The Emergence of Homogeneous Norms in Heterogeneous Populations

Dirk Helbing
Wenjian Yu
Karl-Dieter Opp
Heiko Rauhut

SFI WORKING PAPER: 2011-01-001

SFI Working Papers contain accounts of scientific work of the author(s) and do not necessarily represent the views of the Santa Fe Institute. We accept papers intended for publication in peer-reviewed journals or proceedings volumes, but not papers that have already appeared in print. Except for papers by our external faculty, papers must be based on work done at SFI, inspired by an invited visit to or collaboration at SFI, or funded by an SFI grant.

©NOTICE: This working paper is included by permission of the contributing author(s) as a means to ensure timely distribution of the scholarly and technical work on a non-commercial basis. Copyright and all rights therein are maintained by the author(s). It is understood that all persons copying this information will adhere to the terms and constraints invoked by each author's copyright. These works may be reposted only with the explicit permission of the copyright holder.

www.santafe.edu
The emergence of homogeneous norms in heterogeneous populations

Dirk Helbing,1,2 Wenjian Yu,1 Karl-Dieter Opp,3,4 and Heiko Rauhut1

1ETH Zurich, Chair of Sociology, in particular of Modeling and Simulation, Clausiusstr. 50, 8092 Zurich, Switzerland
2Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA
3University of Leipzig, Institute of Sociology (Emeritus Professor)
4University of Washington, Department of Sociology (Affiliate Professor)

January 14, 2010

Abstract

Social norms are known to establish social order and cohesion if actors commonly agree on them. In heterogeneous populations, however, normative conflict may result and social order may collapse. In this article, we show by means of a computational simulation model that homogeneous social norms may even come about in heterogeneous societies consisting of groups with competing interests. We demonstrate that punishment oriented at conformity can set off enforcement cascades leading to one generally accepted norm. In our model, agents put pressure on others to perform the same public behavior as they show themselves, even if they privately disapprove it. Interestingly, this type of punishment is more effective to form norms than pressuring others to meet their own private preferences. We conclude that group pressure and punishment may be interrelated phenomena, which can lead to homogeneous behaviors even in well diversified societies.

1. Introduction

Social norms can make individuals better off in situations in which different behaviors of interacting individuals are not mutually compatible, not fitting, or create a “collision of wills”. By prescribing “roles” as behavioral “standard solutions”, they can make social interactions more predictable and successful. Hence, behavioral consensus can create a benefit in itself. Moreover, norms guide human behavior in often hardly noticable ways, which has been described as an “invisible hand” kind of phenomenon (Smith 1986 (1776)). To reflect that norms are the basis of social institutions and social order, they have also been paraphrased as “Cement of Society” (Elster 1989b). However, social norms are not unchangeable. They are flexible and negotiable so that they have been described as “The Grammar of Society” (Bicchieri 2006).

When trying to explain the formation of social norms in the following, we consider it desirable to reproduce all of the following stylized facts:

- The occurrence of social norms requires that individual behaviors have (positive or negative) externalities. These can, for example, be material damages (like environmental destruction) or intangible effects (such as being dissimilar to others). These externalities make a coordinated behavior desirable (Bicchieri 2006; Coleman...
The establishment and maintenance of norms typically requires sanctioning efforts, punishing deviations from the norm or rewarding the compliance with it (Axelrod 1986; Cialdini and Trost 1998; Heckathorn 1989; Oliver 1980; Ostrom, Walker and Gardner 1992; Yamagishi 1986).

It can happen that unpopular norms are established which contradict the preferences of the majority (Bicchieri and Fukui 1999; Centola, Willer and Macy 2005; Willer, Kuwabara and Macy 2009).

Norms may have almost any content, but it is largely history- or path-dependent (Greif 1994; North 1990; Young 1996).

Typically, one finds a local consensus and global diversity, i.e. different local norms in different places (Axelrod 1997).

Norms can vary abruptly from one area to another and from one group to another. The separating borders are quite sharp.

While the first two points are directly implemented in our simulation model of norm formation, the last four points are not part of the model ingredients. They are rather results emerging from social interactions. We address situations of populations with heterogeneous preferences (cf. Winter, Rauhut and Helbing 2009). For illustration, consider a situation where new residents move to a village or city and mix with the population that already lives there. Let us further assume that the new residents (“population 2”) have certain preferences and exhibit particular behaviors which differ from those of the indigenous population (“population 1”). For example, there may be different ways of dressing, different habits, different values or religious beliefs, preferences for different kinds of food, or a different language or dialect. While the concept of norms is used in various meanings (Opp 2001), we refer in this paper to a social norm as a commonly shared behavior, discussing the relevance of sanctioning in its second part.

It goes without saying that each person in the two populations wants to perform his or her preferred behavior. However, we assume that there are externalities in favor of a commonly shared behavior or that it is rewarding to show a similar behavior as others. There are thus two conflicting motivations: to perform the preferred behavior and to perform the behavior of the majority. It may thus happen that people choose to perform the behavior they like best or that they choose the behavior that conforms to the behavior of the majority, although this may not be what they prefer to do (Horne 2009). The incentive to conform with the majority may result from the (positive) motivation to be like others or the (negative) motivation to avoid being different. In other words, behaving different or alike produces (negative or positive) externalities, which is supported by the theory of homophily (Flache and Mäs 2008; Lazarsfeld and Merton 1954; McPherson, Smith-Lovin and Cook 2001).

As second model ingredient we assume that individuals put pressure on others to adapt. That is to say, different behaviors are sanctioned. We distinguish two different kinds of sanctioning (or, equivalently, punishing): one, where people punish others if their actual behavior is different (behavior-based punishment), and another one, where sanctioning is applied when others do not choose the behavior that they prefer themselves (preference-based punishment).

Our simulation experiments assume individuals who interact with each other in space. The question that will be addressed is under what conditions the interactions and sanctioning behavior establish a generally accepted norm, i.e. the performance of a commonly shared behavior. Furthermore, we will identify conditions leading to a situation where each
individual performs the behavior that he or she prefers, i.e. a situation characterized by behavioral coexistence or “pluralism”.

Our modeling approach thus describes the conditions for a process of norm formation in a setting of the following kind: there are several (sub-)populations which are distinct with regard to the behavior they prefer. Hence, there is heterogeneity. The members of the different populations interact, which gives them a payoff. During the interaction there is a possibility to sanction the behavior of others. Moreover, individuals tend to imitate others with the same preference, when this promises a higher payoff. Our aim is to reveal under what conditions these assumptions lead to a shared norm or to behavioral coexistence.

The results of our computational model turn out to be plausible and can be summarized as follows: Homophily (i.e. an intrinsic satisfaction to interact with similar others) generally fosters the formation of homogeneous norms, and the outcome is path-dependent (Figures 1 and 2). A sufficient initial over-representation of one behavior in combination with homophily supports the spreading of this behavior over time and its formation as a norm, even if the population preferring that behavior is in a minority in the beginning (Figs. 3+4). When individuals interact within a certain spatial neighborhood, we find the formation of local norms and global diversity (Fig. 5) or the local coexistence of different behaviors (Fig. 6), depending on the model parameters.

The formation of a homogeneous norm can be largely promoted by sanctioning. Interestingly, however, the enforcement of similar behaviors (“behavior-based punishment) supports the formation of homogeneous norms (Fig. 8), while the enforcement of the agents’ preferences (“preference-based punishment”) is often ineffective (Fig. 9). Our results demonstrate that the different effects of these two types of punishment are due to the different kinds of hypocritical punishment associated with them (Fig. 7). We provide illustrative examples of these kinds of hypocritical punishment and demonstrate how important these are to understand the formation of norms in society.

Finally, if individuals apply group pressure on minorities in their neighborhood (Fig. 10), local norms can suddenly show up and spread under conditions, where different behaviors can coexist (Fig. 11). Moreover, once a norm is widely accepted, it persists for a long time, even if the agents stop their punishment efforts (Fig. 12).

The remaining parts of this manuscript are structured as follows: Section 2 presents a short overview of the literature on social norms. Then, Section 3 introduces our main modeling framework of social norms. Section 4 analyzes the simulation results without the consideration of punishment, while Section 5 investigates the effects of punishment and Section 6 the effects of group pressure, assuming that punishment efforts depend on the number of norm violators. Finally, Sec. 7 discusses our perspective on social norms and suggest future research lines.

2. Theories of Norm Emergence

There are numerous theories in the literature that address the formation of norms. In this section we will sketch the most important factors that are deemed relevant for norm emergence, and we show how these factors are related to our model.

Basic factors for the formation of norms. There are two factors that are components of most or perhaps all theories explaining the emergence of norms. One factor is externalities, i.e. the extent to which the behavior of an actor imposes costs or benefits on other actors. Although the term “externality” is often not used, the facts the term refers to are generally regarded as necessary for the emergence of norms. This holds already for Thomas Hobbes
(1962 [first 1651]), who argued that a war of everybody against everybody else (a situation where all actors impose extensive externalities on others) promotes the creation of a state that constrains individual actions. The theories by Demsetz (1967) and Coleman (1990) explicitly address the question when externalities generate norms.

The second factor that almost all theories of norm emergence refer to is sanctioning (or punishment). Examples are the theories by Heckathorn (1988; 1990), Coleman (1990) and Ellickson (1991). For most norms, sanctioning is costly.

Ullmann-Margalit (1977) distinguishes between coordination norms (“behavioral conventions”) and cooperation norms. Coordination norms are self-enforcing, i.e. everybody profits from following them. A typical example is the convention of pedestrians to walk on one side (Helbing 1992; Young 1993, 1996).

Cooperation norms, in contrast, are not self-enforcing, i.e. they do not automatically emerge. Therefore, the establishment of cooperation norms has often been considered to be a public goods or prisoner’s dilemma kind of problem: Cooperation is needed, but unlikely, because it requires that some or even all people overcome selfish behavior, at least temporarily. This theoretical idea is developed, for example, in the work of Demsetz (1967) and Coleman (1990). In our contribution, we will focus on situations, where norms are not self-enforcing because of heterogeneous, incompatible preferences.

Our model considers externalities in terms of positive sanctioning (similar behavior is assumed to be rewarding to some degree) as well as negative sanctioning (punishment efforts are considered in Sections 5 and 6). In contrast to previous approaches, we distinguish between behavior-based and preference-based punishment.

A further ingredient of most theories of norm formation are social networks. Ellickson (1991) argues that in so-called “close-knit groups” the formation and enforcement of cooperation norms is more likely than in groups with only weak social relations. In Coleman’s (1990) theory, social relations significantly contribute to solving the second-order free rider problem (see also Horne 2009; Helbing et al. 2010). Coleman argues that the integration into social networks encourages sanctioning and, consequently, the emergence of norms which ultimately instigate the contribution to the first-order public good (such as a clean environment).

Our model takes social networks into account by assuming that individuals interact with partners in a certain neighborhood. Such spatial interactions can give rise to a spatial clustering of similar behaviors, which is also known as “network reciprocity” (Nowak 2006). The same effect can also promote the existence of metanorms, i.e. the enforcement of punishment (Horne 2009; Helbing et al. 2010).

There are two basic processes of norm emergence: norms may emerge by human design or spontaneously by human action (see e.g. Axelrod 1986; Boyd and Richerson 2005; Hayek 1973; Opp 2002). Laws are an example for the first process, whereas numerous customs of everyday life, such as table manners, exemplify the second process. Most of the norms that govern social life emerge spontaneously. Therefore, we focus our analyses on spontaneous norm emergence in this article.

The basic factors of norm emergence mentioned above are all included in our model. In addition, our model goes beyond the existing theories as we study the dynamic interplay between these factors and the actors’s decision-making, and how they result in the emergence of norms.

**Previous modeling approaches.** Previous attempts to develop mathematical models of social norms can be mainly subdivided into two categories; namely (a) game-theoretical
models and (b) opinion dynamics models. The game-theoretical approaches treat norms primarily as coordination or cooperation problems. They are typically based on ultimatum games (Bicchieri 2006; Güth, Schmittberger and Schwarze 1982; Samuelson 1997), stag hunt games (Skyrms 2005), prisoner's dilemmas (Axelrod 1984; Bendor and Swistak 2001; Ullmann-Margalit 1977), or related concepts (Chalub, Santos and Pacheco 2006). The prisoner’s dilemma approach is probably the most pertinent one, treating cooperation norms as public goods that are unlikely to occur without a cooperation-promoting mechanism such as repeated interactions (Axelrod 1984; Fudenberg and Maskin 1986; Taylor 1976), trust and reputation (Camerer and Weigelt 1988; Neral and Ochs 1992; Raub and Weesie 1990), signaling (Bacharach and Gambetta 2001; Bird and Smith 2005; Gambetta 1994; Gintis, Smith and Bowles 2001; Molm, Takahashi and Peterson 2000; Murphy 2010; Raub 2004; Spence 1974; Van Winden 1998), punishment (Coleman 1990; Fehr and Gintis 2007; Ostrom, Walker and Gardner 1992; Voss 2001), or social control (Rauhut 2009; Rauhut and Krumpal 2008).

The treatment of norms as cooperation problem focuses on the question of how the commitment to a preexisting norm can be reached. Opinion dynamics models, in contrast, try to understand how one of several possible behaviors can establish as a norm. They typically address the issue of local consensus despite global diversity (Axelrod 1997; Deffuant, Huet and Amblard 2005; Ehrlich 2005; Hegselmann and Krause 2002; Kandel 1978; Kerr and Tindale 2004).

While the opinion dynamics models usually do not distinguish actual from preferred behaviors, the game-theoretical models often do not explain, how and why one of several possible norms is established. In the following, we propose a computer model which integrates the game theoretical perspective with the opinion dynamics one, considering interactions of two or more distinct populations. It is noteworthy that our simulation model is consistent with all the stylized facts summarized in the previous paragraph, not just with a subset of them.

3. The Multi-Population Norms Game

In the following, we will describe a computational, agent-based model for the formation of norms between individuals with different preferences. For the sake of illustration, we will focus on a spatial variant of the model rather than network interactions. In accordance with the so-called KISS principle requiring parsimonious modeling, our model is intentionally kept simple in order to be able to reveal cause-and-effect relations more easily. One may certainly imagine many generalizations, but most of them are not crucial to understand the aspects addressed by this work.

In our model, individuals are distributed on a chess-board-like, two-dimensional spatial grid with torus-like boundary conditions, so that everybody has the same number of neighbors. This space can be imagined to represent geographical space. In our simulations, the grid is fully occupied by so-called agents, which represent individuals. Agents do not relocate to other places, but they interact with other agents within a certain range $R$. Agents may show one of several behaviors $b$, and they have certain preferences $p$. These preferences are reflected by the payoffs $P_b$. A high preference means a high payoff (intrinsic satisfaction), if an individual performs her preferred behavior. When a less preferred behavior is shown, the related payoff is lower, zero, or even negative.

For the sake of simplicity, we will focus on the case of two alternative behaviors $b$ only. Then, some agents may prefer behavior $b=1$, while the others are assumed to prefer behavior $b=2$. Hence, based on their respective preferences $p$, agents can be subdivided into two
different (sub-)populations. In accordance with social psychology, behavior and preference (or belief) do not have to be consonant with each other, i.e. individuals may show the behavior they prefer \((b=p)\) or not \((b\neq p)\). Showing the preferred behavior yields a payoff \(P_b=B>0\), which reflects the benefit of doing what the individual prefers. When showing the non-preferred behavior, the intrinsic payoff is assumed to be \(P_b=0\).

In addition, individuals are assumed to have an advantage \(A>0\), when conforming with the behavior \(c\) of an interaction partner. This advantage could reflect homophily (an intrinsic satisfaction to interact with similar others) or positive externalities. This parameter is supported by the fact that an agreement on a certain behavioral “standard” (“role”) often increases the efficiency and success of a social interaction, or reduces related transaction costs. If two interaction partners show different behaviors \(b\) and \(c\), there is no additional payoff.

Our computer simulations start with uniformly distributed preferences and perform a random sequential update in three substeps: (a) Interaction with payoff, (b) imitation, and (c) randomization.

a. **Interaction.** With probability \(1/N\), one of the \(N\) agents (“agent 1”) is randomly chosen as focal agent. Then, an interaction partner (“agent 2”) is chosen within the radius \(R\) around the location of this agent. The interaction between both agents generates a payoff \(P_b\) for the focal agent and a payoff \(P_c\) for the interaction partner, which depends on the behaviors \(b\) and \(c\) of both agents and their respective preferences \(p\) and \(q\) (where \(q\) represents the preference of the interaction partner), see Table 1.

b. **Imitation.** Afterwards, another interaction partner of agent 1 (“agent 3”) is randomly chosen within the radius \(R\), namely among those agents who belong to the same population and, hence, shares the same preferences. If that interaction partner obtained a higher payoff during the last interaction, agent 1 imitates the behavior of that interaction partner with a probability proportional to the payoff difference, where the proportionality factor is \(1/(A+B)\). According to the “proportional imitation rule” (Helbing 1992; Schlag 1998), agent 1 otherwise sticks to the previous behavior. The same imitation step is applied to agent 2, but with a separately chosen interaction partner (“agent 4”).

We restrict this imitation step to in-group interactions, which appears to be justified by homophily, due to which individuals are more easily influenced by people who can serve as “role model”. It would certainly make little sense to copy the behavior of somebody with different preferences, as this would often lead to disappointing results.

c. **Randomization.** In order to take into account trial-and-error behavior, mistakes, or other factors contributing to randomness in decision-making processes, we assume that individuals would turn to the opposite behavior with a small probability \(r\) (“random strategy flipping”). Such a probabilistic model specification avoids artifacts that may easily occur in deterministic models (Helbing, Yu and Rauhut 2009).

After this update of the payoffs and the behaviors of agents 1 and 2, the computer program continues by selecting another focal individual (the new agent 1), performing the same update steps as described before. When \(N\) individuals have been updated (i.e. after updating \(N/2\) focal agents and their interaction partners in the interaction step), the simulation time \(t\) is increased by 1. Individual preferences are assumed to be constant in our model, i.e. they are not changed.
The payoff $P_b$ of an individual depends on whether its behavior $b$ is preferred ($b=p$) or not ($b≠p$) and whether it conforms with the behavior $c$ of the interaction partner ($b=c$) or not ($b≠c$). Further model parameters besides the advantage $A$ of conforming with the behavior $c$ of the respective interaction partner and the benefit $B$ of showing the preferred behavior $p$ are the interaction range $R$ and the rate $r$ of random strategy flipping. $S$ is the share of individuals belonging to population 1 (having the preferred behavior $p=1$), and $p_1$ denotes the average initial commitment of them to their preferred behavior, i.e. the fraction of individuals in population 1 actually showing behavior $b=1$ at time $t=0$. Similarly, $p_2$ is the average initial commitment of population 2 to their preferred behavior $p=2$.

**Table 1: Payoffs of the focal individual in the multi-population norms game.**

<table>
<thead>
<tr>
<th>Conformity of the focal individual with the behavior of the interaction partner</th>
<th>yes</th>
<th>no</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focal individual performs the preferred behavior</td>
<td>$A+B$</td>
<td>$B$</td>
</tr>
<tr>
<td>yes</td>
<td>no</td>
<td>$A$</td>
</tr>
</tbody>
</table>

Summarizing our computer model so far, individuals tend to imitate more successful individuals in the same population (who have the same preferences), and the success is determined by the payoff in interactions with a randomly chosen individual (see Table 1). A focal individual gets the highest payoff $A+B$, when it shows the preferred behavior $p$ and the interaction partner behaves the same ($b=p=c$). If the actual behavior of the focal individual deviates from the preferred one ($b≠p$), an advantage $A$ can be gained by conforming with the behavior $c$ of the interaction partner ($b=c$), otherwise there is no payoff ($P_b=0$). If the focal individual behaves differently from the interaction partner ($b≠c$), showing the preferred behavior ($b=p$) has a benefit $B$.

Given this payoff structure, one would expect an individual to show the preferred behavior if the benefit $B$ of showing the preferred behavior is higher than the advantage $A$ of conforming with the interaction partner ($B>A$). The interesting case is, therefore, characterized by $A>B$, where the payoff for conforming with somebody else is higher than the benefit of showing the preferred behavior. What will happen under such conditions? Will individuals end up showing the same behavior and, if yes, which one, or will they still show their own preferred behavior? We will see that all these variants are possible, and that the macro-level outcome depends on several factors. In any case, the formation of a commonly shared behavior always constitutes a normative dilemma in our model, as individuals with different preferences have an incentive to deviate from the norm whenever $B>0$. Therefore, according to the classification of Ullmann-Margalit (1977), our model studies the formation of cooperation norms, if $B>0$. For $B=0$, however, there are no unilateral incentives to deviate from a commonly shared behavior, and we have the case of a self-enforcing coordination norm, which is also called a “behavioral convention” (Helbing, 1992; Young, 1993).

Before we investigate the role of sanctioning (costly punishment) in our paper (see Sections. 5 and 6), it will be insightful to study first the behavior of the basic model described above. For the time being, we will therefore use the term “norm” in the sense of a commonly shared behavior. The related model will be called the multi-population norms game.
4. Results of Computer Simulations for the Multi-Population Norms Game

Our computer simulations assume that a share \( S \) of all \( N \) individuals prefers behavior 1 (i.e. \( SN \) individuals belong to population 1), while a fraction \( 1-S \) of individuals has a preference for behavior 2 (i.e. \((1-S)N\) individuals belong to population 2). The parameter \( S \) may be considered to reflect the \textit{relative power} (here: size) of both populations. If \( S=1/2 \), both populations are equally powerful, while population 1 is stronger/more powerful than population 2, if \( S>1/2 \). Of course, different power could also result from other factors than population size, such as material resources (money, weapons, etc.), social capital (status, social influence, etc.), charisma or moral persuasion.

In our computer simulations, each site of the spatial grid is occupied by one individual. The two different preferences are distributed over the two-dimensional simulation area in a random and uniform way, i.e. individuals with different preferences are mixed. A proportion \( p_1 \) of individuals belonging to population 1 starts with their preferred behavior and are represented by white squares, while a proportion \( 1-p_1 \) starts with their non-preferred behavior and are represented by a black squares with a white dot in it. Analogously, a proportion \( p_2 \) of individuals belonging to population 2 starts with their preferred behavior and are represented by black squares, while a proportion \( 1-p_2 \) starts with their non-preferred behavior and are represented by white squares with a black dot in it. Hence, the central dot reflects the \textit{preferred} behavior and the color of the surrounding frame to the \textit{actual} behavior. \( p_1 \) may be called the \textit{average initial commitment} of individuals of population 1 to their preferred behavior (and similar for population 2). Values of \( p_1 \) or \( p_2 \) smaller than 1 can reflect situations in which the populations are not fully committed to their preferred behaviors. This can result from path dependencies, random strategy flips, or interactions with individuals pursuing different behaviors.\(^1\) For example, when a social norm is established, the proportion of individuals showing the preferred behavior goes to values close to 0 in one of the populations. Besides, studying cases with \( p_1, p_2<1 \) or \( p_1 \neq p_2 \) reveals particularly interesting kinds of system dynamics.

In our computer simulations, when a commonly shared behavior is formed in the course of time, the great majority of grids either shows white squares (if the behavior preferred by population 1 wins through) or black squares (if the behavior preferred by population 2 wins through). If black and white squares are mixed and mostly do not have frames around them, this means that most individuals do what they prefer, and we have a coexistence of different behaviors.

\textbf{Path dependence.} Our first set of computer simulations assumes \( B<A \), a large interaction range \( R=10 \), equally strong populations \((S=1/2)\), and different initial fractions of individuals showing their preferred behavior in the beginning. We find a variety of possible results of our simulations (see Fig. 1). Even when all model parameters are exactly the same, the final outcome can be very different, depending on the initial conditions: If a majority of individuals in population 1 shows their preferred behavior in the beginning \((p_1>1/2)\), but not so in population 2 \((p_2<1/2)\), most individuals end up with this behavior (see Figs. 1a and 2). However, if a majority of individuals in population 1 initially shows the behavior preferred by population 2 \((p_1<1/2)\) and the individuals of population 2 are committed to their preferred behavior \((p_2>1/2)\), the individuals of population 1 will adjust to this behavior, i.e. population 1

\(^1\) The initial commitment often depends on historical factors. Of course, not only the average commitment to a norm, but also its \textit{content} can change over time, i.e. a norm can change its “character”. In our model, we assume that the character changes much slower than the commitment, so that changes of content can be neglected.
will assimilate to population 2 (see Figs. 1b and 2). Finally, if the fraction of individuals initially showing their preferred behavior is about the same in both populations \((p_1 \approx p_2)\), none of the behaviors will gain the majority. In this case, the majority of individuals ends up doing what they prefer (see Figs. 1c and 2). This is, by the way, the situation, which is always found for \(B>A\).

Figure 1: Path dependence in the establishment of social norms in case of two equally strong, interacting populations \((S=0.5)\). Both preferences are equally distributed in space. (a) If the initial average commitment is the same in both populations \((p_1 = 0.5, p_2 = 0.5)\), members of both populations predominantly do what they prefer. Only a few individuals deviate from their preferences due to interaction effects or random behavioral changes (see squares with dots in the middle). (b) If the initial average commitment to the preferred behavior is high in population 1 \((p_1 = 0.9)\), but small in population 2 \((p_2 = 0.4)\), the preference of population 1 establishes as commonly shared behavior, i.e. as social norm. Only a few individuals deviate, primarily due to random behavioral changes (determined by the rate \(r\) of strategy flips). (c) If the initial average commitment is high in population 2 \((p_2 = 0.9)\), but small in population 1 \((p_1 = 0.4)\), the preferred behavior of population 2 will set the norm. The baseline parameters of all simulations are \(A=1, B=0.5, R = 10\), and \(r=0.01\).

The dependence on the initial condition can be illustrated by a so-called “phase diagram”: Figure 2 reflects the path dependence in the formation of social norms. In addition, it shows that a commonly shared behavior is not formed, if both populations have similar power \((S \approx 1/2)\) and everybody starts with the preferred behavior \((p_1=p_2=1)\) or the average initial commitment in both populations is about the same \((p_1=p_2)\). This, by the way, demonstrates that norms are not self-enforcing in our multi-population norms game without punishment for \(B>0\), even though we have assumed with \(A>B\) that the advantage \(A\) of conforming with the behavior of others is larger than the benefit \(B\) of showing the preferred behavior.

Relevance of power and unfavorable norms. If one population is more powerful than the other, it is natural to expect that the preferred behavior of the more powerful population would win through (if \(B<A\) and \(R=10\), as before). Figure 3 shows simulation results for \(S=0.8\), i.e. 80% of all individuals belong to population 1. In fact, Fig. 3a shows that most individuals of population 2 eventually adapt to the preferred behavior of population 1. This is, by the way, not the case for \(S=0.6\), i.e. when population 1 is only slightly more powerful than population 2. However, it can happen that the behavior preferred by the weaker population 2 finally prevails (see Fig. 3b). For this, the proportion of individuals who initially show their preferred behavior must be significantly higher in population 2 than in population 1 (i.e. \(p_2>>p_1\), see Fig. 3c). In other words, the minority may succeed to establish their preferred behavior as a commonly shared behavior, if they are more committed in the beginning. This offers an
explanation why unfavorable norms, i.e. norms that are not aligned with the preferences of the majority of people, are possible (see Sections 5 and 6 for a further discussion of this case).

Figure 2. So-called phase diagrams showing the finally resulting system behavior (a) as a function of the relative benefit $B/A$ of showing the preferred behavior and the relative strength $S$ of population 1 (for the parameters $R=10$, $r=0.01$), (b) as a function of the initial average commitments $p_1$ and $p_2$ in both populations for the parameters $A=1$, $B=0.5$, $R=10$, $r=0.01$, and $S=0.5$. Three different cases (“phases”) are possible, as indicated in the figure: either, the great majority of people shows the individually preferred behavior, or the behavior preferred by population 1 is established as a norm, or alternatively the behavior of preferred by population 2. The phase diagrams show how the respective result depends (a) on the parameter choices and (b) on the initial conditions.

Figure 3. Due to the path dependence of norm formation, unpopular norms can be established, if the unpopular behavior is initially overrepresented. In our computer simulations, 80% of individuals belong to population 1 and prefer behavior 1 ($S=0.8$). The other model parameters are $A=1$, $B=0.5$, $R=10$, $r=0.01$. (a) If the initial average commitment is the same in both populations ($p_1=p_2=0.5$), the behavior preferred by the majority wins through, as expected. (b) However, if the initial average commitment in the minority population is high ($p_2=0.9$), while it is low in the majority population ($p_1=0.25$), the behavior preferred by the minority establishes as a norm. (c) The dependence of the final outcome on the initial average commitments $p_1$ and $p_2$ can be illustrated by a phase diagram. In this case, the results are for $A=1$, $B=0.5$, $R=10$, $r=0.01$, $S=0.8$.

The possible dominance of a behavior preferred by a minority also implies that the formation of a behavioral norm does not necessarily establish a system optimum, as pointed out by Elster (1989a). This is illustrated more clearly in Figure 4, and it implies that norms are not
necessarily beneficial, or at least not the most beneficial solution for a social or economic system. In fact, norms can be very costly for the majority of people, as the example of female genital mutilation (Mackie 1996) suggest. Nevertheless, it is plausible that norms implement a coordination of behaviors in the interest of some group.

![Figure 4](image)

**Figure 4.** Average payoff of all individuals in the course of time $t$ for the two scenarios illustrated in Fig. 3 (with parameters $A=1$, $B=0.5$, $R=10$, $r=0.01$, $S=0.8$): (a) the behavior preferred by the majority wins through ($p_1=0.5$, $p_2=0.5$), (b) the behavior preferred by the minority establishes the norm ($p_1=0.5$, $p_2=0.5$). Note that, compared to the (initial) coexistence of the two behaviors, the average payoff increases even, if the behavior preferred by the minority wins through.

**Local cultures.** So far, by selecting a large value of the interaction range $R$, we have assumed that every individual could interact with any other individual. In contrast, the following simulations will be performed for $R=1$ or $R=2$. In order to make our results comparable with Fig. 1, we will again assume $A>B$ and equally strong populations ($S=1/2$). As Fig. 5 illustrates, we find similar simulation results in this case as for models describing the formation of local cultures (Axelrod 1997).

The phenomenon of global diversity despite of local conformity is one of the interesting puzzles in the social sciences (Flache 2008; Mäs 2010). It is well-known from the spatial distribution of languages or dialects (Keller 2003), or also of certain kinds of food (“regional specialities”), traditions, or habits. It is remarkable here that we do not need a separate model to derive this observation. It results as a special case of our model for the formation of a behavioral norm, if the interaction range of individuals is small enough.

Note that the regions of uniform behavior are not stable over time. They are ever-changing, i.e. the boundaries between areas with different cultures (i.e. different majority behaviors) are moving in the course of time. Over very long time periods, the final stage of our computer simulations could theoretically be a “monoculture”, where one norm (or set of norms) is shared by everybody all over the world.

It should be underlined that the case of global diversity despite local conformity is to be distinguished from the previously discussed case of behavioral coexistence (“individualism”).
While coexistence appears to occur for large values of $B/A$, local cultures\(^2\) (or, more precisely, local norms) seem to be formed for small values of $B/A$. In the first case, most individuals show their preferred behavior (see Fig. 1c). In the latter case, in contrast, a large number of individuals shows the non-preferred behavior, i.e. they assimilate to the local culture (see Fig. 5), while the predominating behavior changes from one region to another.

![Figure 5. Emergence of local conformity and global diversity, when the interaction range $R$ is small. The displayed snapshots were simulated for two equally strong populations ($S=0.5$) and taken after $t=100$ iterations (a) for $R=1$ and (b) for $R=2$. The other model parameters are $A=1$, $B=0.4$, $S=0.5$, $p_1=p_2=0.5$, $r=0.01$.](image)

![Figure 6. Coexistence of behaviors, when the benefit $B$ of pursuing the individually preferred behavior is larger than the advantage $A$ of conforming with the behavior of the respective interaction partner. The model parameters used in this computer simulation are $A=1$, $B=1.2$, $R=10$, $S=0.5$, $p_1=p_2=0.5$, $r=0.01$, i.e. the only parameter that is different from the ones used in Fig. 1(a) is $B$.](image)

\(^2\) We would like to note that the term “culture” is normally used in a wider sense, representing a whole set of norms (and potentially other human artifacts as well). Nevertheless, a model that creates local norms can be easily extended to consider the interaction of a variety of independent or interrelated norms.
5. A typology of punishment of norm violations

In many cases, the payoff for conforming with others does not exceed the benefit of following the preferred behavior (i.e. \( A < B \)). It seems plausible to expect a co-existence of different norms for this case. This intuition is confirmed by additional simulation analyses (Figure 6). In other words, this payoff structure triggers most people to do what they like so that the coordination on one norm is unlikely. The situation of normative co-existence, however, can be overcome by sanctioning strategies.

We distinguish two different types of punishment: behavior- and preference-based punishment. These two cases can be further subdivided into sincere and hypocritical punishment.

a) Behavior-based punishment assumes that individuals impose a punishment fine on an interaction partner, whose behavior differs from the own behavior.

i. **Sincere** behavior-based punishment describes the case, where the punisher does not prefer and does not show the punished behavior herself.

ii. **Hypocritical** behavior-based punishment describes the case, where the punisher punishes a different behavior, although it actually corresponds to the privately preferred behavior. A typical example is that somebody, who does not dare to follow the own preferences, punishes somebody else for taking the liberty to do so.

There are numerous examples of hypocritical behavior-based punishment, some of which are discussed by Centola et al. (2005):

- **Unintellegible publications**: In certain fields of academia, unintelligible texts are popular because nobody wants the risk to be blamed for trivial science. Those who publicly question the unintelligible scientific work are punished by others who privately disapprove the work as well.

- **Review system and quality management**: The evaluation and review system in science may be subject to hypocritical punishment, as many may dislike it but few oppose it to avoid being regarded as incompetent. Those who oppose it are publicly punished for questioning the review system. A similar thing may apply to the whole evaluation and quality management culture, which has been established in the past years in business, health care, administrations, schools, etc.

- **Sexual freedom and homophobia**: Most individuals dreaming of sexual freedom did not support it publicly or even opposed it before the “summer of love” (Kinsey 1948; Kinsey 1953). Even though this is not an issue anymore in many countries, most people still want to avoid the risk of being blamed as homosexual. Thus, soccer fans, colleagues at work, or school kids with homosexual preferences often engage in jokes against homosexuals.

- **Bullying**: In school-classes, some pupils are targets of bullying and mobbing. Other pupils are afraid to be the next victim. They engage in the punishment of the mobbed victims as well, although they may privately like them or at least do not have objections against them.

b) Preference-based punishment assumes that individuals impose a punishment fine on an interaction partner who behaved differently from the own preference.
i. **Sincere** preference-based punishment describes the case that the punisher does not prefer and does not perform the punished behavior.

ii. **Hypocritical** preference-based punishment describes the case that the punisher sanctions someone for a behavior that does not comply with her own preferences, while actually performing the same behavior. One could describe the situation as one in which the punisher admonishes those who cannot suppress similar weaknesses as she shows herself.

There are also numerous examples for hypocritical preference-based punishment, some of which are discussed by Heckathorn (1990):

- Public punishment of personal weaknesses like promiscuity, smoking, drinking, use of illegal drugs, laziness, untidiness, unpunctuality, or forgetfulness, although the punisher actually has these weaknesses as well. Think of someone who regularly comes too late to meetings, but blames others if they dare to come later to her own meeting.

- A famous example are television evangelists who blamed others for being homosexual or customers of prostitutes, while the same evangelists have later been uncovered to be homosexual or regular customers of prostitutes themselves.

![Figure 7. Illustration of our typology of four different kinds of punishment, distinguishing behavior-based (left) and preference-based punishment (right) on the one hand, and sincere and hypocritical punishment on the other hand.](image)

6. The Multi-Population Norms Game with Punishment

We will now study two different scenarios, assuming that ego (agent 1) punishes alter (agent 2) either if alter shows a different behavior than ego (behavior-based punishment) or if she behaves against ego’s preference (preference-based punishment). Punishment causes two different kinds of costs: There is a punishment cost which is deducted from the payoff of the punisher and a punishment fine, which is deducted from the payoff of the punished individual (punishee). The punishment cost for the punisher and the punishment fine for the punishee matter in so far as they drive the learning dynamics of the model. In particular, a high punishment cost or fine can trigger an individual to change her strategy if other like-minded agents (with the same preference) in her neighborhood reach a higher payoff with the alternative behavior. The punishment is carried out regardless of a cost-benefit analysis, but merely based on the fact whether the respective interaction partner behaved contrary to ego’s behavior or preference. In this sense, our simulation simplifies punishment as an unconditional, normative act, comparable to the notion of the Kantian categorical imperative,
the notion of unconditional normative behavior in the framing literature (Kroneberg, Yaish and Stocke 2010), and to learning models of unconditional punishment of norm violations (Centola, Willer and Macy 2005; Willer, Kuwabara and Macy 2009). It is also consistent with experimental results of Horne (2009), according to which sanctioning can be irrational in the sense that individuals may enforce norms against their own preferences or at unreasonably high punishment costs (larger than the resulting rewards). This is typically an effect of metanorms to punish deviant behavior, and the effect is further amplified by social relations.3

For both types of punishment, the behavior-based and the preference-based one, we assume the investment of a fixed punishment cost \( C \). The punishment fine \( F \) is assumed to be proportional to \( C \), i.e. \( F=kC \). If the punishment effect \( F \) is larger than the punishment effort \( C \), as is usually assumed, the proportionality factor satisfies \( k > 1 \). Moreover, we presuppose that agents do not punish, if their cumulative payoff would become negative.

If \( P_b \) is the payoff for behavior \( b \) and the interaction partner is punished, the focal individual remains with a payoff of \( P_b - C \), and the payoff \( P_c \) of the interaction partner is reduced to \( P_c - F \).

Furthermore, our simulations focus now on the case, where the payoff \( B \) for following the preferred behavior exceeds the payoff \( A \) for conformity, in which both alternative behaviors would coexist without punishment (otherwise, norms could anyway be established). Analytical calculations demonstrate that behavior-based punishment has a similar effect as an increase of the advantage \( A \) of conformity by the average value of \( F + C = (k+1)C \) experienced by the individuals. Hence, we expect that norms can be established, if the punishment level \( L = (k+1)C \) is large enough, which is confirmed by our computer simulations

Figure 8 shows simulation results for \( A < B \) and various values of the punishment level \( L \), assuming that 80% of all individuals belong to population 1 (\( R = 0.8 \)). Figure 8 illustrates clearly that most individuals follow their preferences, if the punishment level \( L \) is small. However, as \( L \) is increased, individuals of the weaker population (i.e. population 2) increasingly conform to the behavior preferred by population 1. For sufficiently large punishment levels \( L \), the great majority of individuals share a common behavior. This demonstrates the formation of a social norm through punishment under conditions, for which everybody would show the personally preferred behavior without negative sanctioning. For high enough punishment levels, there are only a few individuals who show a norm-deviant behavior, and this is just because of the randomness assumed in our model.

It is interesting to see what results are found, when punishment is preference-based, i.e. everybody tries to enforce the own preferred behavior. Performing the same analysis as for behavior-based punishment, we find the results of Fig. 9. Surprisingly, preference-based punishment is much less effective. The fraction of individuals sharing a common behavior increases only very slowly with the punishment level \( L = (k+1)F \). Therefore, behavior-based punishment is expected to be considerably more successful in establishing social norms. This is probably the reason why, despite the many benefits of diversity (Lorge et al. 1958; Page 2007; Surowiecki 2004), group pressure towards conformity is surprisingly common in social systems (Allport 1924; Asch 1956; Bikhchandani, Hirshleifer and Welch 1998; Cialdini and Trost 1998). Interestingly, peer groups often care more about conformity in behavior than about the preferences of their members, and their norms can have almost any content.

---

3 In a certain sense, deviant behavior is considered “harmful”, as it causes extra coordination costs or implies the likelihood of transaction failures.
Figure 8. Proportion of individuals showing behavior 2 as a function of the punishment level $L$, when individuals either apply behavior-based or preference-based punishment. (a) For $S=0.8$, it becomes visible that preference-based punishment is less successful in establishing a commonly shared behavior 1 than behavior-based punishment. (b) Preference-punishment may fail completely, if both populations have a similar strength ($S=0.6$). The other model parameters are $A=1$, $B=1.2$, $R=2$, $p_1=0.5$, $p_2=0.5$, $r=0.01$.

Figure 9. Proportions of sincere and hypocritical punishers as a function of time $t$ in case of (a) behavior-based punishment and (b) preference-based punishment. The model parameters in both simulation scenarios are $A=1$, $B=1.2$, $L=1.5$, $R=2$, $r=0.01$, $S=0.8$, $p_1=0.5$, $p_2=0.5$.

7. Multi-Population Norms Game with Adaptive Group Pressure

One may, of course, ask what would be the best way to specify the punishment level $L$ in our multi-population norms game with sanctioning. It seems natural to assume that it should depend on the relative frequency $N$ of non-conforming behavior within the interaction range $R$. If the frequency of non-conforming behavior is high, individuals would normally not dare to punish a behavior that deviates from the own one. However, if the frequency of non-conforming behavior is small, individuals would be encouraged to punish non-conforming behavior of an interaction partner. This could, for example, be expressed by the following formula: $C = C_0 (1-N)^4[2-(1-N)^2]^2$, where $C_0$ can be called the maximum punishment level (or
the intensity of the metanorm to punish deviant behavior). This function is illustrated by Figure 10. It assumes low punishment levels when the non-conforming behavior is in the majority ($N>1/2$). However, when the conforming behavior prevails within the interaction range $R$ ($N<<1/2$), the punishment level is high. According to this specification, individuals would always have a tendency to apply a certain level of punishment in favor of conformity. The expectation that an interaction partner should conform with the own behavior is assumed to be high only, when it prevails in the interaction neighborhood (i.e. $N<<1/2$).

Figure 10. Adaptive group pressure is introduced in our multi-population norms game by specifying the punishment cost $C = C_0 (1-N)^4[2-(1-N)]^2$, $C_0 = 1$ as a function of the relative frequency $N$ of non-conforming behavior in the interaction neighborhood of range $R$.

Figure 11 shows snapshots of a computer simulations at different times $t$. The simulation studies the interesting case, where most individuals would show their preferred behavior without punishment efforts ($A<B$). Moreover, we assume a maximum punishment level $C_0$ that is big enough to reinforce a social norm when everybody initially shows the same behavior. However, it is chosen so small that it does not support the formation of a social norm when different behaviors are uniformly distributed, as it is assumed in the beginning of our computer simulation ($t = 0$). Hence, for a long time, most individuals show their preferred behavior. Then, however, we discover a very surprising change in the system behavior. In some areas, local norms are suddenly showing up. These local norms eventually spread in the system, and global diversity finally disappears. In conclusion, under conditions where different behaviors would usually coexist, strong enough group pressure will sooner or later establish a norm of one content or another, which is shared by a great majority.
Figure 11. Adaptive group pressure causing the spontaneous birth and spreading of a social norm after many iterations: (a) $t=205$, (b) $t=250$, (c) $t=1000$. The model parameters in both simulation scenarios are $A=1$, $B=1.1$, $C_0=1$, $R=2$, $k=3, r=0.01$, $S=0.5$, $p_1=p_2=1$.

The mechanism underlying this social dynamics is as follows: Due to a certain degree of randomness in the selection of individual behaviors (and interaction partners), it can occur by coincidence that one of the behaviors happens to gain a majority in a certain neighborhood. The local “balance of power” between the two competing behaviors can be broken, when a sufficient number of “strategy flips” occur at the same time and creates a random bias for one of the behaviors. In other words, fluctuations in the individual behaviors have to add up in a way that considerably deviates from the mean value. As deviations of such size are rare, it requires a considerable time for them to occur by coincidence. However, once a large enough deviation has appeared, the local majority increases its punishment efforts, i.e. the group pressure is intensified according to Fig. 10 to support the behavior of the local majority. This further increases the local conformity, i.e. the fraction of individuals showing the same behavior. When the local level of conformity is high enough, it can spread to neighboring areas. Therefore, areas following the same norm are growing. The final competition between areas with different norms depends on the local majorities and random events. Assuming equally strong populations, behavior 1 finally establishes the norm throughout the whole system with a probability of 50%, and behavior 2 wins through in the other 50%. That is, for the assumed parameter values, there will always be a shared behavioral norm in the end. In our model with adaptive group pressure, individuals are only expected to stick forever to their personally preferred behavior if the maximum punishment level $C_0$ is sufficiently small.

Persistence of once established norms. The previous section suggests that, for social interactions with adaptive group pressure, we would always find the formation of a social norm, whatever its content may be (given that $C_0$ is high enough). Once a certain behavior has gained the majority, it would be reinforced and maintained over extremely long time periods. That is, a social system would normally not switch from one norm to another, even if the norm is “outdated” and does not serve a purpose anymore (if it ever did). The only way of overcoming a once established norm in our model is the reduction of the maximum punishment level $C_0$. Figure 12 illustrates such a scenario. It can be seen that an established norm persists over a long time, even if the punishment efforts are suddenly reduced. Despite setting the punishment strength $C_0$ to a small value at some point $t_1$ in time, this causes a rather gradual decay of the support for the previous norm in favor of pursuing the individually preferred behaviors.\footnote{Alternatively, an outdated norm may be gradually transformed into another one, but this is outside of the scope of our model.} Note that the persistence of social norms in our model is a consequence of the norm-creating mechanism. It does not necessarily require an internalization of normative contents (Epstein 2001), but rather the internalization of “normative thinking”, i.e.
a metanorm of applying group pressure against behavior that deviates from the behavior of the majority.

**Figure 12. Persistence of a social norm, when sanctioning efforts are stopped at some point** $t_0$ **in time.** Before this time, behavior 1 is established as social norm. The model parameters used in the computer simulation are $A=1, B=1.1, R=2, r=0.01, S=0.6, p_1=p_2=1, k=3$ and $C_0=1$ until $t_0=1000$, afterwards $C_0=0.25$.

8. Discussion and Outlook

In this contribution, we have proposed a model for the formation of norms in situations in which people have different preferences. The model has been illustrated by the case of two alternative behaviors and no preset norm. Individuals tend to do what they prefer, but rewards or punishment may encourage conformity with the behavior of others. Interestingly, even when each of the alternative behaviors is preferred by the same number of individuals, interaction effects may cause behavioral consensus in favor of one behavior or the other, which corresponds to the establishment of a commonly shared norm. However, for given model parameters, different outcomes may result, including the coexistence of different behaviors. This reflects the path dependence of the social dynamics simulated by our computer model.

We have compared two different scenarios in order to find out, what happens when people punish others. In the first scenario, individuals punish others for behaviors that do not conform with their own displayed behavior (behavior-based punishment), while in the second scenario, they punish behaviors that do not conform with their preferred behavior (preference-based punishment). While the former could promote the formation of social norms, it is surprising that preference-based punishment does not seem to be effective for the establishment of a commonly shared norm. This implies that behavior-based consensus seems harder to reach than value-based consensus, which is relevant for the course that political states may take.

It appears natural to assume that group pressure towards conformity increases when many neighbors share the same behavior. When adaptive group pressure is considered, social norms
may spontaneously be created by random variations in individual behaviors. Starting off with a situation in which everybody shows the personally preferred behavior, local behavioral majorities will sooner or later occur by coincidence. This gives rise to local conformity, i.e. local norms. These are then spreading into areas where everybody was performing the personally preferred behavior before. After a period of global diversity, the system would finally end up with a globally shared norm, but the process of global convergence is very slow, so that the content of the norm is likely to change, creating new variants of norms.

Our model contains a few model parameters only. The relevant ones are the intrinsic benefit $B$ of showing the preferred behavior, the advantage $A$ to conform with the behavior of an interaction partner (due to externalities or rewards), the level $r$ of randomness in behavioral choice processes, the relative size $S$ or, more generally, the relative power of a population of individuals sharing the same preferences, and the punishment cost $C$ (or the maximum intensity of the group pressure, $C_0$). For the case that there are no unilateral incentives to deviate from a commonly shared behavior ($B=0$), our model describes the formation of coordination norms or “behavioral conventions” (Helbing 1992; Young 1993; Young 1996). Finally, note that the resulting outcome in the simulated social system also depends on the proportions $p_1$ and $p_2$ of individuals showing their preferred behavior at time $t = 0$, i.e. on the average initial commitment to the preferred behavior in each population. This establishes the history- or path-dependence of norm formation.

We consider the simplicity of the model as a great advantage, because it allows one to study the effects of punishment rules, the conditions of norm formation, the dependencies between relevant variables and parameters and the dynamical patterns resulting in different scenarios. The model even appears to be accessible to an analytical treatment (Helbing and Johansson 2010). Despite the simplicity of our model, it seems to capture many stylized facts of social norms, such as path-dependence, the possibility of unpopular or dysfunctional norms, the occurrence of local cultures, their spatial segregation and spreading, or their persistence when the punishment of non-conforming behaviors is taken back. Nonetheless, it is possible to extend our model in a number of ways, which will be addressed by forth-coming studies.

References


