Cooperation, Norms, and Conflict: A Unified Approach

Dirk Helbing
Anders Johansson

SFI WORKING PAPER: 2009-09-040

SANTA FE INSTITUTE

Cooperation, Norms, and Conflict: A Unified Approach

Dirk Helbing[1,2,3*] and Anders Johansson[1]
[1] *ETH Zurich, Chair of Sociology, in particular of Modeling and Simulation, UNO D11, Universitätstr. 41, 8092 Zurich, Switzerland*
[2] *Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA*
[3] *Collegium Budapest—Institute for Advanced Study, Szentháromság u. 2, 1014 Budapest, Hungary*
(Dated: September 14, 2009)

Cooperation is of utmost importance to society, 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 heterogeneous preferences. Yet, interactions between people with incompatible interests often yield conflict, since behavior that is considered cooperative by *one* population might be perceived as non-cooperative by another. To understand the outcome of such competitive interactions, we study game-dynamical replicator equations for multiple populations with incompatible preferences and different power to reveal, for example, what mechanisms can foster the evolution of behavioral norms? When does cooperation fail, leading to conflict or even to revolutions? What incentives are needed to reach peaceful agreements? Our quantitative analysis reveals some striking results, significant for society, law, and economics.

In order to gain a better understanding of factors preventing or promoting cooperation among individuals, 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) [1–3]. These games have in common that a certain fraction of people or even everyone is expected to behave uncooperatively (see Supporting Information (SI) [36]). Therefore, a large amount of research has focused on how cooperation can be supported by mechanisms such as [3] repeated interactions [1], reputation [4], clusters of cooperative individuals [5], costly punishment [6–12], or success-driven migration [13].

Unfortunately, comparatively little attention has been devoted to the problem of cooperation between groups with different preferences (e.g. people of different gender, status, age, or cultural background). Yet, what constitutes cooperative behavior for one group might be considered non-cooperative by another. For example, men and women appear to have different preferences many times, but 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 working days and free days, food one may eat or should avoid, 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 [36]. Is it possible to identify factors determining whether two such populations go their own

*Electronic address: dhelbing@ethz.ch

way, find a common agreement, or end up in conflict? And what is the relevance of power in the rivalry of populations? Differences in power can, for example, result from different sizes of the interacting populations, their material resources (money, weapons, etc.), social capital (status, social influence, etc.), and other factors (charisma, moral persuasion, etc.).

I. MODEL

As a mathematical approach to this problem, we propose game-dynamical replicator equations for multiple populations [14, 15]. The crucial point is to adjust them in a way that reflects interactions between individuals with incompatible preferences (see Methods and Supporting Information [36]). These equations 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. The success depends on the "payoffs" resulting from social interactions, i.e., on the own behavior and the behavior of the interaction partners.

In order to reflect incompatible interests of both populations, we assume that population 1 prefers behavior 1 (e.g. everybody should be undressed at the beach), while population 2 prefers behavior 2 (everybody should be properly dressed). If an interaction partner shows the behavior preferred by oneself, this behavior is called "cooperative", otherwise uncooperative. In other words, 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 *same* behavior, we call this behavior "coordinated". Finally, if the *majority* of individuals in *both* populations shows a coordinated behavior, we speak of "normative behavior" or a "behavioral norm".

To establish a behavioral norm, the individuals belonging to *one* of the populations have to act against their own preferences, in particular as we assume that preferred behaviors and population sizes do not change. (Otherwise, identical behavior in both populations could simply result from the adaptation of preferences or group membership, which can, of course, further promote consensus in reality.)

It is very interesting to study under what conditions interactions within and between populations with incompatible preferences can lead to cooperation, conflict, or "normative behavior". To address this question, we adjust the theoretical framework of game-dynamical replicator equations to the case of two populations playing 2×2 games, which are represented by four payoffs T, R, P, and S. In the prisoner's dilemma, for example, 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. The related game-dynamical replicator equations read

$$\frac{dp(t)}{dt} = p(t)[1-p(t)]F\big(p(t),q(t)\big) \tag{1}$$

and

$$\frac{dq(t)}{dt} = q(t)[1-q(t)]G\big(p(t),q(t)\big)\,, \tag{2}$$

where the terms $p(1-p)$ and $q(1-q)$ can be interpreted as saturation factors. They make sure that the proportions $p(t)$ and $q(t)$ of individuals pursuing their preferred strategies stay within the range from 0 to 1. $F(p,q)$ and $G(p,q)$ are functions reflecting the interactions be-

tween individuals. They include terms describing "in-group" interactions ("self-interactions", reflecting encounters with individuals of the *same* population) and "out-group" interactions (among individuals belonging to *different* populations) (see Methods).

For simplicity, we will focus on the case where both populations play the same game. Then, the functions F and G only depend on the payoff-dependent parameters $B = S - P$, $C = R - T$, and the relative power f of population 1 (see Methods). Furthermore, we specify the power by the relative population size (see [36] for details). The parameter C may be interpreted as gain of coordinating on one's own preferred behavior (if greater than zero, otherwise as loss). B may be viewed as gain when giving up coordinated, but non-preferred behavior. Despite the simplifications made, this model for two populations with incompatible preferences shows very interesting features, and it can be generalized in numerous ways (to consider more populations, to treat heterogeneous, but compatible interactions, to reflect that payoffs in out-group interactions may differ from in-group interactions, to consider migration between populations, spatial interactions, learning, punishment, etc.).

II. RESULTS

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) The case $B < 0$ and $C < 0$ applies to the *multi-population prisoner's dilemma (MPD)*, (2) in the *multi-population harmony game (MHG)*, we have $B > 0$ and $C > 0$, (3) the *multi-population snowdrift game (MSD)* is characterized by $B > 0$ and $C < 0$, and (4) in the *multi-population stag hunt game (MSH)*, we have $B < 0$ and $C > 0$. In the multi-population harmony game, everybody shows cooperative behavior, while in the multi-population prisoner's dilemma, everybody is uncooperative in the end, as one may expect (see [36]). However, as we will show in the following, the dynamics and outcome of the multi-population stag hunt and snowdrift games are in marked contrast to the one-population case. This can be demonstrated by mathematical analysis of the stationary solutions of Eqs. [1] and [4] and their stability properties (see Methods and SI [36]). However, our results can be more intuitively illustrated by figures and movies showing the evolutionary equilibria of the games for various parameter values B and C, their basins of attraction, and representative flow lines. Details are discussed below and in the captions of Figs. 1+2 (see also the supporting information [36], particularly Movies S1–S3).

A. Evolution of normative behavior in the stag hunt game

The one-population stag hunt game is characterized by an equilibrium selection problem [16]: Everyone is finally expected to cooperate, if the initial fraction of cooperative individuals is above $p_0 = |B|/(|B| + |C|)$, otherwise everybody will behave uncooperative in the end [17]. The same applies to non-interacting populations (see Movie S1). However, in the multi-population stag hunt game with incompatible preferences and no self-interactions, it *never* happens 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 behavioral norm [18].

If self-interactions are taken into account as well, the case where everybody or nobody cooperates in both populations is possible, but rather exceptional (see Fig. 1 and Movie S3). It requires that both populations have similar strengths ($f \approx 1/2$) and the initial levels of cooperation are comparable as well (see yellow area in Fig. 1B). Under such conditions, both populations may develop separate subcultures. Normally, however, both populations establish a commonly shared norm and either end up with behavior 1 (green area in Fig. 1) or with behavior 2 (blue area).

Due to the payoff structure of the multi-population stag hunt game, 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. 1A,C,D). *Who* adapts to the preferred strategy of the other population essentially depends on the *initial* fractions of behaviors. The majority behavior 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. 1B+C).

Note that the evolution of norms is one of the most fundamental challenges in the social and economic sciences. Norms are crucial for society, as they reduce uncertainty, bargaining efforts, and 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 [18].

Researchers from various disciplines have worked on the evolution of norms, often utilizing game-theoretical concepts [19–23], but it has been hard to reveal the conditions under which behavioral consensus[24] is established. This is so, because cooperation norms, in contrast to coordination norms, are not self-enforcing. In other words, there are incentives for unilateral deviance. Considering the fact that norms require people to constrain self-interested behavior [25] and to perform socially prescribed roles, the ubiquity of norms is quite surprising. Yet, widespread cooperation-enhancing mechanisms such as group pressure can transform prisoner's dilemmas into stag hunt interactions [3, 17, 26] (see also Methods and SI [36]). This creates a natural tendency towards the formation of norms, whatever their content may be.

Our model sheds new light on the problem of whether, why and how a norm can establish. In particular, it reveals that the dynamics and finally resulting state of the system is not just determined by the payoff structure. It also crucially depends on the power of populations and even on the initial proportions of cooperative individuals (the "initial conditions").

B. Occurence of conflict in the snowdrift game

In the one-population *snowdrift game*, there is one stable stationary point, corresponding to a fraction $p_0 = |B|/(|B| + |C|)$ of cooperative individuals [17]. If this were transferable to the multi-population case, 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:

(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. 2A). 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 always end up with non-cooperative behavior by everybody (if $p_0 < 1/2$, see Fig. 2B), 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 a change of the model parameters may induce a "revolutionary" transition. 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 (see Methods), 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: In the weaker population 2, *everyone* shows behavior 1 for $p_0 < 1/2$ (see Fig. 2C), otherwise everyone shows behavior 2 (see Fig. 2D). 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 [27] 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 S1 and SI [36]). History and science [28] 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" offers a quantitative scientific approach to interpret such revolutions as the outcome of social interactions [29]. Here, their recurrence becomes understandable in a unified and simple game-theoretical framework.

III. DISCUSSION

Multi-population game-dynamical replicator equations provide an elegant and powerful approach to study the dynamics and outcomes expected for groups with incompatible interests. A detailed mathematical analysis reveals how interactions within and between groups can substantially change the dynamics of various game theoretical dilemmas. Generalizations to more than 2 behaviors or groups and to populations with heterogeneous preferences are easily possible.

When populations with incompatible preferences interact among and between each other, the signs of the payoff-dependent parameters B and C determine the character of the game. The snowdrift and stag hunt games show a particularly rich and interesting dynamics. 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 can be crucial.

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 can 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 face 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 disontinuous transitions in the system behavior can occur (see SI [36]): (1) the breakdown of cooperation, (2) in-group cooperation (the formation of "sub-cultures"), (3) societal polarization and conflict with the possibility of discontinuous regime shifts ("revolutions"), and (4) the evolution of shared behavioral norms.

In the past, the problem of the evolution of norms has often been addressed by studying one-population prisoner's dilemma situations [18, 21], and by assuming cooperation-enhancing mechanisms such as repeated interactions (a "shadow of the future", as considered in the mechanism of direct reciprocity) [1], or the sanctioning of non-conforming behavior ("punishment") [7–12, 23, 30, 31] These mechanisms can, in fact, transform prisoner's dilemmas into stag hunt games [20, 26] (see Methods and SI [36]), which connects our approach with previous work addressing norms. However, our model goes beyond studying the circumstances under which people follow a *preset* norm, it considers situations where it is not clear from the beginning what behavior would eventually establish as a norm (or whether any of the behaviors would become a norm *at all*). When studying multi-population settings with incompatible interests, we do not only have the problem of how *cooperative* behavior can be promoted, as in the prisoner's dilemma. We also have a *normative* dilemma by the circumstance that *the establishment of a behavioral norm requires one population to adjust to the preferred behavior in the other population—against its own preference.*

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* [17] (see SI [36]). Therefore, our findings suggest that 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 social conflict. Obviously, this has crucial implications for society, law and economics [16, 32], where conflicts need to be avoided or solved, and norms and standards are of central importance.

Take language as another example—probably the most distinguishing trait of humans [33, 34]. Successful communication requires the establishment of a norm, how words are used (the "evolution of meaning"). It will, therefore, be intriguing 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 due to human agglomeration in settlements, the development of reputation mechanisms, and the invention of sanctioning institutions should have largely accelerated cultural evolution [35].

Acknowledgments

The authors would like to thank for partial support by the EU Project QLectives and the ETH Competence Center "Coping with Crises in Complex Socio-Economic Systems" (CCSS) through ETH Research Grant CH1-01 08-2. They are grateful to Thomas Chadefaux, Ryan Murphy, Carlos P. Roca, Stefan Bechtold, Sergi Lozano, Heiko Rauhut, Wenjian Yu and further colleagues for valuable comments and to Sergi Lozano for drawing Fig. S3. D.H. thanks Thomas Voss for his insightful seminar on social norms.

Author Contributions: D.H. developed the concept and model of this study, did the analytical calculations and wrote the manuscript. A.J. performed the computer simulations, and prepared the figures and movies.

---

[1] Axelrod R (1984) *The Evolution of Cooperation* (Basic Books, New York).
[2] Gintis H (2000) *Game Theory Evolving* (Princeton University, Princeton, NJ).
[3] Nowak MA (2006) Five rules for the evolution of cooperation. *Science* 314, 1560–1563.
[4] Milinski M, Semmann D, Krambeck HJ (2002) Reputation helps solve the "tragedy of the commons". *Nature* 415, 424–426.
[5] Nowak MA, May RM (1992) Evolutionary games and spatial chaos. *Nature* 359:826-829.
[6] Fehr E, Gächter S (2002) Altruistic punishment in humans. *Nature* 415, 137–140.
[7] Hauert C, Traulsen A, Brandt H, Nowak, MA, Sigmund K (2007) Via freedom to coercion: The emergence of costly punishment. *Science* 316, 1905–1907.
[8] Rockenbach B, Milinski M (2006) The efficient interaction of indirect reciprocity and costly punishment. *Nature* 444, 718–723.
[9] Gurerk O, Irlenbusch B, Rockenbach B (2006) The competitive advantage of sanctioning institutions. *Science* 312, 108–111.
[10] Henrich J. *et al.* (2006) Costly punishment across human societies. *Science* 312, 1767–1770.
[11] Fowler JH (2005) Altruistic punishment and the origin of cooperation. *Proc Natl Acad Sci USA* 102, 7047–7049.

[12] Boyd R, Gintis H, Bowles S, Richerson PJ (2003) The evolution of altruistic punishment. *Proc Natl Acad Sci USA* 100, 3531–3535.
[13] Helbing D, Yu W (2009) The outbreak of cooperation among success-driven individuals under noisy conditions. *Proc Natl Acad Sci USA* 106(8), 3680–3685.
[14] 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.
[15] 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.
[16] Samuelson L (1998) *Evolutionary Games and Equilibrium Selection* (The MIT Press), Chap. 5: The Ultimatum Game.
[17] Helbing D, Lozano S (2009) Routes to cooperation and herding effects in the prisoner's dilemma game, e-print http://arxiv.org/abs/0905.3671.
[18] Axelrod R (1986) An evolutionary approach to norms. *American Political Science Review* 80(4), 1095–1111.
[19] Hechter M, Opp KD (Eds.) (2001) *Social Norms* (Russell Sage, New York), particularly Chap. 4 by Voss T, Game-theoretical perspectives on the emergence of social norms, pp. 105–136.
[20] Bicchieri C, Jeffrey R, Skyrms B (Eds.) (2009) *The Dynamics of Norms* (Cambridge University, Cambridge).
[21] Bendor J, Swistak P (2001) The evolution of norms. *American Journal of Sociology* 106(6), 1493–1545.
[22] Chalub FACC, Santos FC, Pacheco JM (2006) The evolution of norms. *J Theor Biol* 241, 233–240.
[23] Ostrom E (2000) Collective action and the evolution of social norms. *Journal of Economic Perspectives* 14(3), 137–158.
[24] Ehrlich PR, Levin SA (2005) The evolution of norms. *PLoS Biol* 3(6), 0943–0948.
[25] Keizer K, Lindenberg S, Steg L (2008) The spreading of disorder. *Science* 322, 1681–1685.
[26] Skyrms B (2003) *The Stag Hunt and the Evolution of Social Structure* (Cambridge University, Cambridge).
[27] Opp KD (2009) *Theories of Political Protest and Social Movements* (Routledge, London).
[28] Kuhn TS (1962) *The Structure of Scientific Revolutions* (University of Chicago, Chicago).
[29] Weidlich W, Huebner H (2008) Dynamics of political opinion formation including catastrophe theory. *Journal of Economic Behavior & Organization* 67, 1–26.
[30] Fehr E, Fischbacher U, Gächter S (2002) Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature* 13, 1–25.
[31] Whiten A, Horner, V, de Waal FBM (2005) Conformity to cultural norms of tool use in chimpanzees. *Nature* 437, 737–740.
[32] Binmore K (2005) *Natural Justice* (Oxford University, New York).
[33] Nowak MA, Komarova NL, Niyogi P (2002) Computational and evolutionary aspects of language. *Nature* 417, 611–617.
[34] Loreto V, Steels L (2007) Emergence of language. *Nat Phys* 3, 758–760.
[35] Boyd R, Richerson PJ (2005) *The Origin and Evolution of Cultures* (Oxford University, Oxford).
[36] Supporting Information can be sent upon request.

Methods: Multi-population game-dynamical replicator equations describe the temporal evolution of the proportions $p_i^a(t)$ of individuals showing behavior i at time t in population a. They assume that more successful behaviors spread, as these are imitated by individuals of the same population at a rate proportional to the gain in the expected success. The expected success is determined from the frequency of interactions between two behaviors i and j, and by the associated payoffs A_{ij}^{ab} (see [36]). 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, respectively. 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.

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 $q(t) = p_2^2(t)$ of cooperative individuals in population 2 are given by Eqs. [1] and [2] with

$$F(p, q) = b_1 f + (c_1 - b_1) f p(t) + C_1(1 - f) + (B_1 - C_1)(1 - f) q(t) \tag{3}$$

and

$$G(p, q) = b_2(1 - f) + (c_2 - b_2)(1 - f) q(t) + C_2 f + (B_2 - C_2) f p(t) \tag{4}$$

(see [36]). Here, we have used the abbreviation $f = f_1 = 1 - f_2$. Moreover, $b_a = s_a - p_a$, $B_a = S_a - P_a$, $c_a = r_a - t_a$, and $C_a = R_a - T_a$ are payoff-dependent model parameters, which can be positive, negative, or zero. Note that the above equations describe the *general* case for two populations with interactions and/or self-interactions playing *any* kind of 2×2 game. When setting $p_a = P_a = P$, $r_a = R_a = R$, $s_a = S_a = S$, and $t_a = T_a = T$ (i.e. $b_a = B_a = B$ and $c_a = C_a = C$), both populations play the same game. Moreover, 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. That is, in- and out-group interactions yield the same payoff, and the preference of the interaction partner does not matter for it.

Sanctioning of non-conforming behavior. Individuals often apply *group pressure* to support conformity and discourage uncoordinated behavior. This could be modeled by subtracting a value δ from the off-diagonal payoffs S and T or by adding δ to the diagonal elements R and P, resulting in the effective model parameters $b_a = B_a = B - \delta$ and $c_a = C_a = C + \delta$. 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 $c_a = C_a > 0$.

Summary of the main analytical results (see [36] for more general formulas). All of the

multi-population games with interactions and self-interactions studied by us have the four stationary solutions $(p_1, q_1) = (0, 0)$, $(p_2, q_2) = (1, 1)$, $(p_3, q_3) = (1, 0)$ and $(p_4, q_4) = (0, 1)$, corresponding to the four corners of the p-q-space. Their stability properties depend on the eigenvalues $\lambda_l = (1 - 2p_l)F(p_l, q_l)$ and $\mu_l = (1 - 2q_l)G(p_l, q_l)$, where $l \in \{1, 2, 3, 4\}$. Stable (attractive) fix points require $\lambda_l < 0$ and $\mu_l < 0$, unstable (repulsive) fix points $\lambda_l < 0$ and $\mu_l < 0$. $\lambda_l \mu_l < 0$ implies a saddle point. In the multi-population prisoner's dilemma (MPD), the only stable fix point is $(0, 0)$, while in the multi-population harmony game (MHG), it is $(1, 1)$. In both games, $(1, 0)$ and $(0, 1)$ are always saddle points. For $B, C < 0$ (the MPD) and $B, C > 0$ (the MHG), no further stationary points exist (see [36], Fig. S4). For the multi-population stag hunt game (MSH) with $B < 0$ and $C > 0$, we find:

– (0,1) and (1,0) are always stable fix points, see Fig. 1.

– (1,1) is a stable fix point for $C/|B| > \max[f/(1-f), (1-f)/f]$ (see Fig. 1B).

– (0,0) is a stable fix point for $C/|B| < \min[f/(1-f), (1-f)/f]$.

For the multi-population snowdrift game (MSD) with $B > 0$ and $C < 0$ we have:

– (1,0) and (0,1) are always unstable fix points (see Fig. 2).

– (0,0) is a stable fix point for $C/|B| > \max[f/(1-f), (1-f)/f]$ (see Fig. 2B).

– (1,1) is a stable fix point for $C/|B| < \min[f/(1-f), (1-f)/f]$.

Moreover, if B and C have different signs, further stationary points (p_l, q_l) with $l \in \{5, 6, 7, 8\}$ may occur on the boundaries, while inner points (p_9, q_9) with $0 < p_9 < 1$ and $0 < q_9 < 1$ can only occur for $B = -C$ (see Figs. 1A and 2A). As $|C|$ is increased from 0 to high values, we find the following additional stationary points for the MSH and MSD, where we use the abbreviations $p_5 = [Bf + C(1-f)]/[(B-C)f]$, $p_6 = B/[(B-C)f]$, $q_7 = B/[(B-C)(1-f)]$, and $q_8 = [B(1-f) + Cf]/[(B-C)(1-f)]$:

– $(p_5, 0)$ and $(0, q_8)$, if $f \geq 1/2$ and $|C|/|B| \leq (1-f)/f$ or if $f \leq 1/2$ and $|C|/|B| \leq f/(1-f)$.

– $(p_5, 0)$ and $(p_6, 1)$, if $f \geq 1/2$ and $(1-f)/f < |C|/|B| < f/(1-f)$, or $(1, q_7)$ and $(0, q_8)$ if $f \leq 1/2$ and $f/(1-f) < |C|/|B| < (1-f)/f$ (see Figs. 1C+D and 2C+D).

– $(p_6, 1)$ and $(1, q_7)$, if $f \geq 1/2$ and $|C|/|B| \geq f/(1-f)$ or if $f \leq 1/2$ and $|C|/|B| \geq (1-f)/f$ (see Figs. 1B and 2B).

For $B < 0 < C$ (the MSH), these fix points are unstable or saddle points, while they are stable or saddle points for $C < 0 < B$ (the MSD). Obviously, there are transitions to a qualitatively different system behavior at the points $|C|/|B| = (1-f)/f$ and $|C|/|B| = f/(1-f)$. Moreover, there are discontinuous, "revolutionary" transitions, when $|C|$ crosses the value of $|B|$, as the stability properties of pairs of fix points are then *interchanged* (see Movie S3). This can be followed from the fact that the dynamic behavior and final outcome for the case $|B| > |C|$ can be derived from the results for $|B| < |C|$, namely by applying the transformations $B \leftrightarrow -C$, $p \leftrightarrow (1-p)$, and $q \leftrightarrow (1-q)$, which do not change the game-dynamical equations (see Eqs. [1] to [4]). Generally, *discontinuous transitions in the system behavior may occur, when the sign of* $1 - |B|/|C|$ *changes, or if the sign of* B *or*

C changes (which modifies the character of the game, for example from a MPD to a MSH game).

Fraction of cooperators in the multi-population snowdrift game. When $(p_5, 0)$ is the stable stationary point, the *average* fraction of cooperative individuals in *both* populations from the perspective of the stronger population 1 can be determined as the fraction of cooperative individuals in population 1 times the relative size f of population 1, plus the fraction $1 - q_5 = 1$ of non-cooperative individuals in population 2 (who are cooperative from the point of view of population 1), weighted by its relative size $(1 - f)$:

$$p_5 \cdot f + (1 - q_5) \cdot (1 - f) = \frac{Bf + C(1 - f)}{(B - C)f} \cdot f + 1 \cdot (1 - f) = \frac{B}{B - C} \,.$$

Considering $C < 0$, this corresponds to the expected fraction $p_0 = |B|/(|B| + C)$ of cooperative individuals in the one-population snowdrift game [17].

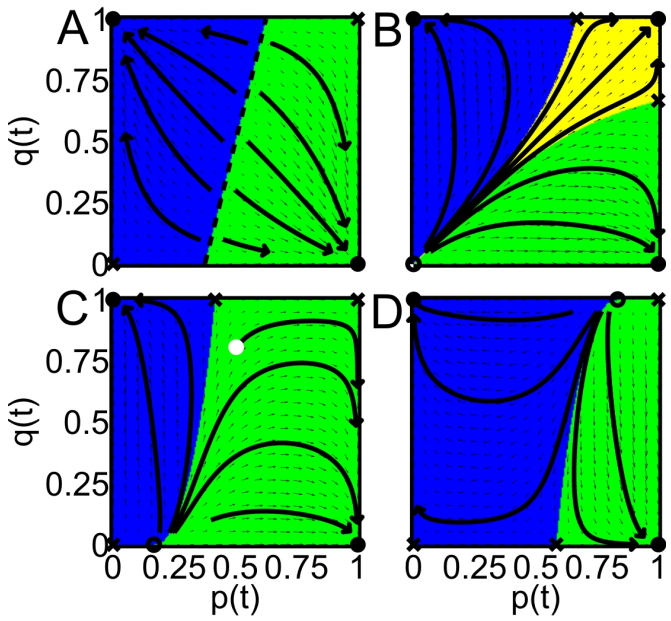

FIG. 1: Vector fields (small arrows), sample trajectories (large arrows) and phase diagrams (colored areas) for two interacting populations with incompatible preferences, playing stag hunt games with $B < 0$ and $C > 0$ (right). p is the fraction of individuals in population 1 showing their preferred, cooperative behavior 1, and q is the fraction of cooperative individuals in population 2 showing their preferred behavior 2. 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 or dashed lines) and are attracted towards stable stationary points (black circles or solid diagonal lines). The colored areas represent the basins of attraction, i.e. all initial conditions $(p(0), q(0))$ leading to the same fix point [yellow = (1,1), blue = (0,1), green = (1,0)]. Saddle points (crosses) are attractive in one direction, but repulsive in another. The model parameters are as follows: (A) $|B| = |C| = 1$ and $f = 0.8$, i.e. 80% of all individuals belong to population 1, (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 $f = 0.8$. In the multi-population stag hunt game (MSH), 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 grows in the beginning, but later on it goes down dramatically. Therefore, a momentary trend does not allow to easily predict the final outcome of the struggle between two interest groups.

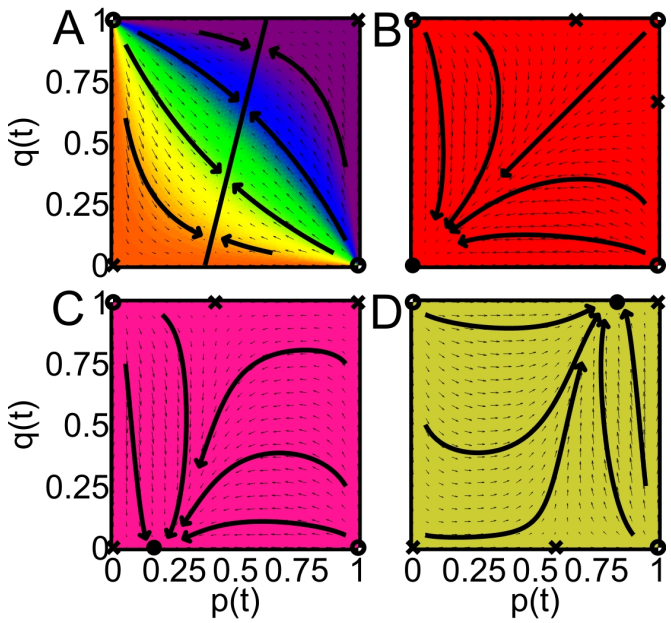

FIG. 2: Vector fields (small arrows), sample trajectories (large arrows) and phase diagrams (colored areas) for two interacting populations with incompatible preferences, playing snowdrift games with $B > 0$ and $C < 0$ (right). The representation is the same as in Fig. 1. In particular, the colored areas represent again the basins of attraction, i.e. all initial conditions $(p(0), q(0))$ leading to the same 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. 80% of all individuals belong to population 1, (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 solidarizes 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 occurence of a discontinuouos ("revolutionary") transition of the evolutionary equilibrium from $(u, 0)$ to $(v, 1)$.