OPEN ACCESS Freely available online

PLoS one

Cooperation, Norms, and Revolutions: A Unified Game-Theoretical Approach

Dirk Helbing[1,2,3]*, **Anders Johansson**[1,4]

1 ETH Zurich, Zurich, Switzerland, **2** Santa Fe Institute, Santa Fe, New Mexico, United States of America, **3** Collegium Budapest-Institute for Advanced Study, Budapest, Hungary, **4** Centre for Advanced Spatial Analysis, University College London, London, United Kingdom

Abstract

Background: Cooperation is of utmost importance to society as a whole, but is often challenged by individual self-interests. While game theory has studied this problem extensively, there is little work on interactions within and across groups with different preferences or beliefs. Yet, people from different social or cultural backgrounds often meet and interact. This can yield conflict, since behavior that is considered cooperative by one population might be perceived as non-cooperative from the viewpoint of another.

Methodology and Principal Findings: To understand the dynamics and outcome of the competitive interactions within and between groups, we study game-dynamical replicator equations for multiple populations with incompatible interests and different power (be this due to different population sizes, material resources, social capital, or other factors). These equations allow us to address various important questions: For example, can cooperation in the prisoner's dilemma be promoted, when two interacting groups have different preferences? Under what conditions can costly punishment, or other mechanisms, foster the evolution of norms? When does cooperation fail, leading to antagonistic behavior, conflict, or even revolutions? And what incentives are needed to reach peaceful agreements between groups with conflicting interests?

Conclusions and Significance: Our detailed quantitative analysis reveals a large variety of interesting results, which are relevant for society, law and economics, and have implications for the evolution of language and culture as well.

Citation: Helbing D, Johansson A (2010) Cooperation, Norms, and Revolutions: A Unified Game-Theoretical Approach. PLoS ONE 5(10): e12530. doi:10.1371/journal.pone.0012530

Editor: Enrico Scalas, University of East Piedmont, Italy

Received June 7, 2010; **Accepted** August 2, 2010; **Published** October 12, 2010

Funding: Partial support was provided by the Future and Emerging Technologies programme FP7-COSI-ICT of the European Commission through the project QLectives (grant no.: 231200) and the ETH Competence Center "Coping with Crises in Complex Socio-Economic Systems" (CCSS) through ETH Research Grant CH1-01 08-2 (Eidgenoessische Technische Hochschule/Swiss Federal Institute of Technology). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: dhelbing@ethz.ch

Introduction

In order to gain a better understanding of factors preventing or promoting cooperation among humans or other species, biologists, economists, social scientists, mathematicians and physicists have intensively studied game theoretical problems such as the prisoner's dilemma and the snowdrift game (also known as chicken or hawk-dove game) [1–5]. In all these games, a certain fraction of people or even everyone is expected to behave uncooperatively (see Fig. 1). Therefore, a large amount of research has focused on how cooperation can be supported by mechanisms such as

- repeated interactions [1,6–8],
- reputation [9–12],
- clusters of cooperative individuals [13,14],
- sanctioning [15–21],
- success-driven migration [22], or
- economic incentives [23].

For a discussion and classification of cooperation-promoting mechanisms within an evolutionary game-theoretical framework see Refs. [3,24,25].

Many game-theoretical studies of social cooperation are based on models, in which all individuals are assumed to have the same properties. In reality, however, individuals are different. To investigate the relevance of this for the resulting outcome and dynamics of social interactions, we will consider that people of different gender, status, age, or cultural background may have heterogeneous preferences (e.g. due to framing effects, see [27–30]. We will focus here on the case where the preferences are not only *gradually* different, but where we have two interacting populations with mutually incompatible preferences, which cannot be satisfied at the same time. For example, men and women appear to have incompatible interests many times. Nevertheless, they normally interact among and between each other on a daily basis. It is also more and more common that people with different religious beliefs live and work together, while their religions request some mutually incompatible behaviors (in terms of the working days and free days, the food one may eat or should avoid, the headgear, or appropriate clothing, etc.). A similar situation applies, when people with different mother tongues meet or businessmen from countries with different business practices make a deal. In this contribution, we are interested in identifying factors, which determine whether two such populations go their own way, find a common

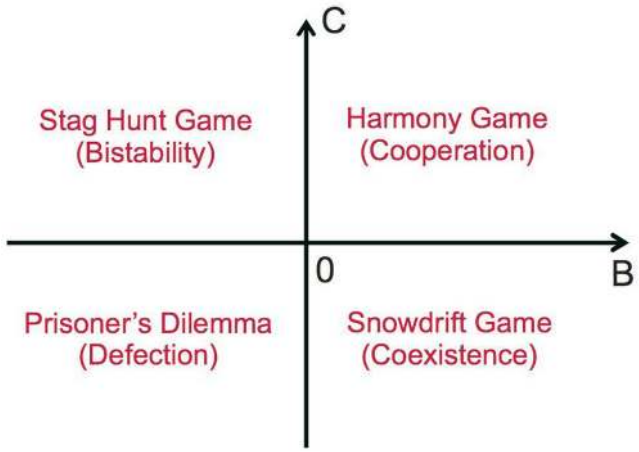

Figure 1. Illustration of the types and outcomes of a game-dynamical treatment of *one*-population symmetrical 2×2 games as a function of the two model parameters *B* and *C* (see, e.g., [26], pp. 28ff). *B* and *C* are related to the four payoffs *P*, *R*, *S*, and *T* of symmetrical 2×2 games via the relations $B=S-P$ and $C=R-T$, see Sec. *Methods* for details. The payoff is *R*, if two interacting individuals show behavior 1, but *P*, if both of them show behavior 2. When the interaction partners show different behaviors, the one showing behavior 1 receives the payoff *S* and the other one the payoff *T*. In the prisoner's dilemmas (PD), behavior 1 corresponds to cooperation and behavior 2 to defection. *R* is then called the "reward" for mutual cooperation, *P* the "punishment" for mutual defection, *T* the "temptation" for unilateral defection, and *S* the "sucker's payoff" for unilateral cooperation. As the prisoner's dilemma is characterized by the inequalities $T>R>P>S$, we have $B<0$ and $C<0$, and defection ("free-riding", "cheating") is the dominant strategy. For snowdrift games (SD) with $T>R>S>P$, which are also called chicken or hawk-dove games, we have $B>0$ and $C<0$, and uncooperative behavior is tempting. The stable stationary solution corresponds to a coexistence of a fraction $p_0=|B|/(|B|+|C|)$ of cooperators with a fraction $1-p_0$ of defectors (uncooperative individuals). For harmony games (HG) with $R>T>S>P$, we have $B>0$ and $C>0$, and everybody will eventually cooperate. Finally, for stag hunt games (SH) with $R>T>P>S$, which are also called assurance games, we have $B<0$ and $C>0$. Here, cooperation is uncertain, as the situation is bistable: If the initial fraction p_0 of cooperators is larger than $p_0=|B|/(|B|+|C|)$, everybody is expected to cooperate in the end, otherwise everybody will eventually behave uncooperatively.
doi:10.1371/journal.pone.0012530.g001

agreement, or end up in conflict [31]. Moreover, we want to understand the relevance of power in the rivalry of populations [32].

Our treatment of heterogeneous preferences is based on multi-population games [33–37]. The simpler case where individuals of two *different* populations interact with each other, while they do not interact when belonging to the *same* population, has been nicely summarized by [26], pp. 182ff. An earlier publication on the self-regulation of behavior in animal societies also considers self-interactions within each population [38]. In fact, most applications of multi-population models in evolutionary game theory so far were oriented at the interaction of *biological* species and the study of *ecosystems* [39–43]. Compared to these publications, our treatment focuses on *social* systems, and effects of differences in the power of interacting populations are considered. The problem of conflicts between social groups has been studied before by bimatrix games [44] such as the "battle of sexes" (see, e.g., [35] and by so-called hypergames (see, e.g., [45]. However, the related models appear to be less versatile than the one proposed in the following. The main difference to previous approaches is that we study *social* interactions *between* and *within* populations, considering that the power of the involved populations may be different. This generates interesting kinds of system dynamics, which do not appear when self-interactions (between individuals of the same population) are neglected or if all populations are equally strong. For example, we find that it may not only depend on the payoffs, but also on the initial condition, whether the individuals of two populations with incompatible preferences finally show a commonly shared behavior (see Sec. *Evolution of normative behavior in the stag hunt game*).

Note that this paper presents more than "just another model". First of all, our approach fits particularly well into widely established modeling concepts. Second, it bridges between two different modeling worlds by unifying features of game-theoretical and opinion dynamics models (see Sec. *Discussion of previous literature on norms*). Third, the model is analytically tractable [46]. Fourth, it contains very few parameters, while it describes a variety of different system behaviors (although it was not explicitly constructed for this). Despite 3 parameters only (which can be further reduced, since only the signs and quotient of the parameters *B* and *C* matter), the model shows a surprisingly rich behavior and can reproduce a variety of phenomena observed in social systems: (1) the breakdown of cooperation, (2) the coexistence of different behaviors (the establishment of "subcultures"), (3) the evolution of commonly shared behaviors ("social norms"), and (4) the occurrence of social polarization, conflict, or revolutions. The approach can also be cast into an agent-based model. In fact, agent-based models for the establishment of norms have found a lot of interest, recently [47–60].

Modeling Approach

The crucial point of our modeling approach is to adapt the game-dynamical replicator equations for multiple populations [35,38,46,61] in a way that reflects interactions between individuals with incompatible preferences (see Sec. *Methods*). The resulting equations (1) and (2) describe the time evolution of the proportions $p(t)$ and $q(t)$ of cooperative individuals in populations 1 and 2, respectively, as individuals imitate more successful behaviors in their own population. Their success depends on the "payoffs" quantifying the results of social interactions, i.e., on the own behavior and the behavior of the interaction partner(s).

In order to reflect incompatible interests of both populations, we assume that population 1 prefers behavior 1, while population 2 prefers behavior 2. If an interaction partner shows the behavior preferred by oneself, we call this behavior "cooperative", otherwise uncooperative. Accordingly, behavior 1 is cooperative from the viewpoint of population 1, but uncooperative from the viewpoint of population 2 (and vice versa). Furthermore, if an individual of population 1 interacts with an individual of population 2 and both display the *same* behavior, we call this behavior "coordinated". Finally, if the great *majority* of individuals in *both* populations shows a coordinated behavior (in case of a commonly shared behavior), we speak of "normative behavior" or a "behavioral norm". To establish a social norm, one of the populations has to give up its preferred behavior.

What will be the resulting dynamics and outcome of such interactions? Under what conditions will we find "normative behavior" (although neither the relative sizes of both populations nor their incompatible preferences change in our model)? To answer these questions, the payoffs from social interactions in 2×2 games are represented by *T*, *R*, *P*, and *S*, as usual (see Sec. *Methods* for details). In the prisoner's dilemma, the meaning of these parameters is "Temptation" to behave non-cooperatively, "Reward" for mutual cooperation, "Punishment" for mutual non-cooperative behavior and "Sucker's payoff" for a cooperative

individual meeting an uncooperative one (see Fig. 1). The related game-dynamical replicator equations contain two payoff-dependent parameters, $B=S-P$ and $C=R-T$. C may be interpreted as gain by coordinating on one's own preferred behavior (if greater than zero, otherwise as loss). B reflects the gain when giving up coordinated, but non-preferred behavior. Equations (1) and (2) contain a further parameter f, which can be interpreted as "(relative) power" of population 1 (while $1-f$ would correspond to the relative power of population 2). The relative power may represent the relative size of the populations, but also differences in their material resources (money, weapons, etc.), social capital (status, social influence, etc.), and other factors (charisma, moral persuasion, etc.). It reflects how much influence a population has on the behavioral choice of individuals. When a population has a greater relative power than another one, we call it "stronger", if it has less power, we call it "weaker". Details of the model and some generalizations are provided in *Methods*.

Results

We have solved Eqs. (1) and (2) by numerical simulation for different parameter values B, C, and f and different initial conditions $p(0)$ and $q(0)$. In contrast to the computer-based analysis presented here, a mathematical analysis of the stationary (i.e. time-invariant) solutions and their stability properties has been carried out in a complementary paper [46]. While the linear stability analysis reveals the sensitivity to stochastic fluctuations (random effects, "noise"), the sensitivity to parameter variations is captured by so-called phase diagrams. Here, we will not go into these technicalities, but rather discuss representative examples of the different *kinds* of system dynamics and their relevance for social systems.

We find that social interactions with incompatible interests do not necessarily produce conflict—they may even promote mutual coordination. Depending on the signs of B and C, which determine the character of the game, we have four archetypical situations:

1. In games like the *multi-population prisoner's dilemma (MPD), we have* $B<0$ and $C<0$.
2. In the *multi-population harmony game (MHG)*, we have $B>0$ and $C>0$.
3. $B<0$ and $C>0$ applies to games like the *multi-population stag hunt game (MSH)*.
4. The *multi-population snowdrift game (MSD)* is characterized by $B>0$ and $C<0$.

In a multi-population prisoner's dilemma with incompatible preferences (MPD), everybody behaves non-cooperatively in the end (see Fig. 2). This does not even change, if one population is stronger than the other one (i.e. $f \neq 1/2$, or if the interaction rate *between* populations is different from the interaction rate *within* populations. This disappointing outcome results despite the fact that non-cooperative behavior in one population corresponds to cooperative one from the perspective of the other. However, as non-cooperative individuals earn a high payoff (the temptation T), when meeting a non-cooperative individual of the other population (it is cooperative from the own perspective), there is no incentive to give up defection in their own population—on the contrary.

In contrast, in the multi-population harmony game, everybody finally shows a cooperative behavior in the *own* population, but due to the different preferences, the behaviors in both populations are not coordinated. (Every population just does what it likes, as if both populations had their own "subcultures", see Fig. 2).

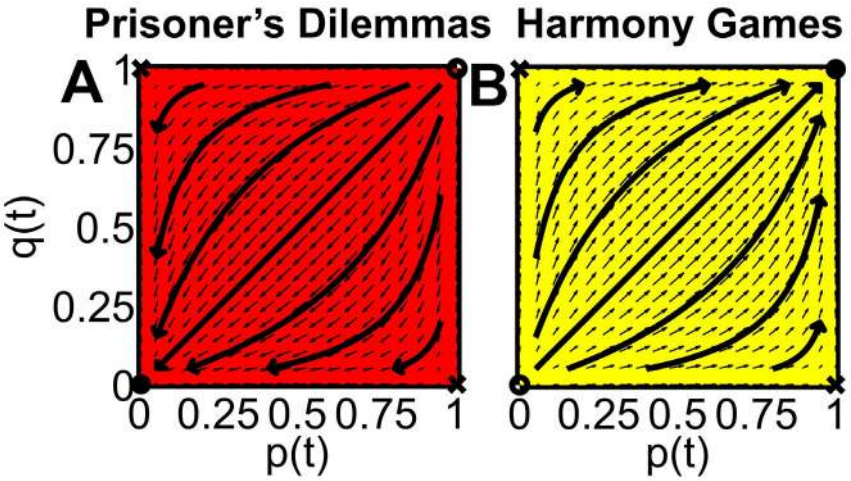

Figure 2. Simulation results for two interacting populations with self-interactions and incompatible preferences playing prisoner's dilemmas or harmony games. The outcome is visualized by vector fields (small arrows) and sample trajectories (large arrows) for $f=0.8$ (i.e. population 1 is assumed to be stronger than population 2). p denotes the fraction of individuals in population 1 showing their preferred, cooperative behavior 1, and q is the fraction of cooperative individuals in population 2 showing their preferred behavior 2. A fraction $1-q$ of individuals in population 2 shows the non-preferred behavior 1, and a fraction $1-p$ of individuals in population 1 shows behavior 2. The vector fields show (dp/dt,dq/dt), i.e. the direction and size of the expected temporal change of the behavioral distribution, if the fractions of cooperative individuals in populations 1 and 2 are $p(t)$ and $q(t)$. Sample trajectories illustrate some representative flow lines ($p(t)$,$q(t)$) as time t passes. The flow lines move away from unstable stationary points (empty circles) and are attracted towards stable stationary points (black circles). Saddle points (crosses) are attractive in one direction, but repulsive in another. The colored areas represent the "basins of attraction", i.e. all initial conditions ($p(0)$,$q(0)$) leading to the same stable fix point [red = (0,0), yellow = (1,1)]. Intuitively, the initial conditions quantify the influence of the previous history. (A) If $B=C=-1$, the individuals in each population are facing prisoner's dilemma interactions and end up with non-cooperative behavior. (B) If $B=C=1$, individuals in each population are playing a harmony game instead, and everybody eventually behaves cooperatively. The results look similar when the same two-population games are played with different values of f, $|B|$ or $|C|$.
doi:10.1371/journal.pone.0012530.g002

As we will show in the following, the dynamics and outcome of the multi- population stag hunt and snowdrift games with incompatible preferences are more complicated and in marked contrast to the corresponding one-population games. This can be demonstrated by systematically exploring the parameter space with computer simulations. In the following, we will illustrate typical simulation results by figures and movies showing the stationary solutions (fix points, evolutionary equilibria) of the games, their basins of attraction, and representative flow lines. Details are discussed below and in the captions of Figs. 2–5 (see also Movies S1, S2, and S3 and *Methods*).

Evolution of normative behavior in the stag hunt game

The *one-population* stag hunt game is characterized by an equilibrium selection problem [62]: *Everyone* is finally expected to cooperate, if the initial fraction of cooperative individuals is above $p_0 = |B|/(|B|+|C|)$, otherwise *nobody* will behave cooperatively in the end (see Fig. 1). The same applies to *non-interacting* populations (see Movie S1). For *interacting* populations without self-interactions, however, it *never* happens in the multi-population stag-hunt game with incompatible preferences that everybody or nobody cooperates in both populations (otherwise there should be yellow or red areas in the second part of Movie S2). Although both populations prefer *different* behaviors, all individuals end up coordinating themselves on a commonly shared behavior (corresponding to the blue and green areas in Movie S2). This can be interpreted as self-organized evolution of a social norm (see below).

Note that the previously discussed cases, which neglect interactions *between* populations or *within* populations, are applicable to *particular* social systems only. Normally, however, there are interactions between *different* populations and, at the same time, interactions between individuals of the *same* population (''self-interactions''). If this is taken into account, the case where everybody or nobody cooperates in both populations is still possible, but it requires that both populations have similar power ($f \approx 1/2$) and that the initial levels of cooperation, $p(0)$ and $q(0)$, are comparable as well. Under such conditions, both populations may develop separate, coexisting norms (see yellow area in Fig. 3B and

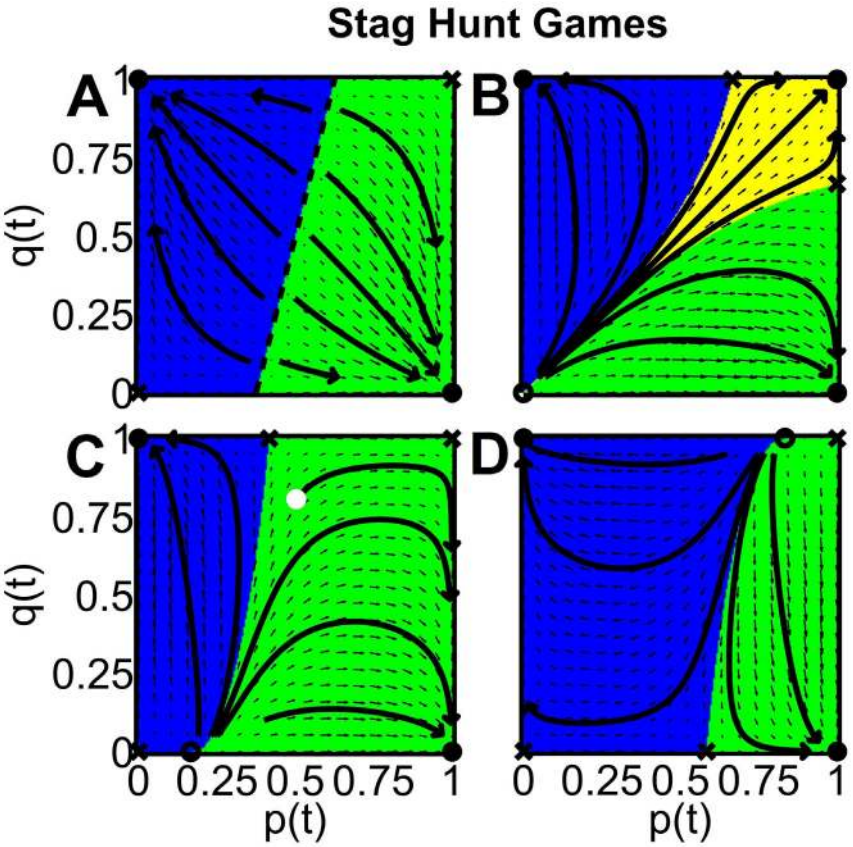

Figure 3. Simulation results for two interacting populations with self-interactions and incompatible preferences, playing stag hunt games. The corresponding vector fields (small arrows), sample trajectories (large arrows) and phase diagrams (colored areas) were determined for $B<0$ and $C>0$. The representation is the same as in Fig. 2. In particular, the colored areas represent the basins of attraction, i.e. all initial conditions $(p(0),q(0))$ leading to the same stable fix point (stationary solutions) [yellow = (1,1), blue = (0,1), green = (1,0)]. The dashed diagonal line represents an infinite number of unstable fix points. The model parameters are as follows: (A) $|B| = |C| = 1$ and $f = 0.8$, i.e. population 1 is more powerful than population 2, (B) $|C| = 2|B| = 2$ and $f = 1/2$, i.e. both populations are equally strong, (C) $|C| = 2|B| = 2$ and $f = 0.8$, (D) $2|C| = |B| = 2$ and $f = 0.8$. Due to the asymptotically stable fix points at (1,0) and (0,1), all individuals of both populations finally show the behavior preferred in population 1, when starting in the green area, or the behavior preferred in population 2, when starting in the blue area. This case can be considered to describe the evolution of a shared behavioral norm. Only for similarly strong populations ($f \approx 1/2$) and similar initial fractions $p(0)$ and $q(0)$ of cooperators in both populations (yellow area), both populations will end up with population-specific norms (''subcultures''), corresponding to the asymptotically stable point at (1,1). The route towards the establishment of a shared norm may be quite unexpected, as the flow line starting with the white circle shows: The fraction $q(t)$ of individuals in population 2 who are uncooperative from the viewpoint of population 1 may grow in the beginning, but later on go down dramatically. Therefore, a momentary trend does not allow one to easily predict the final outcome of the struggle between two interest groups.
doi:10.1371/journal.pone.0012530.g003

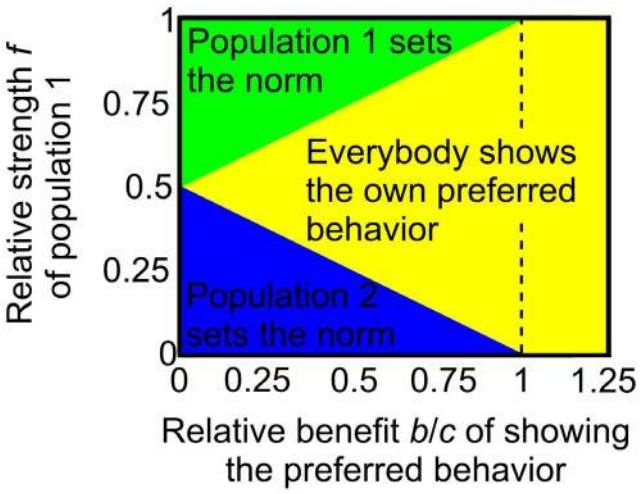

Figure 4. Illustration of the finally resulting system state for the two-population stag hunt game with interactions and self-interactions, and incompatible preferences. The outcome is displayed as a function of the relative strength f of population 1 and the ratio b/c between the benefit b of showing the preferred behavior and the benefit c of showing coordinated behavior. If $b>c$, individuals always show the behavior they prefer. However, if $b\leq c$, a commonly shared behavior (i.e. a "social norm") may result, depending on the respective initial conditions. The initial commitments $p(0)$ and $q(0)$ to the respectively preferred behavior determine also, who sets the norm. If all individuals initially show the behavior they prefer (i.e. $p(0)=q(0)=1$, as assumed in the above illustration), the stronger population sets the norm, given the difference between both population strengths, f and $(1-f)$, is larger than a certain threshold. Otherwise, individuals in each population show their preferred behavior. The threshold depends on the relative strength f and the size of b/c, i.e. the fraction of the relevant payoffs b and c. It can, in principle, be determined from analytical results derived in Ref. [46].
doi:10.1371/journal.pone.0012530.g004

Movie S3), as for the multi-population harmony game. Normally, however, both populations establish a commonly shared norm and either end up with behavior 1 (green area in Fig. 3) or with behavior 2 (blue area).

The behavior of Eqs. (1) and (2) becomes better understandable for multi-population games with the payoffs $P=c$, $R=b+c$, $S=b$, and $T=0$, where the payoff b reflects the benefit of showing the *preferred* behavior, while c is the payoff for showing *coordinated* behavior (reflecting the reward for conforming with the behavior of the interaction partner). While these payoffs do not correspond to a stag hunt game, for $c>0$ and $-c<b<c$ they *also* lead to $B=b-c<0$ and $C=b+c>0$, which implies exactly the same solutions. One advantage of this formulation besides the better interpretation is the possibility to extend it to simultaneous interactions with several players.

For the sake of illustration of this specification, assume that individuals of population 1 like to be properly dressed at the beach and individuals of population 2 enjoy to be naked ($b>0$), but even more than doing what they like, everybody prefers to conform with the behavior of the interaction partners ($c>b$). In situations, when naked people and those wearing a swimming suit interact with each other at the same part of the beach, our equations allow one to identify conditions under which one population eventually sets the behavioral standards or under which a mixture of both behaviors persists. (In contrast to the related multi-population game *without* self-interactions, see Movie S2, there is a possibility that the two populations do *not* coordinate their behaviors, and everybody ends up doing what he or she likes.) Note that, if both populations interact in *space*, dressed people and nudists may segregate, even when there was no disapproval between both behaviors. As a consequence, there may be different "(sub-) cultures" in different parts of the beach, as is often observed. This becomes understandable by the circumstance that the effect of mutual *disapproval* of the non-preferred behavior (which may be described by the payoffs $P=0$, $R=b$, $S=b-c$, and $T=-c$) *again* leads to $B=b-c<0$ and $C=b+c>0$. Therefore, the same kind of dynamics results as in the case where there is a tendency to conform with others.

In conclusion, due to the payoff structure of the multi-population stag hunt game and other multi-population games with $B<0$ and $C>0$, it can be profitable to coordinate oneself with the prevailing behavior in the other population. Yet, the establishment of a norm requires the individuals of one population to give up their *own* preferred behavior in favor of the one preferred by the *other* population. Therefore, it is striking that the preferred behavior of the *weaker* population can actually win through and finally establish the norm (see blue areas in Figs. 3A,C,D). *Who* adapts to the preferred strategy of the other population essentially depends on the *initial* fractions of behaviors (and, thereby, on the previous history). The majority behavior in the beginning is likely to determine the resulting behavioral norm, but a powerful population is in a favorable position: The area of possible histories leading to an establishment of the norm preferred by population 1 tends to increase with power f (compare the size of the green areas in Figs. 3B+C).

Discussion of the equilibrium selection problem

As was indicated already, when two populations with incompatible preferences interact among and between each other, the behavior of the stag hunt game changes completely: Then, the values of the payoff-dependent parameters B and C have an influence on the stable stationary solutions, and inner stationary points (p,q) with $0<p,q<1$ disappear, if $|B|\neq|C|$ [46].

It is noteworthy that, *without* interactions between populations, the stag hunt game implies an interesting equilibrium selection problem [6,100–102], since it has *several* stable solutions. These are classified as payoff-dominant solution (which maximizes the individual payoff, if the interaction partner decides in the same way) and risk-dominant solution (which "minimizes the maximum damage", i.e. maximizes the individual payoff for the worst-case choice of the interaction partner and corresponds to non-cooperative behavior). Which of these solutions is selected depends on the initial conditions, and there is a monotonous increase or decrease of the fraction $p(t)$ of cooperative individuals in the course of time t. In the one-population stag hunt game, the payoff-dominant solution corresponds to cooperative behavior by everybody ($\lim_{t\to\infty} p(t)=1$). It is selected, if the initial fraction $p(0)$ of cooperative individuals is above the value $p_0=|B|/(|B|+|C|)$ of the unstable stationary solution. Instead, the risk-dominant solution results for $p(0)<p_0$ and corresponds to non-cooperative behavior by everyone ($\lim_{t\to\infty} p(t)=0$).

In the multi-population games with interactions *across* populations, the risk-dominance concept is not sufficient to understand the dynamics and outcome of the game. For example, when two populations with incompatible preferences play stag hunt games *without* self-interactions, the game is of the "battle of sexes" type, and there are no thresholds that would separate payoff-dominant from risk-dominant solutions [35,46]. When both, interactions *and* self-interactions are considered, the inner stationary point disappears whenever $|B|\neq|C|$. The unstable solution is rather located at the boundary or in one of the corners. Moreover, as Figs. 3C and 3D illustrate, the incompatibility of preferences can

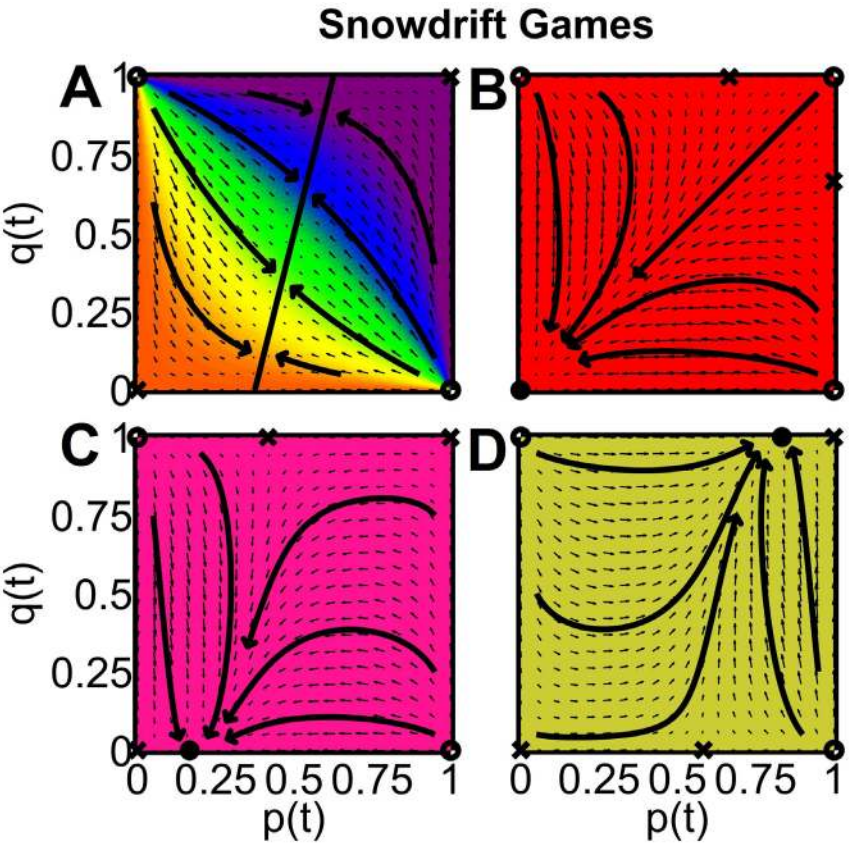

Figure 5. Two interacting populations with self-interactions and incompatible preferences, playing snowdrift games. The corresponding vector fields (small arrows), sample trajectories (large arrows) and phase diagrams (colored areas) were determined for $B>0$ and $C<0$. The flow lines move away from unstable stationary points (empty circles) and are attracted towards stable stationary points (black circles and solid diagonal line). Saddle points (crosses) are attractive in one direction, but repulsive in another. The representation is the same as in Fig. 2. In particular, the colored areas represent the basins of attraction, i.e. all initial conditions ($p(0)$,$q(0)$) leading to the same stable fix point [red = (0,0), salmon = (u,0), mustard = (v,1), rainbow colors = (u,v), with $0<u$, $v<1$]. The model parameters are as follows: (A) $|B|=|C|=1$ and $f=0.8$, i.e. population 1 is more powerful than population 2, (B) $|C|=2|B|=2$ and $f=1/2$, i.e. both populations are equally strong, (C) $|C|=2|B|=2$ and $f=0.8$, (D) $2|C|=|B|=2$ and $f=0.8$. (A) In the multi-population snowdrift game (MSD), a mixture of cooperative and uncooperative behaviors results in both populations, if $|B|=|C|$. (B) For $|B|<|C|$ and equally strong populations, everybody ends up with non-cooperative behavior in each population. (C) For $|B|<|C|$ and $f-1/2\gg 0$, the weaker population 2 forms a ''tacit alliance'' with the minority of the stronger population 1 and opposes its majority. (D) Same as (C), but now, all individuals in the weaker population 2 show their own preferred behavior after the occurrence of a ''revolutionary'' transition, during which the stable stationary solution (the evolutionary equilibrium) changes discontinuously from (u,0) to (v,1).
doi:10.1371/journal.pone.0012530.g005

destabilize the solutions (0,0) or (1,1). Even an initial increase in the fractions $p(t)$ and $q(t)$ of cooperative individuals (i.e. $p'(0)>0$ and $q'(0)>0$, where $p'(t)=dp(t)/dt$ and $q'(t)=dq(t)/dt$) does not imply that the system will end up in the stationary solution (1,1). While there still exists a payoff- and a risk-dominant solution for the stronger population, there is no threshold behavior for the weaker population, as Figs. 3B+C show.

In the limiting case where the relative size f of population 1 goes to one, the resulting fraction of cooperative individuals in population 2 is completely determined by the initial fraction $p(0)$ of cooperative individuals in population 1, while the initial fraction $q(0)$ of cooperative individuals in population 2 does not have any influence. No matter whether population 1 selects the payoff-dominant solution (for large enough values of $p(0)$) or the risk-dominant solution (for small values of $p(0)$), the behavior in population 2 is always coordinated with population 1 in the end [46]. The risk-dominant case can be interpreted such that population 2 effectively manages to set the norm. Figure 4 shows the parameter dependence for the general case, using the alternative parametrization $B=b-c$ and $C=b+c$, where b is the benefit of showing the preferred behavior and c is the benefit of showing coordinated behavior. The classical coordination game, where individuals *always* form a behavioral convention (i.e. coordinate on a behavioral norm) results for $b=0$ [61].

Discussion of previous literature on norms

Note that the subject of social norms is a multi-faceted research field, and there is no single, generally accepted definition of what norms are [63–66]. Definitions range from the concept of ''oughtness'' [67] to a behavioral regularity or shared behavior with a sanctioning (''punishment'') of non-conforming behavior [20,68–77]. However, not all authors agree on the necessity of the sanctioning element [78,79]. In our manuscript, we define ''normative behavior'' or a ''behavioral norm'' as a situation in which behaviors are shared among a large majority of individuals.

The question of how behavioral consensus may evolve has been addressed by opinion dynamics models such as voter models [55] or models of in- and out-group interactions [59,80]. The currently most common approach to behavioral norms is based on game theory and relates the issue to ultimatum games [62,65,78], stag

hunt games [81,82], prisoner's dilemmas [15,47,69,83,84], or related concepts [58]. For overviews see [63,83,85]. Yet, the majority of these models investigate conditions under which people comply with a *preset* norm. Repeated interactions [1] and the sanctioning of non-conforming behavior [86] are two such conditions. It was less clear, however, whether and how a commonly shared norm would be established in situations, where the involved populations prefer *different* norms.

Our own approach to understanding behavioral consensus relates to evolutionary game theory, but in contrast to most models for the evolution of norms, it assumes a heterogeneity of individual preferences and therefore involves several populations. In *Methods* and *Discussion*, we furthermore consider the role of sanctioning as a mechanism that supports the evolution of norms, but point out that there are other mechanisms which are expected to support the evolution of norms as well. Note that the term "evolution" is used by us in the sense of "temporal evolution" or "eventual establishment", not in the sense of "biological evolution" or "spontaneous emergence".

Examples and classification of norms

In the following, we will illustrate the concept of social norms by some examples. In his book "The Cement of Society", Jon Elster [79] discusses consumption norms, norms against behavior 'contrary to nature', norms regulating the use of money, norms of reciprocity, medical ethics, codes of honor, norms of retribution, work norms, norms of cooperation, and norms of distribution. Norms underlying common neighborhood or business practices have been analyzed by Macaulay [87], Ostrom [88] and Ellickson [69].

When discussing norms, it is useful to distinguish between coordination norms and cooperation norms [83]. *Coordination norms* are self-enforcing. They are established, when it is advantageous for people to show a coordinated behavior, but it does not matter which of the behavioral options people agree upon. In that case, one also speaks of behavioral conventions, and one-population models are often sufficient to describe the underlying dynamics [61,89]. Examples of conventions are the preference of pedestrians to walk on one side [61,90–93] (for example, the right-hand side in continental Europe or the left-hand side in Japan), the direction of writing, the way people greet each other (whether one gives a hand and which one, whether one hugs or kisses the person and how many times), the way people eat, the color of clothes worn by political movements, and signs used by followers of certain ideas or tastes to identify each other (e.g. tattoos or hanky codes).

In contrast to coordination norms, *cooperation norms* are not self-enforcing, since there are incentives for unilateral deviance. In our paper, we analyze situations, where people have *different preferences*, so that at least a certain fraction of people is tempted to show non-conforming behavior. Gender norms may serve for illustration. Just imagine a "battle of the sexes" in a *group* of friends (rather than between two players), where men prefer to watch soccer and women prefer to see a cultural performance, to discuss a stereotypical example. Note that, in our model, interactions occur not only between men and women, but also among men and among women, so the outcome will depend on their relative power.

Religious norms constitute another case, where people with incompatible preferences interact with each other. A similar thing applies to legal norms, when people believing in a pluralistic civil law system interact with people believing in a religious law system. It is well-known that these law systems have incompatible implications with regard to certain issues. A similar situation applies, when businessmen from countries with different business practices make a deal or people with different mother languages meet. In our opinion, communicating in a language is not just a coordination problem. Most people have a clear preference for their mother tongue, and it shapes even the way of thinking and of social interactions. Therefore, when people with different mother tongues meet, there is an incentive to unilaterally deviate from speaking the same language (e.g., due to differences in proficiency). Nevertheless, a common ("normative") language *can* establish, as is impressively shown, for example, by the unification of regionally spoken dialects in Germany triggered by the Luther bible. Note, however, that proper language use does not seem to be fully self-enforcing, otherwise lexica, schools, and related legal regulations would not be needed.

Besides coordination and cooperation norms, it appears to make sense to distinguish a third class of "*hybrid norms*", which share features of both kinds of norms. This case occurs when it is *costly* to switch the behavior (i.e. when transaction costs are high). Technological norms may serve as an example. Customers will usually profit from shared technical standards concerning, for example, the type of keyboard (QWERTY or Dvorak) [94], the kind of operating system (Windows vs. Mac OS or Linux), the technology of video players (VHS vs. Beta MAX) [95] or high resolution DVD players (blue-ray vs. HD DVD). In such cases, customers do not have incentives to deviate from a technological standard, once it has established everywhere. In the beginning, however, a common standard does not evolve by itself, as customers buy different technologies and are reluctant to give up the technology they have invested in. Therefore, the use of a single technology is not self-enforcing in the beginning. Once there is a majority standard, however, most people will join it after some time, and their preferences change accordingly.

One should also mention that *some* conventions or norms are set by law, e.g. the driving side [89], or may involve an intentional segregation from other groups (e.g. when groups develop their own dress-codes or symbols such as tattoos). This touches issues of group dynamics, which are beyond the scope of this paper. The novel contribution of our model is that it sheds new light on the problem of whether a norm can establish (under what conditions) and how (in terms of the dynamics). There are even exact mathematical results for this [46]. In particular, our model reveals that the dynamics and finally resulting state of the system is not only determined by the payoff structure. It also depends on the power of populations and even on the initial proportions of cooperative individuals (the initial conditions or previous history).

Within our model of the evolution of norms, one could say that Figs. 3A,C,D represent the formation of *coordination norms*, as one behavioral norm is *always* established (reflecting self-enforcement). Figure 3B, in contrast, describes situations where two different behaviors can coexist in a stable way (see the yellow basin of attraction). Due to this lack of self-enforcement, it makes sense to attribute this case to the problem of establishing a *cooperation norm*. This relevant case can *only* occur, when taking into account self-interactions in multi-population games. It is also interesting to note, that the application of group pressure can transform the problem of establishing a *cooperation* norm into the problem of establishing a *coordination* norm (see Sec. *Methods*). Finally, the case of *hybrid norms* can be treated by considering transaction costs in our model.

Occurence of social polarization in the snowdrift game

Let us now turn our attention to the discussion of snowdrift games. In the *one-population* case, there is *one* stable stationary point, corresponding to a fraction $p_0 = |B|/(|B|+|C|)$ of cooperative individuals (see Fig. 1). If this would be transferable to the multi-

population case we are interested in, we should have $p = q = p_0$ in the limit of large times $t \to \infty$. Instead, we find a variety of different outcomes, depending on the values of the model parameters B, C, and f (see Fig. 5):

(a) *The interactions between both populations shift the fraction of cooperative individuals in each population to values different from* p_0. If $|B| = |C|$, we discover a line of *infinitely many* stationary points, and the actually resulting stationary solution uniquely depends on the initial condition (see Fig. 5A). This line satisfies the relation $q = p$ only if $f = 1/2$, while for most parameter combinations we have $q \neq p \neq p_0$. Nevertheless, the typical outcome in the case $|B| = |C|$ is characterized by a finite fraction of cooperative individuals in each population.

(b) *Conflicting interactions between two equally strong groups destabilize the stationary solution* $q = p = p_0$ *of the one- population case, and both populations lose control over the final outcome.* For $|B| \neq |C|$, all stationary points are *discrete* and located on the *boundaries*, and only *one* of these points is an evolutionary equilibrium. If both populations have *equal* power ($f = 1/2$), we either end up with non-cooperative behavior by everybody (if $p_0 < 1/2$, see Fig. 5B), or everybody is cooperative (if $p_0 > 1/2$). Remarkably, there is no mixed stable solution between these two extremes.

(c) *The stronger population gains control over the weaker one, but shows polarization itself, and a change of the model parameters may induce a revolution.* If $|B| \neq |C|$ and population 1 is much *stronger* than population 2 (i.e., $f - 1/2 \gg 0$), we find a finite fraction of cooperative individuals in the stronger population, while either 0% or 100% of the individuals are cooperative in the weaker population. A closer analysis reveals that the resulting overall fraction of cooperative individuals fits exactly the expectation p_0 of the stronger population [46], while from the perspective of the weaker population, the overall fraction of cooperative individuals is largely different from $p_0 = |B|/(|B|+|C|)$. Note that the stronger population alone can not reach an overall level of cooperation of p_0. The desired outcome can only be produced by effectively *controlling* the behavior of the weaker population. This takes place in an unexpected way, namely by polarization: The stronger *population* splits up into fractions of people showing different behaviors, which may give rise to social differentiation, inequality, and conflict. In the weaker population 2, everyone either shows behavior 1 (namely for $p_0 < 1/2$, see Fig. 5C), otherwise everyone shows behavior 2 (see Fig. 5D). There is no solution in between these two extremes (apart from the special case $p_0 = 1/2$ for $|B| = |C|$).

It comes as a further surprise that the behavior in the weaker population is always coordinated with the *minority* behavior in the stronger population. Due to the payoff structure of the multi-population snowdrift game, it is profitable for the weaker population to oppose the majority of the stronger population, which creates a tacit alliance with its minority. Such antagonistic behavior is well-known from protest movements [96] and finds here a natural explanation.

Moreover, when $|C|$ changes from values greater than $|B|$ to values smaller than $|B|$, there is an unexpected, discontinuous transition in the weaker population 2 from a state in which everybody is cooperative from the point of view of population 1 to a state in which everybody shows the *own* preferred behavior 2 (see Movie S3). History and science [97] have seen many abrupt regime shifts of this kind. Revolutions caused by class conflict provide ample empirical evidence for their existence. Combining the theory of phase transitions with "catastrophe theory" [98] offers a quantitative scientific approach to interpret such revolutions as the outcome of social interactions [99]. Here, their recurrence becomes understandable in a unified and simple game-theoretical framework.

Discussion

Multi-population game-dynamical replicator equations provide an elegant and powerful approach to study the dynamics and outcomes expected for populations with incompatible interests. A detailed analysis reveals how combining interactions within and between populations and considering differences in their power can substantially change the dynamics of various game theoretical dilemmas (compare Movies S2 and S3 with Movie S1). Generalizations to *more* than 2 behaviors or groups and to different payoffs for in- and out-group interactions are easily possible (see Sec. *Methods*).

When two populations with incompatible preferences interact among and between each other, we find the same stationary points for the prisoner's dilemma and the harmony game as for the corresponding non-interactive games. In particular, interactions across populations do not change the attractive solutions (Nash equilibria) of these games. However, the behavior of the snowdrift game and the stag hunt game changes completely, and their dynamics is particularly interesting. For the interactive case, the signs of the payoff-dependent parameters B and C do not only determine the character of the game, but also the location and stability of the stationary solutions, and the basins of attraction.

In the multi-population snowdrift game, for example, there is a discontinuous ("revolutionary") transition, when $1 - |B|/|C|$ changes its sign. On top of this, the power f has a major influence on the outcome, and the initial distribution of behaviors (and, consequently, the previous history) can be crucial, also for the multi-population stag hunt game. Note that such a rich system behavior is already found for the *simplest* setting of our model and that the concept of multi-population game-dynamical equations may be generalized in various ways to address a number of challenging questions in the future: How can we gain a better understanding of a clash of cultures, the outbreak of civil wars, or conflicts with ethnic or religious minorities? How can we analytically study migration and group competition? When do social systems become unstable and experience a polarization of society? How can we understand the emergence of fairness norms in bargaining situations?

Another interesting aspect of our model is that it makes a variety of quantitative predictions. Therefore, it could be tested experimentally with iterated games in the laboratory, involving several groups of people with random matching and sufficiently many iterations. Suitable changes in the payoff matrices should be able to confirm the mathematical conditions under which different archetypical types of social phenomena or discontinuous transitions in the system behavior can occur: (1) the breakdown of cooperation, (2) in-group cooperation (the formation of "subcultures"), (3) the evolution of shared behavioral norms, and (4) societal polarization or conflict with the possibility of a revolutionary regime shift. The findings are particularly important to understand interactions between human populations with different ethnic, cultural or religious backgrounds. However, they are also relevant for social features within animal societies [38,103–105] or even for interactions among bacteria [4,5].

The significant influence of the respective payoffs of social interactions on the resulting outcome has crucial implications for society, law and economics [62,106–115]. There, conflicts need to

be avoided or solved, and norms and standards are of central importance. For society, norms are equally important as cooperation, since they reduce uncertainty, bargaining efforts, and (potentially) also conflict in social interactions. They are like social forces guiding our interactions in numerous situations and subtle ways, creating an "invisible hand" kind of self-organization of society [47]. Nevertheless, their ubiquity is quite surprising, as norms require people to constrain self-interested behavior [116] and to perform socially prescribed roles. Yet, widespread cooperation-enhancing mechanisms such as direct reciprocity due to repeated interactions [1] and indirect reciprocity based on reputation [10] can transform a prisoner's dilemma into stag hunt interactions, see Fig. 6 [3,82,117].

This suggests a natural tendency towards the formation of norms, whatever their content may be. Costly punishment can support the evolution of norms in prisoner's dilemma situations as well (see Fig. 7A). Another way of promoting the preferred coordinated behavior as commonly accepted norm is to transform a prisoner's dilemma situation into a stag hunt game in the *own* population and to make sure that the population interacts with another one with incompatible preferences and prisoner's dilemma interactions (see Fig. 7B). Accordingly, the sanctioning of non-conforming behavior (see Sec. *Methods* for details) is not the *only* mechanism to support the evolution of norms. Other cooperation-enhancing mechanisms such as kin selection (based on genetic relationship) and group selection tend to transform a prisoner's dilemma into a *harmony game* (see Fig. 6). Therefore, genetic relatedness and group selection are *not* ideal mechanisms to establish shared behavioral norms. They rather support the formation of subcultures. Moreover, the transformation of prisoner's dilemma interactions into a *snowdrift game* is expected to cause polarization or conflict (see Fig. 6).

The evolution of language [119] is another example for the importance of norm-establishing social interactions, since successful communication requires norms, how words are used (the "evolution of meaning") [120,121]. In this connection, it is interesting to study whether the explosive development of language and culture in humans is due to their ability to transform interactions into norm-promoting stag hunt interactions. From this point of view, repeated interactions thanks to human settlements, the development of reputation mechanisms, and the invention of punishment institutions should have largely accelerated cultural evolution [122–124].

Another interesting research direction relates to the circumstance that people do not only *follow* norms—at the same time, they also *create* norms. This touches the issue of norm *emergence* [64,73,80,126–128]. In order to address it, one also has to answer questions such as the following: How is the "content" of norms generated or selected, i.e. how does the prescription of a behavioral role or normative behavior come about [129]? How and why do people start sanctioning non-conforming individuals, although this is costly, and why is there a tendency towards conformity at all [130]? This is beyond the scope of this paper. The same applies to related questions such as the following: Do norms always establish a "Pareto" or a "system optimum" [56,131–133]? Do norms always emerge, when they would be "functional" or beneficial [79,134]? Do they disappear, when they are not beneficial anymore? Do norms reduce or produce conflicts [31]? How can one explain local conformity, global diversity, and punctuated equilibria [51]? These points will be addressed in a forthcoming publication, while the goal of *this* manuscript was to develop a unified theoretical concept allowing one to study the interaction of populations with incompatible interests. This became possible by analysis of multi-population game-dynamical (replicator) equations, which have been used here to address, besides the evolution of norms, the occurrence of polarization or conflict, the outbreak of revolutions and several other relevant questions, like the importance of the power of a population and of

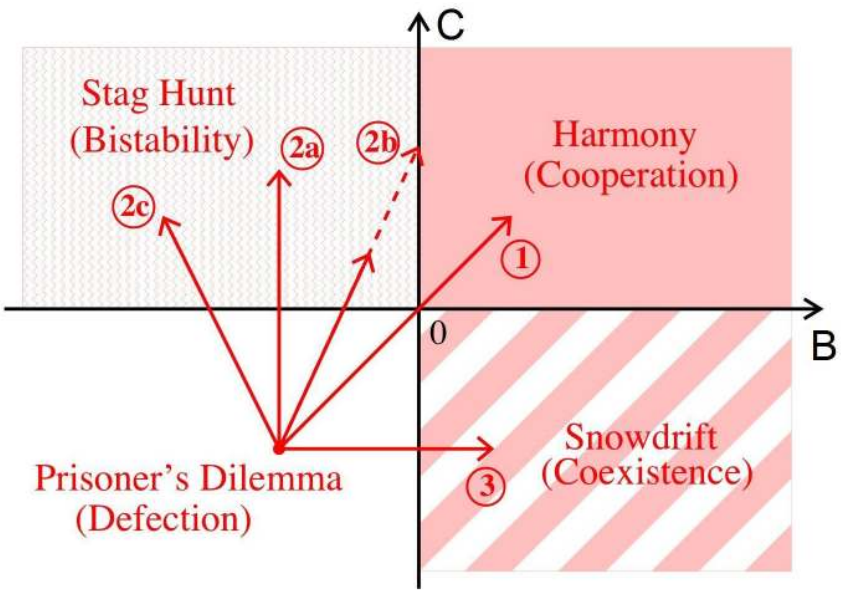

Figure 6. Illustration of different routes to cooperation (through arrows), assuming particular reproduction-selection mechanisms [3]. The direction of the arrows can be mathematically calculated [117]. Route 1 reflects the way in which the payoff-dependent parameters *B* and *C* of the game (see Fig. 1) are effectively modified by kin selection (genetic relationship), network reciprocity (clustering of individuals showing the same behavior), or group selection (competition between different groups). Route 2a corresponds to the effect of direct reciprocity (due to the "shadow of the future" through the likelihood of future interactions). Route 2b belongs to the mechanism of indirect reciprocity (based on reputation effects), and route 2c reflects costly punishment. Route 3 results for certain kinds of network interactions [3,118].
doi:10.1371/journal.pone.0012530.g006

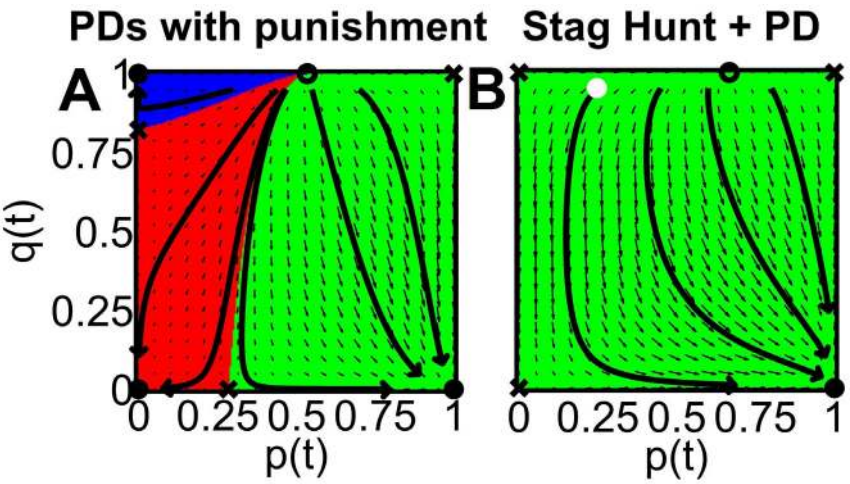

Figure 7. Illustration of different ways to establish a behavioral norm. (A) Prisoner's dilemma games with model parameters $B=-1$, $C=-2$, when population 1 is much stronger than population 2 ($f=0.8$) and individuals of both populations apply costly punishment, whenever the respective focal individual behaves cooperatively, but (from his or her perspective) the interaction partners does not [125]. The punishment cost was assumed to be $\beta=2.5$, the punishment fine $\gamma=5\beta$ (see Sec. *Methods* for details). When starting with an initial condition in the blue or green areas, costly punishment can establish a behavioral norm, corresponding to the asymptotically stable points at (0,1) and (1,0). When starting in the red area, however, everybody in both populations will finally behave in a non-cooperative way, as indicated by the asymptotically stable point at (0,0). (B) Population 1 playing a stag hunt game with model parameters $B=-1$ and $C=2$, while individuals in population 2 experience prisoner's dilemma interactions with $B=-1$ and $C=-2$, assuming equally strong populations ($f=0.5$). All individuals are expected to end up with behavior 1, which is preferred by population 1. However, as the flow line starting with the white circle and ending up in the asymptotically stable point (1,0) illustrates, the evolution of the behavioral norm can take a long time and unexpected detours.
doi:10.1371/journal.pone.0012530.g007

the previous history ("initial conditions") for the outcome of social interactions.

Methods

Multi-population game-dynamical replicator equations

Multi-population game-dynamical replicator equations [38,46,61] describe the temporal evolution of the proportions $p_i^a(t)$ of individuals showing behavior i at time t in population a, assuming that more successful behaviors spread, as these are imitated by individuals of the same population at a rate proportional to the increase in the expected success [61,92,135]. The expected success is determined from the frequency of interactions between two behaviors i and j, and by the associated payoffs A_{ij}^{ab}. Focusing on the above-mentioned social dilemmas, in the case of two interacting populations a, $b \in \{1, 2\}$ and two behavioral strategies i, $j \in \{1, 2\}$, we assume the following for interactions within the *same* population a: If two interacting individuals show the *same* behavior i, both will either receive the payoff r_a or p_a. If we have $r_a \neq p_a$, we call the behavior with the larger payoff r_a "preferred" or "cooperative", the other behavior "non-cooperative" or "uncooperative". When one individual chooses the cooperative behavior and the interaction partner is uncooperative, the first one receives the payoff s_a and the second one the payoff t_a. To model conflicts of interests, we assume that population $a=1$ prefers behavior $i=1$ and population 2 prefers behavior 2. Therefore, if an individual of population 1 meets an individual belonging to population 2 and both show the same behavior $i=1$, the first one will earn R_1 and the second one P_2, as behavior $i=1$ is considered uncooperative in population 2. Analogously, for $i=2$ they earn P_1 and R_2. If the interaction partners choose *different* behaviors i and j, they earn S_a, when the behavior corresponds to their cooperative behavior, otherwise they earn T_a [46]. In mathematical notation, the payoff matrix (A_{ij}^{11}) for individuals belonging to population 1 in interactions with other individuals of the same population is

		Interaction partner's behavior	
		$j=1$ (preferred)	$j=2$
Focal agent's behavior	$i=1$ (preferred)	r_1	s_1
	$i=2$	t_1	p_1

while the payoff matrix (A_{ij}^{12}) for interactions with individuals of population 2 is

		Interaction partner's behavior	
		$j=1$	$j=2$ (preferred)
Focal agent's behavior	$i=1$ (preferred)	R_1	S_1
	$i=2$	T_1	P_1

To reflect incompatible preferences of both populations, we assume that the payoff matrix of individuals belonging to population 2 is "inverted" or "mirrored". When interacting with individuals of population 1, the related payoff matrix (A_{ij}^{21}) is

		Interaction partner's behavior	
		$j=1$ (preferred)	$j=2$
Focal agent's behavior	$i=1$	P_2	T_2
	$i=2$ (preferred)	S_2	R_2

while the payoff matrix (A_{ij}^{22}) when interacting with individuals of

population 2 is

		Interaction partner's behavior	
		$j=1$	$j=2$ (preferred)
Focal agent's behavior	$i=1$	p_2	t_2
	$i=2$ (preferred)	s_2	r_2

Assuming constant preferences and fixed relative population strengths f_a, the resulting coupled game-dynamical replicator equations for the temporal evolution of the proportion $p(t)=p_1^1(t)$ of cooperative individuals in population 1 and the fraction $q(t)=q_2$ of cooperative individuals in population 2 become

$$\frac{dp(t)}{dt}= \underbrace{p(t)[1-p(t)]}_{\text{saturation factors}} \quad \underbrace{[b_1f+(c_1-b_1)fp(t)+C_1(1-f)+(B_1-C_1)(1-f)q(t)]}_{\text{growth factor } F(p,q) \text{ containing interaction effects}} \quad (1)$$

and

$$\frac{dq(t)}{dt}= \underbrace{q(t)[1-q(t)]}_{\text{saturation factors}} \quad \underbrace{[b_2(1-f)+(c_2-b_2)(1-f)q(t)+C_2f+(B_2-C_2)fp(t)]}_{\text{growth factor } G(p,q) \text{ containing interaction effects}} \quad (2)$$

Here, we have used the abbreviation $f=f_1=1-f_2$. As indicated above, this parameter can reflect the relative population size of population 1. More generally, however, $A_{ij}^{ab}f_b$ can be considered to represent "effective payoffs", and f can be used to model the "power" of population 1 (which may not only depend on population size, but also on education, the availability of weapons or technologies, and other factors). $b_a=s_a-p_a$, $B_a=S_a-P_a$, $c_a=r_a-t_a$, and $C_a=R_a-T_a$ are payoff-dependent model parameters, which can be positive, negative, or zero. When setting $b_a=B_a=B$ and $c_a=C_a=C$ for simplicity, the payoff depends on the own behavior i and the behavior j of the interaction partner only, but not on the population he/she belongs to (i.e. in- and out-group interactions just determine whether an interaction partner may be imitated or not, but they do not influence the payoff). Given the values of B and C used in our computer simulations, it is easily possible to construct related payoff matrices. Fixing values for P (e.g. $P=0$) and for $D=R-P$, we have $R(P,D)=P+D$, $T(P,D)=R-C=P+D-C$, and $S(P,D)=P+B$.

In reality, in-group and out-group interactions may, of course, affect the payoff as well. Such situations can be treated by choosing different values for the lower-case and upper-case parameters. It is obvious that this creates a multitude of additional cases, which deserve to be investigated in detail. However, before doing so, one first has to understand the basic case addressed in this study, which is already quite complicated.

Specification of Costly Punishment and Effects of Group Pressure

Let us now consider costly punishment analogously to the way it was specified in Ref. [125]. Then, we have $s_1=S-\beta_1$, $t_1=T-\gamma_1$, $S_1=S-\beta_1-\gamma_2$, $S_2=S-\beta_2-\gamma_1$, $t_2=T-\gamma_2$, $s_2=S-\beta_2$. This specification assumes that someone who receives the low "sucker's payoff" S in the event of unilateral cooperation (from his/her point of view), will punish the respective interaction partner immediately, which modifies the payoffs. The punishment performed by an individual of population a reduces the payoff of his/her interaction partner by the fine $\gamma_a>0$. However, punishment is costly (it needs some punishment effort), which reduces the punisher's payoff by $\beta_a>0$. Usually one assumes $\gamma_a>\beta_a$. The correspondingly changed payoffs imply the parameters $b_1=B-\beta_1$, $c_1=C+\gamma_1$, $B_1=B-\beta_1-\gamma_2$, $C_1=C$ and $b_2=B-\beta_2$, $c_2=C+\gamma_2$, $B_2=B-\beta_2-\gamma_1$, $C_2=C$, which must be inserted into Eqs. (1) and (2). Therefore, punishment transforms the prisoner's dilemma *within* a population into a stag hunt game, when $c_a=C+\gamma_a>0$, i.e. $\gamma_a>|C|$. The interaction with the *other* population remains a prisoner's dilemma, since $B_a<0$ and $C_a<0$. Altogether, costly punishment results in the modified two-population game-dynamical equations

$$\frac{dp(t)}{dt}= p(t)[1-p(t)][F(p,q)-\beta_1f[1-p(t)]+\gamma_1fp(t)-(\beta_1+\gamma_2)(1-f)q(t)] \quad (3)$$

and

$$\frac{dq(t)}{dt}= q(t)[1-q(t)][G(p,q)-\beta_2(1-f)[1-q(t)]+\gamma_2(1-f)q(t)-(\beta_2+\gamma_1)fp(t)] \quad (4)$$

These can generate $dp/dt\geq 0$ and $dq/dt\geq 0$ even for the prisoner's dilemma with $B<0$ and $C<0$. While punishment in a *one*-population prisoner's dilemma can lead to *cooperation* [125], in the *multi*-population case it can cause *normative behavior* (see green and blue areas in Fig. 7A).

Rather than considering costly punishment as discussed before, one may also consider that individuals can apply *group pressure* to support conformity and discourage discoordinated behavior [19,70]. That could be reflected by subtracting a value δ from the off-diagonal payoffs S and T or by adding δ to the diagonal elements R and P. This results in the effective model parameters $b_a=B_a=B-\delta$ and $c_a=C_a=C+\delta$ [117]. Therefore, if the group pressure δ is large enough (namely, $\delta>|C|$), a prisoner's dilemma with $B<0$ and $C<0$ is transformed into a stag hunt game with $b_a=B_a<0$ and $c_a=C_a>0$.

Discussion of Implicit Model Assumptions

Any model has some underlying model assumptions, and even though they may not be exactly fulfilled, the resulting model can be a useful approximation. The evolutionary game theoretical model of this paper assumes that individuals show a certain behavior and stick to it until they change it due to social learning (e.g. imitation) or mutations (e.g. trial-and-error behavior). This appears applicable to situations of routine choice [124] and to situations, where individuals do not spend much time on analyzing situations, but rather orient at what the others do. Crowd behavior and certain kinds of public opinion formation seem to be good examples for this [61]. The approach should also be applicable to many kinds of culturally acquired behaviors, including a considerable number of behavioral conventions and norms.

In contrast to evolutionary game theory, classical game theory assumes complex, strategic decision-making processes based on utility maximization. These decision-making processes usually consider individual preferences of interaction partners and other aspects. In spite of this difference, both approaches lead to

mutually consistent outcomes in many cases. For this reason, it should not matter for our main conclusions whether the analysis is based on classical or evolutionary game theory. Indeed both kinds of game-theoretical analysis are expected to show the four archetypical types of system behaviors identified in our paper. In this connection, it is interesting to note the following [2]:

1. Every Nash equilibrium is a fixed point of the replicator dynamics and the game-dynamical equation. (At a Nash equilibrium, which may also correspond to a mixed strategy, no player can improve the payoff by changing the strategy unilaterally.)
2. Every (asymptotically or neutrally) stable fix point of the replicator equation is a Nash equilibrium.
3. The fix points (and their stability properties) imply the main conclusions of our paper, as they determine the features of the system dynamics, which are reflected by the basins of attraction and the flow lines.

Further assumptions underlying the multi-population evolutionary game-theoretical model become obvious in the mathematical derivation of the equations. One- or multi-population replicator equations may describe the spreading of more successful individuals via a higher reproduction rate (see, e.g. [35]. However, they may also reflect social learning, namely by the imitation of more successful behaviors [46,61,92]. The above model equations result in case of proportional imitation [61,135]. They assume that interactions take place in large, well-mixed populations, usually between different individuals. If effects of repeated interactions (and, thereby, a "shadow of the future") shall be taken into account, this can be done by modifying the payoffs accordingly [3,117]. A similar thing applies to reputation effects, network reciprocity, group selection, etc. (see also Fig. 6). It is furthermore possible to generalize the approach to finite populations [136] and to populations with spatial or network interactions (see, for example, [14,137,138].

Entities belonging to different populations differ in two aspects: They earn different payoffs, and they imitate different entities (namely, better-performing entities belonging to the *own* population, assuming that it would not necessarily be wise to copy behaviors of individuals with different preferences and payoff functions). If, additionally, in-group interactions shall be distinguished from out-group interactions (in the sense that not only the *actual* behavior, but also the *preferred* behavior of the interaction partner matters), one has to specify the parameters b_a and c_a differently from the parameters B_a and C_a. In this way, one can even consider cases, where both populations play different games (see Fig. 7B).

Supporting Information

Movie S1 Vector fields (small arrows) and phase diagrams (colored areas) for two interacting populations with incompatible preferences (conflicting interactions), when population 1 is more powerful than population 2 ($f = 0.8$), i.e. population 1 is assumed to be more powerful. The movie shows the situation for the two-population snow-drift game (first half of the movie) and the two-population stag hunt game (second half), when in-group interactions are considered, while interactions between populations are neglected ($b_a = b$, $c_a = c$, $B_a = 0 = C_a$). Therefore, the dynamics in each population is independent of the dynamics in the other population. The size of the parameters B and C is varied according to the relation $C = -B^3$. This serves to demonstrate the parameter-dependence of the fix points and dynamics of both games. The small moving dots illustrate trajectories. One can clearly see the discontinuous transitions in the system behavior when one of the parameters B, C, or $1-|B|/|C|$ changes its sign. In the snowdrift game, we find a stable fraction $p_0 = |B|/(|B|+|C|)$ of cooperative individuals in each population, i.e. $p = p_0 = q$. This stationary fix point corresponds to the large black circle moving along the diagonal line. In the stag hunt game, the fix point located on the diagonal line is unstable (see empty circle). Therefore, trajectories move away from it. If the fraction of cooperative individuals in a population is larger than p_0, it will grow further, otherwise it will continuously shrink. That is, each population will either end up with 0% or 100% cooperative individuals, depending on the initial conditions. Therefore, $2^2 = 4$ stable fix points are possible - one in each corner. Further details: p is the fraction of individuals in population 1 showing their preferred, cooperative behavior 1, and q is the fraction of cooperative individuals in population 2 showing their preferred behavior 2. A fraction $1-q$ of individuals in population 2 shows the non-preferred behavior 1, and a fraction $1-p$ of individuals in population 1 shows behavior 2. The vector fields displays (dp/dt, dq/dt), i.e. the direction and size of the expected temporal change of the behavioral distribution, if the fractions of cooperative individuals in populations 1 and 2 are p(t) and q(t). Trajectories are representative flow lines (p(t), q(t)) as time t passes. The flow lines move away from unstable stationary points (empty circles) and are attracted towards stable stationary points (black circles). The colored areas represent the basins of attraction, i.e. all initial conditions (p(0), q(0)) leading to the same fix point [red = (0,0), yellow = (1,1), blue = (0,1), green = (1,0), salmon = (u, 0), mustard = (v, 1), other colors = (u, v), with $0<u, v<1$]. Saddle points (crosses) are attractive in one direction, but repulsive in another.
Found at: doi:10.1371/journal.pone.0012530.s001 (6.66 MB AVI)

Movie S2 Same as Movie S1, but while interactions between both populations are considered, self-interactions are neglected ($b_a = 0 = c_a$, $B_a = B$, $C_a = C$). The contrast to Movie S1 is pronounced: In the snowdrift game (first half of the movie), everybody is either cooperative or non-cooperative in both populations now, corresponding to the stable fix points at (0,0) and (1,1) (see black circles). In contrast, in the stag hunt game (second half of the movie), the evolutionary equilibria are located at (p, q) = (1, 0) and (p, q) = (0, 1). $p = 1$ means that 100% of the individuals in population 1 show behavior 1, while $q = 0$ implies that 0% of the individuals in in population 2 show behavior 2 (i.e. all of them show behavior 1 as well). Therefore, we find the establishment of a commonly shared behavior (the formation of a behavioral norm).
Found at: doi:10.1371/journal.pone.0012530.s002 (8.07 MB MPG)

Movie S3 Same as Movie S1, but considering both, interactions within and between the two populations. Assuming no difference between in-group and out-group interactions, we have $b_a = B_a = B$ and $c_a = C_a = C$. While the multi-population stag hunt game (first half of the movie) shows a tendency to establish a commonly shared behavior ("behavioral norm"), the snowdrift game (second half) rather delineates situations of conflict between both populations. It is known that conflicts between two populations may sometimes cause a "revolution". According to our interpretation, this corresponds to the discontinuous transition of the evolutionary equilibrium, when the background color turns from salmon to mustard. The abrupt change of the q-coordinate from 0 to 1 means that all individuals in the weaker population show the non-preferred behavior before the revolution, but their preferred

behavior afterwards. The discontinuous transition occurs, when |B| and |C| in the multi-population snowdrift game become the same. (Note that there is no such revolutionary transition, when individuals have compatible preferences.) The dynamics for two interacting populations without self-interactions is clearly less differentiated (see Movie S2). In particular, Movie S2 shows no revolutionary transition in the snowdrift game. It also lacks cases where the phase diagram of the stag hunt game displays three different basins of attraction at the same time, corresponding to a coexistence of three stable fix points. While two of them correspond to the establishment of a commonly shared behavior (a behavioral norm), the third point represents the formation of different behaviors (separate "subcultures") in each population.
Found at: doi:10.1371/journal.pone.0012530.s003 (10.15 MB MPG)

Acknowledgments

The authors are grateful to Thomas Chadefaux, Andreas Flache, Ryan Murphy, Carlos Roca, Stefan Bechtold, Sergi Lozano, Heiko Rauhut, Wenjian Yu, and further colleagues for valuable comments and to Sergi Lozano for drawing Fig. 6. They would also like to acknowledge concrete suggestions of the anonymous referees regarding possible improvements of the manuscript. Furthermore, D.H. thanks Thomas Voss for his insightful seminar on social norms.

Author Contributions

Conceived and designed the experiments: DH. Performed the experiments: AFJ. Analyzed the data: DH AFJ. Wrote the paper: DH.

References

1. Axelrod R (1984) The Evolution of Cooperation Basic Books, New York.
2. Gintis H (2000) Game Theory Evolving Princeton University Press, Princeton, NJ.
3. Nowak MA (2006) Five rules for the evolution of cooperation. Science 314: 1560–1563 and related Supporting Online Material.
4. Ben Jacob E, Becker I, Shapira Y, Levine H (2004) Bacterial linguistic communication and social intelligence. Trends in Microbiology 12(8): 366–372.
5. Griffin AS, West SA, Buckling A (2004) Cooperation and competition in pathogenic bacteria. Nature 430: 1024–1027.
6. Harsanyi JC, Selten R (1988) A General Theory of Equilibrium Selection MIT Press, Cambridge, MA.
7. Macy MW (1998) Social order in artificial worlds. Journal of Artificial Societies and Social Simulation 1, no. 1.
8. Macy MW, Flache A (2002) Learning dynamics in social dilemmas. Proc Natl Acad Sci (USA) 99, Suppl. 3: 7229–7236.
9. Raub W, Weesie J (1990) Reputation and efficiency in social interactions: An example of network effects. American Journal of Sociology 96(3): 626–654.
10. Milinski M, Semmann D, Krambeck HJ (2002) Reputation helps solve the "tragedy of the commons". Nature 415: 424–426.
11. Castelfranchi C, Conte R, Paolucci M (1998) Normative reputation and the costs of compliance. Journal of Artificial Societies and Social Simulation 1, no. 3.
12. Buskens V (2002) Social Networks and Trust Kluwer Academic, Dordrecht.
13. Nowak MA, May RM (1992) Evolutionary games and spatial chaos. Nature 359: 826–829.
14. Flache A, Hegselmann R (2001) Do irregular grids make a difference? Relaxing the spatial regularity assumption in cellular models of social dynamics. Journal of Artificial Societies and Social Simulation 4, no. 4.
15. Heckathorn DD (1990) Collective sanctions and compliance norms: A formal theory of group-mediated social control. American Sociological Review 55(3): 366–384.
16. Kandori M (1992) Social norms and community enforcement. Rev Econ Stud 59: 63–80.
17. Bendor J, Mookherjee D (1990) Norms, third-party sanctions, and cooperation. Journal of Law. Economics, and Organization 6: 33–63.
18. Posner RA, Rasmussen EB (1999) Creating and enforcing norms, with special reference to sanctions. International Review of Law and Economics 19(3): 369–382.
19. Fehr E, Gächter S (2002) Altruistic punishment in humans. Nature 415: 137–140.
20. Fehr E, Fischbacher U (2004b) Third party punishment and social norms. Evolution and Human Behavior 25: 63–87.
21. Helbing D, Szolnoki A, Perc M, Szabó G (2010) Evolutionary establishment of moral and double moral standards through spatial interactions. PLoS Computational Biology 6(4): e1000758.
22. Helbing D, Yu W (2009) The outbreak of cooperation among success-driven individuals under noisy conditions. Proceedings of the National Academy of Sciences (USA) 106(8): 3680–3685.
23. Lindbeck A, Nyberg S, Weibull J (1999) Social norms and economic incentives in the welfare state. Quarterly Journal of Economics 114: 1–35.
24. Lehmann L, Keller L (2006) The evolution of cooperation and altruism – A general framework and a classification of models. J Evol Biol 19: 1365–1376.
25. Fletcher JA, Doebeli M (2009) A simple and general explanation for the evolution of altruism. Proc Roy Soc B 276: 13–19.
26. Weibull JW (1995) Evolutionary Game Theory MIT Press, Cambridge, MA.
27. Sugden R (1995) A theory of focal points. The Economic Journal 105: 533–550.
28. Bacharach M, Bernasconi M (1997) The variable frame theory of focal points: An experimental study. Games and Economic Behavior 19: 1–45.
29. Bacharach M, Stahl DO (2000) Variable-frame level-n theory. Games and Economic Behavior 32: 220–246.
30. Bacharach M (2006) Beyond Individual Choice Princeton University, Princeton, NJ.
31. Stouffer SA (1949) An analysis of conflicting social norms. American Sociological Review 14(6): 707–717.
32. Saam N, Harrer A (1999) Simulating norms, social inequality, and functional change in artificial societies. Journal of Artificial Socieites and Social Simulation 2, no. 1.
33. Cressman R, Dash AT, Akin E (1986) Evolutionary games and two species population dynamics. Journal of Mathematical Biology 23: 221–230.
34. Cressman R, Garay J, Hofbauer J (2001) Evolutionary stability concepts for N-species frequency-dependent interactions. Journal of Theoretical Biology 211: 1–10.
35. Hofbauer J, Sigmund K (1998) Evolutionary Games and Population Dynamics Cambridge University, Cambridge.
36. Cressman R (1995) Evolutionary game theory with two groups of individuals. Games and Economic Behavior 11: 237–253.
37. Cressman R (1996) Frequency-dependent stability for two-species interactions. Theoretical Population Biology 49: 189–210.
38. Schuster P, Sigmund K, Hofbauer J, Gottlieb R, Merz P (1981) Selfregulation of behaviour in animal societies. III. Games between two populations with selfinteraction. Biological Cybernetics 40: 17–25.
39. de Oliveria VM, Fontanari JF (2000) Random replicators with high-order interactions. Physical Review Letters 85: 4984–4987.
40. de Oliveria VM, Fontanari JF (2002) Complementarity and diversity in a soluble model ecosystem. Physical Review Letters 89: 148101.
41. Diederich S, Opper M (1989) Replicators with random interactions: A solvable model. Physical Review A 39: R4333–R4336.
42. Sato Y, Akiyama E, Crutchfield JP (2005) Stability and diversity in collective adaptation. Physica D 210: 21–57.
43. Galla T (2006) Random replicators with asymmetric couplings. J Phys A: Math Gen 39: 3853–3869.
44. Cressman R (2003) Evolutionary Dynamics and Extensive Form Games MIT Press, Cambridge, MA.
45. Kanazawa T, Ushio T, Yamasaki T (2007) Replicator dynamics of evolutionary hypergames. IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans 37(1): 132–138.
46. Helbing D, Johansson A (2010) Evolutionary Dynamics of Populations with Conflicting Interactions: Classification and Analytical Treatment Considering Asymmetry and Power. Physical Review E 81: 016112.
47. Axelrod R (1986) An evolutionary approach to norms. American Political Science Review 80(4): 1095–111.
48. Conte R, Castelfranchi C (1995) Understanding the functions of norms in social groups through simulation. In: NGilbert, RConte, eds. Artificial Societies, The Computer Simulation of Social Life (UCL Press, London). pp 252–267.
49. Conte R, Falcone R, Sartor G (1999) Agents and norms: How to fill the gap? Artificial Intelligence and Law 7: 1–15.
50. Dignum F (1999) Autonomous agents with norms. Artificial Intelligence and Law 7: 69–79.
51. Epstein JM (2001) Learning to be thoughtless: Social norms and individual computation. Computational Economics 18: 9–24.
52. Flentge F, Polani D, Uthmann T (2001) Modelling the emergence of possession norms using memes. Journal of Artificial Societies and Social Simulation 4, no. 4.
53. Thébaud O, Locatelli B (2001) Modelling the emergence of resource-sharing conventions: An agent-based approach. Journal of Artificial Societies and Social Simulation 4, no. 2.
54. Nakamaru M, Levin SA (2004) Spread of two linked social norms on complex interaction networks. Journal of Theoretical Biology 230: 57–64.
55. Ehrlich PR, Levin SA (2005) The evolution of norms. PLoS Biology 3(6): 0943–0948.

56. Centola D, Willer R, Macy M (2005) The Emperor's Dilemma: A Computational Model of Self-Enforcing Norms. American Journal of Sociology 110(4): 1009–40.
57. Galan JM, Izquierdo LR (2005) Appearances can be deceiving: Lessons learned re-implementing Axelrod's 'Evolutionary Approach to Norms'. Journal of Artificial Socieites and Social Simulation 8, no. 3.
58. Chalub FACC, Santos FC, Pacheco JM (2006) The evolution of norms. Journal of Theoretical Biology 241: 233–240.
59. Fent T, Groeber P, Schweitzer F (2007) Coexistence of social norms based on in- and out-group interactions. Advances of Complex Systems 10(2): 271–286.
60. Neumann M (2008) Homo Socionicus: A case study of simulation models of norms. Journal of Artificial Societies and Social Simulation 11, no. 4 6.
61. Helbing D (1992) A mathematical model for behavioral changes by pair interactions, in Haag G, Mueller U, Troitzsch KG, eds. Economic Evolution and Demographic Change Springer, Berlin. pp 330–348.
62. Samuelson L (1998) Evolutionary Games and Equilibrium Selection The MIT Press, Chap. 5: The Ultimatum Game.
63. Voss T (2001) Game theoretical perspectives on the emergence of social norms. In: MHechter, K.-DOpp, eds. Social Norms (Russell Sage Foundation, New York). pp 105–136.
64. Opp K-D (2001) How do social norms emerge? An outline of a theory. Mind and Society 2: 101–128.
65. Bicchieri C (2006) The Grammar of Society: The Nature and Dynamics of Social Norms Cambridge University, New York.
66. Bicchieri C, Jeffrey R, Skyrms B, eds. (2009) The Dynamics of Norms Cambridge University Press, Cambridge.
67. Homans GC (1974) Social Behavior Its Elementary Forms. Harcourt, New York.
68. Popitz H (1980) Die normative Konstruktion von Gesellschaft Mohr, Tübingen.
69. Ellickson R (1991) Order without Law Harvard University Press, Cambridge, MA.
70. Cialdini RB, Trost MR (1998) Social influence: Social norms, conformity, and compliance. In: Gilbert DT, Fiske ST, Lindzey G, eds. The Handbook of Social Psychology, 4th ed., Vol. II McGraw-Hill, Boston, MA, Chap. 21. pp 151–192.
71. Ostrom E (2000) Collective action and the evolution of social norms. Journal of Economic Perspectives 14(3): 137–158.
72. Fehr E, Gächter S (2000) Cooperation and punishment in public goods experiments. American Economic Review 90(4): 980–994.
73. Horne C (2001a) Sociological perspectives on the emergence of norms. In: Hechter M, Opp K-D, eds. Social Norms (Russell Sage, New York).
74. Horne C (2008) Norm enforcement in heterogeneous groups. Rationality and Society 20(2): 147–172.
75. Boyd R, Gintis H, Bowles S, Richerson PJ (2003) The evolution of altruistic punishment. Proc Natl Acad Sci USA 100: 3531–3535.
76. Fehr E, Fischbacher U (2004a) Social norms and human cooperation. Trends in Cognitive Science 8(4): 185–190.
77. Henrich J, McElreath R, Barr A, Ensminger J, Barrett C, et al. (2006) Costly punishment across human societies. Science 312: 1767–1770.
78. Elster J (1989a) The Cement of Society Cambridge University Press, Cambridge.
79. Elster J (1989b) Social norms and economic theory. Journal of Economic Perspectives 3(4): 99–117.
80. Kitts J (2006) Social influence and the emergence of norms amid ties of amity and enmity. Simulation Modelling Practice and Theory 14: 407–422.
81. Skyrms B (1996) Evolution of the Social Contract Cambridge University, Cambridge.
82. Skyrms B (2003) The Stag Hunt and the Evolution of Social Structure Cambridge University, Cambridge.
83. Ullmann-Margalit E (1977) The Emergence of Norms Oxford University, Oxford.
84. Bendor J, Swistak P (2001) The evolution of norms. American Journal of Sociology 106(6): 1493–1545.
85. Hechter M, Opp K-D, eds. Social Norms Russell Sage, New York.
86. Fehr E, Fischbacher U, Gächter S (2002) Strong reciprocity, human cooperation, and the enforcement of social norms. Human Nature 13: 1–25.
87. Macaulay S (1963) Non-contractual relations in business: A preliminary study. Americal Sociological Review 28: 55–67 (1963).
88. Ostrom E (1990) Governing the Commons The Evolution of Institutions for Collective Action. Cambridge University Press, New York.
89. Young HP (1993) The evolution of conventions. Econometrica 61: 57–84.
90. Helbing D (1991) A mathematical model for the behavior of pedestrians. Behavioral Science 36: 298–310.
91. Helbing D (1994) A mathematical model for the behavior of individuals in a social field. Journal of Mathematical Sociology 19(3): 189–219.
92. Helbing D (1996) A stochastic behavioral model and a 'microscopic' foundation of evolutionary game theory. Theory and Decision 40: 149–179.
93. Moussaïd M, Helbing D, Garnier S, Johansson A, Combe M, et al. (2009) Experimental study of the behavioural mechanisms underlying self-organization in human crowds. Proceedings of the Royal Society B 276: 2755–2762.
94. David P (1985) Clio and the economics of QWERTY. American Economic Review 75(2): 332–337.
95. Arthur WB (1989) Competing technologies, increasing returns, and lock-in by historical events. The Economic Journal 99: 116–131.
96. Opp K-D (2009) Theories of Political Protest and Social Movements Routledge, London.
97. Kuhn TS (1962) The Structure of Scientific Revolutions University of Chicago, Chicago.
98. Zeeman EC (1977) Catastrophe Theory Addison-Wesley, London.
99. Weidlich W, Huebner H (2008) Dynamics of political opinion formation including catastrophe theory. Journal of Economic Behavior & Organization 67: 1–26.
100. Hofbauer J, Sorger G (2002) A differential game approach to evolutionary equilibrium selection. International Game Theory Review 4(1): 17–31.
101. Van Damme E (1994) Evolutionary game theory. European Economic Review 38(3–4): 847–858.
102. Blume A (1998) Communication, risk, and efficiency in games. Games and Economic Behavior 22(2): 171–202.
103. Clutton-Brock T (2009) Cooperation between non-kin in animal societies. Nature 462: 51–57.
104. Wilson EO (2000) Sociobiology The Belknap Press, Cambridge, MA.
105. Bonner T (1980) The Evolution of Culture in Animals Princeton University Press, Princeton, NJ.
106. Sugden R (1986) The Evolution of Rights, Cooperation and Welfare Blackwell, New York.
107. Sugden R (1989) Spontaneous order. The Journal of Economic Perspectives 3(4): 85–97.
108. Sugden R (1998) Normative expectations: The simultaneous evolution of institutions and norms. In: Economics, Values, and Organizations (Cambrdige University, Cambridge). pp 73–100.
109. Koford KJ, Miller JB (1991) Social Norms and Economic Institutions (University of Michigan, Ann Arbour, MI.
110. Sethi R (1996) The evolution of social norms in common property resource use. American Economic Review 86(4): 766–788 (1996).
111. Binmore K, Samuelson L (1994) An economist's perspective on the evolution of norms. Journal of Institutional and Theoretical Economics 150: 45–63.
112. Binmore K (2005) Natural Justice Oxford University, New York.
113. Platteau J-P (2000) Institutions, Social Norms, and Economic Development Routledge, London.
114. Ellickson RC (2001) The evolution of social norms: A perspective from the legal academy. In: Hechter M, Opp K-D, eds. Social Norms (Russell Sage Foundation, New York). pp 35–75.
115. Bohnet I, Frey BS, Huck S (2001) More order with less law: On contract enforcement, trust, and crowding. American Political Science Review 95(1): 131–144.
116. Keizer K, Lindenberg S, Steg L (2008) The spreading of disorder. Science 322: 1681–1685.
117. Helbing D, Lozano S (2010) Phase transitions to cooperation in the prisoner's dilemma. Physical Review E 81(5): 057102.
118. Ohtsuki H, Nowak MA (2006) The replicator equation on graphs. Journal of Theoretical Biology 243: 86–97 (2006).
119. Nowak MA, Komarova NL, Niyogi P (2002) Computational and evolutionary aspects of language. Nature 417: 611–617.
120. Castello X, Eguiluz VM, Miguel MS (2006) Ordering dynamics with two non-excluding options: Bilingualism in language competition. New Journal of Physics 8: 308.
121. Baronchelli A, Loreto V, Steels L (2008) In-depth analysis of the naming game dynamics: The homogeneous mixing case. International Journal of Modern Physics C 19(5): 785–812.
122. Boyd R, Richerson PJ (1985) Culture and the Evolutionary Process The University of Chicago, Chicago.
123. Boyd R, Richerson PJ (1994) The evolution of norms: An anthropological view. Journal of Institutional and Theoretical Economics 150(1): 72–87.
124. Gintis H (2009) The Bounds of Reason Game Theory and the Unification of the Behavioral Sciences. Princeton University Press, Princeton.
125. Traulsen A, Hauert C, De Silva H, Nowak MA, Sigmund K (2009) Exploration dynamics in evolutionary games. Proceedings of the National Academy of Sciences (USA) 106(3): 709–712.
126. Opp K-D (1979) The emergence and effects of social norms. Kyklos 32: 775–801.
127. Opp K-D (1982) The evolutionary emergence of norms. British Journal of Social Psychology 21: 139–149.
128. Horne C (2001b) Sex and sanctioning: Evaluating two theories of norm emergence. In: Hechter M, Opp K-D, eds. Social Norms (Russell Sage Foundation, New York). pp 305–324.
129. Boyd R, Richerson PJ (2005) The Origin and Evolution of Cultures Oxford University, Oxford.
130. Oliver P (1980) Rewards and punishments as selective incentives for collective action. American Journal of Sociology 85: 1356–1375.
131. Banfield EC (1967) The Moral Basis of a Backward Society Free Press, New York.
132. Putnam RD, Leonardi R, Nanetti RY (1994) Making Democracy Work: Civic Traditions in Modern Italy Princeton University, New Jersey.
133. Posner E (2000) Law and Social Norms Harvard University Press, Cambridge, MA.

134. Coleman JS (1990) Foundations of Social Theory Harvard University, Cambridge, MA.
135. Schlag KH (1998) Why imitate, and if so, how? A boundedly rational approach to multi-armed bandits. Journal of Economic Theory 78(1): 130–156.
136. Taylor C, Fudenberg D, Sasaki A, Nowak MA (2004) Evolutionary game dynamics in finite populations. Bulletin of Mathematical Biology 66(6): 1621–1644.
137. Roca CP, Cuesta JA, Sánchez A (2009) Effect of spatial structure on the evolution of cooperation. Physical Review E 80: 046106.
138. Szabó G, Fáth G (2007) Evolutionary games on graphs. Physics Reports 446(4–6): 97–216.