thus $p = p_1$ globally stable
Case (iv) $p = 1$ is unstable, $p_1 < 0$ and thus $p = 0$ globally stable. ■

8 Appendix B (Endogenous Norm-strength)

In order to state the proof for Proposition 6 first note that $p = 0 \Rightarrow s = 1 - x$ and $p = 1 \Rightarrow s = 1$. Denote

$$\frac{(1 - w(s) - a) - x(1 - d - a)}{x(a + d - 1)} =: \Gamma(p)$$

and $\widetilde{p}$ the solution to $\Gamma(p) = p$ with corresponding norm strength $w(\widetilde{s})$.The following Lemma shows existence of such a solution:

Lemma 1 *There exists $\widehat{x} \in [0, 1]$ s.th. if $x \geq \widehat{x}$ there is a unique fixed point $\widetilde{p}$ (solving $\Gamma(p) = p$) with corresponding norm-strength $w(\widetilde{s}) \in [1 - a, d]$.*

Proof. First note that as $w(s) \in C^2, w(0) = 0$ and $w(1) = 1$ there exists $\widetilde{s}$ s.th. $w(\widetilde{s}) \in [1 - a, d]$. Assume that

$$x \geq \frac{a + w(\widetilde{s}) - 1}{a + d - 1} =: \widehat{x} \in [0, 1] \tag{24}$$

Furthermore note that $w(s) \in [1-a, d]$ implies $p \in [1 - \frac{1-w^{-1}(1-a)}{x}, 1 - \frac{1-w^{-1}(d)}{x}]$. Define

$$\Psi(p) = \Gamma(p) - p$$

Obviously $\Psi(p)$ is a continous function of p. Under condition (24) we have that $\Psi(p)$ maps the non-empty, compact and convex interval $[1-\frac{1-w^{-1}(1-a)}{x}, 1-\frac{1-w^{-1}(d)}{x}]$ into $\mathbb{R}$. Furthermore $\Psi(\ 1-\frac{1-w^{-1}(1-a)}{x}) = \frac{1-w^{-1}(1-a)}{x} > 0$. And $\Psi(\ 1-\frac{1-w^{-1}(d)}{x}) = \frac{-w^{-1}(d)}{x} \leq 0$ under condition (24). Consequently $\exists p^* \in [1-\frac{1-w^{-1}(1-a)}{x}, 1-\frac{1-w^{-1}(d)}{x}]$ s.th. $\Psi(p)=0$. Uniqueness can be seen by noting that

$$\Psi'(p) = \frac{-w'(s)}{(a+d-1)} - 1 < 0$$

i.e. that $\Psi(p)$ is strictly decreasing. ■

Proof of Proposition 6

Proof. From Propositions 1 and 3 it follows that given $a+d>1$ the interior zero p_2 will always be unstable independently of the strength of the norm. Note also that $a+d>1 \Leftrightarrow 1-d/a < (a-d)/(1-d)$.

Next examine the stability of the three candidates $p=0$, $p=1$ and $p=\widetilde{p}$:

Focus first on the case where $p=0$. Then we have that if $x > 1-w^{-1}(1-a)$ A^w corresponds to a Prisoner's dilemma payoff-matrix and consequently $p=0$ is unstable. If $x \in [1-w^{-1}(d), 1-w^{-1}(1-a)]$ A^w represents a stag-hunt game. In this case $p=0$ is stable iff $x \in [1-\frac{d}{a}, \frac{a+w(\widetilde{s})-1}{a+d-1}]$. Finally if $x < 1-w^{-1}(d)$ cooperation is a dominant strategy in game (2). Remember that in this case $p=0$ is locally stable iff $x > 1-\frac{d}{a}$. Noting that $\frac{a+w(\widetilde{s})-1}{a+d-1} > 0$ iff $x < 1-w^{-1}(1-a)$ we can summarize that $p^*=0$ is locally stable iff

$$x \in [1-\frac{d}{a}, \frac{a+w(\widetilde{s})-1}{a+d-1}] \wedge \Delta < \Delta_1 \tag{25}$$

On the other hand $p=1$ implies $w = w(1) = 1 > \max\{1-a, d\}$. Then it is clear that $p^*=1$ is locally stable iff

$$x < \frac{a-d}{1-d} \vee \left\{ x > \frac{a-d}{1-d} \wedge \Delta > \Delta_2 \right\} \tag{26}$$

Finally noting that $\frac{a+w(\widetilde{s})-1}{a+d-1} < 1 \Leftrightarrow x > 1-w^{-1}(d)$ we have for $\Delta \to 0$ that $\widetilde{p}$ is locally stable iff

$$x > \frac{a+w(\widetilde{s})-1}{a+d-1} \tag{27}$$

Comparing conditions (25), (26) and (27) it can be seen that in

Case (i): $p=0$ and $\widetilde{p}$ are unstable and $p=1$ is globally stable

Case (ii): both monomorphic equilibria are locally stable while $\widetilde{p}$ is unstable

Case (iii): $p=1$ and $p=\widetilde{p}$ are unstable and thus $p=0$ globally stable

Case (iv): $p=0$ is unstable while the other candidates are locally stable

Case (v): the monomorphic equilibria are unstable and $\widetilde{p}$ globally stable. ■

Proof of Proposition 7

Proof. Observe first that $(a-d)/(1-d) < 1-\frac{d}{a}$ in this parameter region. Consider the equilibrium $p=0$: In this case whenever $x > 1-w^{-1}(d)$ A^w corresponds to a Prisoner's dilemma payoff-matrix and consequently $p=0$ is

unstable. If $x \in [1 - w^{-1}(1-a), 1 - w^{-1}(d)]$ A^w represents a chicken game. In this case $p = 0$ is stable iff $x > 1 - \frac{d}{a}$. Finally if $x < 1 - w^{-1}(1-a)$ cooperation is a dominant strategy in game (2). Remember that in this case $p = 0$ is locally stable iff $x > 1 - \frac{d}{a}$.Summarizing thus $p^* = 0$ is locally stable iff

$$\Delta < \Delta_1 \wedge x \in [1 - \frac{d}{a}, 1 - w^{-1}(d)] \tag{28}$$

By contrast $p = 1$ is locally stable iff

$$x < \frac{a-d}{1-d} \ \vee \ \left\{ x > \frac{a-d}{1-d} \wedge \Delta > \Delta_2 \right\} \tag{29}$$

Observe then that in

Case (i): $p = 1$ is globally stable (independently of norm strength)

Case (ii): $p = 1$ and $p = 0$ are unstable and $\forall p \in [0, 1]$ the norm is either strict or intermediate. Consequently $p = p_1$ is globally stable (Proposition 1)

Case (iii): $p = 0$ is globally stable (as $1 - w^{-1}(d) > x > 1 - \frac{d}{a} > \frac{a-d}{1-d}$).

Case (iv): If $p = 0$ the norm is weak and consequently $p = 0$ is unstable, just as $p = 1$ (as $x > \frac{a-d}{1-d}$). We have that $\forall p$ s.th. $w(s) > d : \dot{p} < 0$. Whereas $\forall p$ s.th. $w(s) < d : \dot{p} > 0$.The globally stable equilibrium is thus the polymorphic state where the norm switches from being weak to being intermediate. This is the state where $w(s) = d$ or equivalently where $p = 1 - \frac{1-w^{-1}(d)}{x} =: \widehat{p}$. ■

Statement of result Case $\mathbf{a + d = 1}$

For this parameter constellation only two situations can arise: Whenever $w < 1 - a$ defection is a dominant strategy for both types. Whereas whenever $w > 1 - a$ defection is a dominant strategy for a 0-type and cooperation for a w-type.[39] We have: If $\Delta \to 0$ and

(i) $0 < x < \frac{a-d}{1-d} (= 1 - \frac{d}{a})$ the globally stable equilibrium is $p^* = 1$

(ii) $x \in [\frac{a-d}{1-d}, 1 - w^{-1}(d)]$ the globally stable equilibrium is $p^* = 0$

(iii) $x > 1 - w^{-1}(d)$ the globally stable equilibrium is $p^* = 1 - \frac{1-w^{-1}(1-a)}{x}$

Proof of Proposition 8

Proof. Simply look at the exspected material payoffs any non-zero measure of agents receives in each of the locally (or globally) stable equilibria: In the equilibrium $p = 0$ we have that $\Pi^0_{|p=0} = d$. While at $p = 1$ we have $\Pi^w_{|p=1} = a$.

At $p = \widetilde{p}$, $\Pi^w_{|\widetilde{p}} = (1 - (1 - \widetilde{p})x)a < a = \Pi^w_{|p=1}$and $\Pi^0_{|\widetilde{p}} = \widetilde{p}x + (1 - \widetilde{p}x)d$.

Note that $\Pi^0_{|\widetilde{p}} > a \Leftrightarrow -\frac{a^2 - a(d+x(1-d)) - (1-d)(1-d-w-x(1-d))}{-1+a+d} > 0$

$\Leftrightarrow w \geq \frac{(1-d)^2(1-x) + (1-d)ax - a(a-d)}{1-d} \in [1-a, d]$

At $p = \widehat{p}$, $\Pi^w_{|\widehat{p}} = (\sigma^{w*}_X)^2(1-(1-p)x)a + (1-\sigma^{w*}_X)(\sigma^{w*}_X(1-(1-p)x)(1-d)+d) < a = \Pi^w_{|p=1}$and $\Pi^0_{|\widehat{p}} = px\sigma^{w*}_X + ((1-px) + px(1-\sigma^{w*}_X))d$.

Note that $\Pi^0_{\widehat{p}} > a \Leftrightarrow w(s) \geq \frac{a(1-x(1-p)-a^2(1-(1-p)x)-(1-d)d(1-x)}{(1-d)px} \in [d, 1-a]$.

[39] If $w = 1 - a = d$ the bilateral game represented by A^w is trivial as all payoffs (matrix-entries) are equal.

At $p = p_1, \Pi^w_{|p_1} = (1-(1-p_1)x)a < \Pi^w_{|\widehat{p}} < a$ and $\Pi^0_{|p_1} = p_1 x + (1-p_1 x)d$. The latter expression is larger than $\Pi^0_{\widehat{p}}$ for high enough Δ and some $w(s)$. ■

Proof of Proposition 9

Proof. (i) Consider first the case where $a+d>1$: Note that in this case all w-types cooperate in the polymorphic equilibrium where $p = \widetilde{p}$. Average material payoff is thus given by

$$\bar{\Pi}(p) := p(1-(1-p)x)a + (1-p)px + (1-p)(1-px)d$$

We have that

$$p_{\min} = \frac{1}{2}(1 - \frac{a-d}{(a+d-1)x}) \in (0,1)$$

minimizes $\bar{\Pi}(p)$. In addition $a = \bar{\Pi}(1) > \bar{\Pi}(0) = d$. Consequently $\bar{\Pi}(p)$ is maximized for $p = 1$.

(ii) In the case where $a+d<1$ average material payoff in any polymorphic equilibrium is given by

$$\begin{aligned}\bar{\Pi}(p) &= p\{(\sigma_X^{w*})^2(1-(1-p)x)a + (1-\sigma_X^{w*})(\sigma_X^{w*}(1-(1-p)x)(1-d)+d)\} \\ &\quad +(1-p)\{px\sigma_X^{w*} + ((1-px) + px(1-\sigma_X^{w*}))d\} \\ &= d + \frac{\widehat{p}(w-d)(1-d-w)}{(1-a-d)(1-(1-p)x)}\end{aligned}$$

This payoff is larger than $a = \bar{\Pi}(1)$ whenever
$\widehat{p} > \frac{(a-d)(1-a-d)(1-x)}{w(1-w)-(1-a)ax-d(1-d-x(1-d))} \in (0,1) \forall w \in [d, 1-a)$. ■