# Coevolution of actions, personal norms, and beliefs about others in social dilemmas

Sergey Gavrilets

Department of Ecology and Evolutionary Biology, Department of Mathematics, National Institute for Mathematical and Biological Synthesis, Center for the Dynamics of Social Complexity, University of Tennessee, Knoxville, TN 37996 USA

April 3, 2021

**Abstract**

Human decision-making is affected by a diversity of factors including material cost-benefit considerations, normative and cultural influences, learning, and conformity with peers and external authorities (e.g., cultural, religious, political, organizational). Also important are their dynamically changing personal perception of the situation and beliefs about actions and expectations of others as well as psychological phenomena such as cognitive dissonance, and social projection. To better understand these processes, I develop a modeling framework describing the joint dynamics of actions and attitudes of individuals and their beliefs about actions and attitudes of their groupmates. I consider which norms get internalized and which factors control beliefs about others. I predict that the long-term average characteristics of groups are largely determined by a balance between material payoffs and the values promoted by the external authority. Variation around these averages largely reflects variation in individual costs and benefits mediated by individual psychological characteristics. The efforts of an external authority to change the group behavior in a certain direction can, counter-intuitively, have an opposite effect on individual behavior. I consider how various factors can affect differences between groups and societies in tightness/looseness of their social norms. I show that the most important factors are social heterogeneity, societal threat, effects of the authority, cultural variation in the degree of collectivism/individualism, the population size, and the subsistence style. My results can be useful for achieving a better understanding of human social behavior, historical and current social processes, and in developing more efficient policies aiming to modify social behavior.

## Introduction

Human groups at various scales of social organization repeatedly face situations when engaging in an individually costly collective action or refraining from an individually beneficial behavior can help bring larger benefits or avoid certain disastrous outcomes. Examples range from cooperating in hunting or agricultural production in small-scale societies to mobilizing against social injustice to modifying collective behavior of the population to stop a pandemic or decrease global warming. Such situations commonly lead to social dilemmas when individual and group interests come into a conflict. In the scientific literature, they come under various names including collective action problem (1, 2), the tragedy of the commons (3, 4), social traps (5), many-person Prisoner's Dilemma (6, 7), and collective risk dilemma (8).

Human decision-making in social dilemmas is affected by a diversity of factors including genetically-informed biological instincts, material cost-benefit considerations, normative and cultural influences, and conformity with peers or external authorities (e.g., cultural, religious, political, organizational). Human actions also depend on their personal perception of the situation and on beliefs about actions and expectations of their peers. The beliefs and expectations can change as a result of learning and other psychological processes. For example, cognitive dissonance (i.e. a feeling of mental discomfort experienced when the person's attitudes, beliefs, or behaviors conflict) can cause changes in behaviors but also in attitudes or beliefs (9). To predict the intents and beliefs of others, people may use the "theory of mind" (10, 11) and social projection, which is the tendency to assume that others are similar to oneself (12). Therefore changing personal attitudes can also change predictions about others.

Due to this complexity, modeling human behavior is notoriously difficult. Nevertheless several approaches successfully capturing certain aspects of human decision-making have been developed. These include classical (13), evolutionary (14), mean-field (15), and quantum (16, 17) game theories, social influence models focusing on the dynamics of consensus formation (or fragmentation) in social networks as a result of social learning and imitation (18–24), models of strategic deliberation (25), models of normative behavior (26–28), and models of foresight (29, 30). I will build on this earlier work to develop a novel theoretical approach explicitly integrating multiple material, cognitive and social forces shaping human behavior.

I posit that individuals are motivated by both material factors and immaterial values and norms, that their actions are driven by their interpretation of what they observe, and that their interpretations and beliefs change dynamically as social interactions unfold. My approach allows for irrationality and captures some emotional and cognitive processes inherent to humans. Moreover I explicitly allow for and study the effects of differences between individuals in various physical and psychological characteristics as well as in their attitudes and beliefs. Starting with individual-level social and psychological processes affecting agents' intent and beliefs under information uncertainty I aim to predict group dynamics. I will also consider which norms get internalized and which factors control beliefs about others.

My starting point is what is known in social psychology as the "Thomas theorem" which states that "If men define situations as real, they are real in their consequences" (31). In other words, our actions often depend on our interpretation of a situation rather than on its objective reality. In my models, I will capture this "theorem" by postulating that individual decisions in social situations are based on individual beliefs about the current situation as well as beliefs about others and their beliefs. Individuals will revise their actions, attitudes, and beliefs according to not only the information they receive but also according to some psychological processes governing their thinking and emotions (32, 33). The general structure of my model is illustrated in Figure 1.

Below after introducing my approach and describing main results, I illustrate them by considering different types of social interactions including those stylized by Coordination, Public Goods, Tragedy of the Commons, Common Pool Resource, continuous Prisoner's Dilemma, Dictator and "Us vs. nature" games.

## Model

I consider a group of people repeatedly engaged in a particular type of social interaction. For example, individuals can contribute efforts to a joint production or maintenance of a public good (e.g., an irrigation canal) or harvest from a common pool of resources (e.g., fishing from a pond). Individuals care about their own material costs and benefits. They do not like to be disapproved

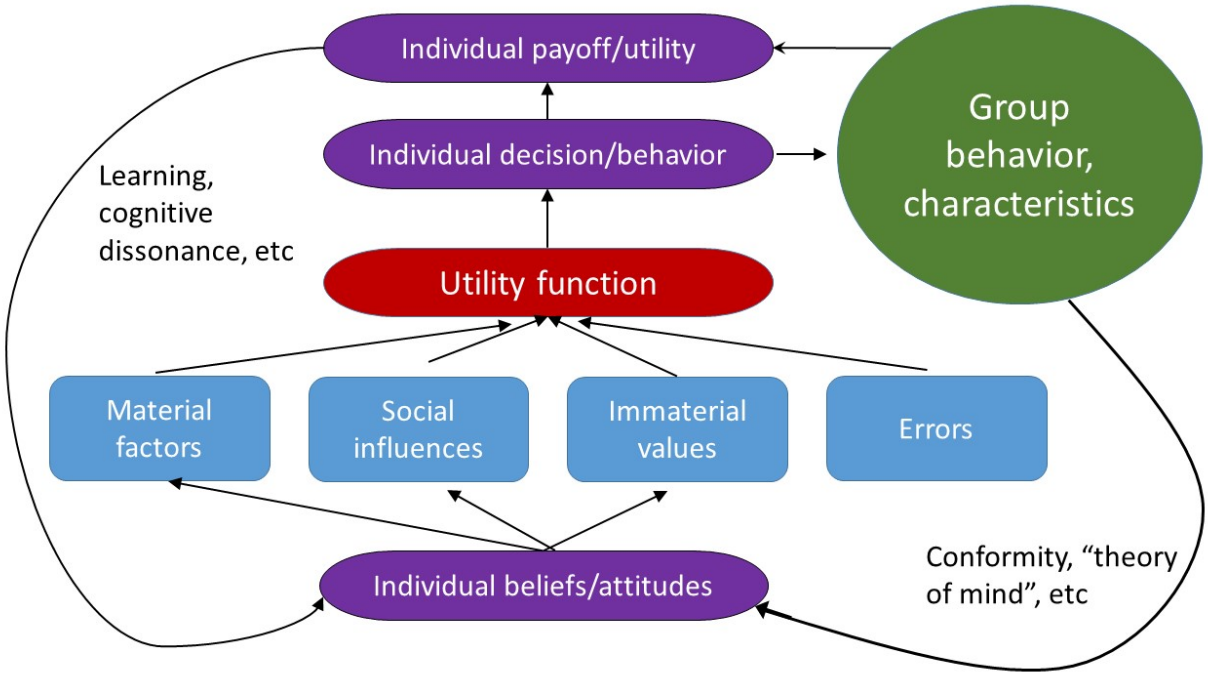



**Figure 1:** Model structure. The model integrates material factors, social influences (both by peers and an external authority), immaterial values, and errors (the blue boxes) into a general utility function (the red shape) which individuals attempt to maximize when making decisions (the middle violet shape). Individual behavior is a part of group behavior (the green shape) which together define the actual payoff/utility of the individual (top violet shape). The latter together with the observed group behavior feed back into individual beliefs and attitudes (bottom violet shape). In my approach, the strength of various factors, as perceived by individuals, will vary between them depending on the information available as well as on the individual's attitudes and beliefs. My approach allows for attitudes and beliefs to (rapidly) change in time as a consequence of different actions taken by individuals and the groups they belong to, of the information they receive, and the emotions they experience.

by peers (or an external authority) but they also prefer to do what they personally think is appropriate. Individuals observe (and learn from) the actions of others and make inferences about others' attitudes (preferences) and beliefs but they do not know them exactly. How can they find the right action? What happens to their preferences, beliefs, and behaviors as social interactions dynamically unfold?

I will treat time as discrete. Let a continuous variable $x$ specify an action chosen by a focal individual. Each individual is characterized by an attitude $y$ which gives his personal belief about the most appropriate action in a given social situation. Each individual also has a belief (an expectation) $\tilde{x}$ about the average action of peers as well as a second order belief $\tilde{y}$ about the average attitude of their peers. Experiments show that people represent the preferences and beliefs of others separately from their own (34, 35). In the social psychology literature, variables $y, \tilde{x}$, and $\tilde{y}$ would be called a personal norm (or value), an empirical expectation, and a normative expectation, respectively (36–38). I will occasionally use this terminology below. Individuals are also subject to influence by an external authority promoting a particular action $G$. I assume $x, y, \tilde{x}, \tilde{y}, G$ are nonnegative. Individuals form their beliefs about others on the basis of their actions they observe and some cognitive and psychological processes (which I discuss below).

**Utility function**. I postulate that each individual chooses (via myopic best response) an action

$x$ in an attempt to maximize the utility function $u$. I write it as a sum of several terms:

$$u = \underbrace{\pi(x,\tilde{x})}_{\text{material payoff}} - \underbrace{\frac{1}{2}\,A_1(x-y)^2}_{\text{cognitive dissonance}} - \underbrace{\frac{1}{2}\,A_2(x-\tilde{y})^2}_{\text{disapproval by peers}} - \underbrace{\frac{1}{2}\,A_3(x-\tilde{x})^2}_{\text{conformity w/ peers}} - \underbrace{\frac{1}{2}\,A_4(x-G)^2}_{\text{conformity w/ authority}}. \quad (1)$$

The first term in equation (1) specifies a material payoff to a focal individual performing action $x$ under the expectation that his peers' average action is $\tilde{x}$. The second term in equation (1) captures the psychic costs due to cognitive dissonance (9) incurred when the action $x$ chosen deviates from the personal norm $y$. The third term captures the expected psychic costs of disapproval (or material costs of punishment) by others who are expected to have expectation $\tilde{y}$ regarding the behavior of the focal individual (39, 36). The fourth term in equation (1) captures the psychic costs of nonconformity with the expected actions of others (40, 41). For example, the fact that peers choose a particular action may indicate that this action is most beneficial. So acting differently may cause additional psychic costs not related to disapproval or punishment by peers (captured by the third term). The last term in equation (1) captures the expected costs of material punishment or psychic costs of disapproval by the external authority promoting an action at a "standard" level $G$ which I will treat as a constant(42, 40). Some studies show stable variation between people in following the "rules" (43, 44).

I assume parameters $A_1, A_2, A_3$ and $A_4$ are non-negative individual-specific constants. This assumption aims to capture the fact that people differ in their personalities, cultural background, and other characteristics affecting their emotions, feelings, psychology and behavior. Parameters $A_2$ and $A_3$ may depend on the group size, so that individuals whose actions deviate from the expected behavior or beliefs of others suffer bigger costs in larger groups. Parameter $A_4$ may depend on the degree of legitimacy of the external authority and on individual self-identification.

My approach is particularly simple when the function $\pi(x,\tilde{x})$ specifying the material payoff is a linear, quasi-linear, or a quadratic function of $x$ and $\tilde{x}$. For such cases, the first derivative of $\pi(x,\tilde{x})$ (i.e., marginal payoff) with respect to $x$ is a linear function of $x$ and $\tilde{x}$, which I will write as

$$\frac{\partial \pi(x,\tilde{x})}{\partial x} = D_0 - D_1\tilde{x} - D_2 x, \quad (2)$$

where $D_0, D_1$ and $D_2$ are constant individual-specific parameters. For example, individuals may differ in their strengths, valuation (or shares received) of the collectively produced goods, costs, or availability of information regarding the material consequences of the game. [For simplicity of notation, for now I do not use explicitly any indices in the equations to specify the individual. This will change later when I discuss specific social situations and games.]

Below I will use a composite parameter of the material payoff function

$$\theta = \frac{D_0}{D_1 + D_2}, \quad (3)$$

which can be interpreted as the best response action for a focal individual who believes that the average action of his social partners will always match his own action (i.e. $\tilde{x} = x$). In several games to be considered below, $\theta$ can also be viewed as a measure of the material benefit-to-cost ratio; in some games $\theta$ is the Nash equilibrium for the individual effort. As I show below, the distribution

of $\theta$ in the society strongly affects the long-term dynamics of the model. When I use agent-based simulations, I will also allow for errors in decision-making.

**Best response action**. The action $x$ maximizing the utility function $u$ of the focal individual can be found by computing the derivative $\frac{\partial u}{\partial x}$. Since $u$ is a quadratic function, the best response action given an attitude $y$ and beliefs $\tilde{x}$ and $\tilde{y}$ can be found in a straightforward way. I will write it as

$$x = \max(0, B_0 + B_1 y + B_2 \tilde{y} + B_3 \tilde{x} + B_4 G), \tag{4}$$

where $B_0, \ldots, B_4$ are re-scaled individual-specific parameters measuring the effects of material and immaterial forces on individual actions (see the Supporting Information, SI). I assume that all individuals in the group take their own best response actions simultaneously.

**The dynamics of attitudes and beliefs**. After taking their own action and observing the actions of their groupmates, each individual revises their attitudes and beliefs. To capture these changes, I adapt an approach standard in social influence models describing the dynamics of publicly expressed opinions. Specifically I postulate that attitudes and beliefs of a focal individual change according to a system of linear recurrence equations:

$$y' = y + \underbrace{C_{11}(x-y)}_{\text{cognitive dissonance}} + \underbrace{C_{12}(X-y)}_{\text{conformity w/ peers}} + \underbrace{C_{13}(G-y)}_{\text{conformity w/authority}}, \tag{5a}$$

$$\tilde{y}' = \tilde{y} + \underbrace{C_{21}(y-\tilde{y})}_{\text{social projection}} + \underbrace{C_{22}(X-\tilde{y})}_{\text{learning about others}} + \underbrace{C_{23}(G-\tilde{y})}_{\text{conformity w/ authority}}, \tag{5b}$$

$$\tilde{x}' = \tilde{x} + \underbrace{C_{31}(\tilde{y}-\tilde{x})}_{\text{logic constraints}} + \underbrace{C_{32}(X-\tilde{x})}_{\text{learning about others}} + \underbrace{C_{33}(G-\tilde{x})}_{\text{conformity w/ authority}}, \tag{5c}$$

where the prime means the next time step, $X$ is the average action of groupmates as observed by the focal individual (so that different individuals are characterized by different $X$), and $C_{ij}$ are non-negative individual-specific constant coefficients. Here the "cognitive dissonance" term acts to reduce the mismatch of the ego's actions and their beliefs about themselves. The "social projection" term captures the ego's believe that others are probably similar to themselves (10, 12). The "logic constraints" term reduces a mismatch between the ego's beliefs about actions and beliefs of others (cf., Ref.(20)). The "conformity w/ peers" and two "learning about others" terms move the corresponding attitude and beliefs closer to the observed average behavior $X$ among peers(45). The "conformity w/ authority" terms move the corresponding attitudes and beliefs closer to the promoted "standard" $G$. Note that cognitive dissonance makes individuals to choose action $x$ closer to their attitude $y$ (as implied by equation 1) and simultaneously changes their attitude $y$ to justify the action previously chosen (as described by the first term in equation (5a) (cf., Ref.(46)). The authority effectively changes the utility function (1) and simultaneously affects attitudes and beliefs (equations 5) which then feed back into the utility function and behavior. For a group of $n$ individuals I thus end up with $3n$ recurrence equations of type (5) which are coupled via terms $X$ which are the observed average actions of groupmates. Below in deriving analytical approximations I will assume that $n$ is sufficiently large so that individual values $X$ are approximately the same (and equal to the actual average action of the group).

Below I will use normalized parameters $\alpha_i = \frac{C_{i1}}{\sum_j C_{ij}}$, $\beta_i = \frac{C_{i2}}{\sum_j C_{ij}}$, $\gamma_i = \frac{C_{i3}}{\sum_j C_{ij}}$, with $\alpha_i + \beta_i + \gamma_i = 1$ for all $i$. Parameters $\alpha_i$ characterize the relative strengths of cognitive factors (i.e., related to the cognitive dissonance, the social projection, and the logic constraint, respectively). Parameters $\beta_i$ and $\gamma_i$ characterize the relative strengths of two types of social factors: learning from/about peers and complying with external influences, respectively. All these coefficients are individual-specific;

they may depend on individual psychology, cultural and education background, etc. They may also depend on social and cultural factors acting in the group. For example, increased efforts to promote certain ideas by an authority may translate in increased values of parameters $\gamma_i$ while strongly conformist or collectivistic communities may be characterized by higher values of parameters $\beta_i$. Parameters $B_4$ and $\gamma_i$ can depend on trust in the authority and its legitimacy. Intuitively, cognitive factors work to align individual actions, attitude and beliefs, learning from/about peers works to align those between individuals, while external influence works to shift them towards a promoted standard.

Before proceeding further it is instructive to compare my approach with already existing models. First, classical, evolutionary, and mean-field game-theoretic models focus exclusively on the material payoff component $\pi(x)$ of the utility model disregarding all other terms (13–15, 47). Note also that in contrast to standard game-theoretic models where individuals choose best responses to the previous action of their mates which they know exactly, in my approach they best respond to their expectation $\tilde{x}$ of the action of their group-mates in this round. Some game-theoretic models add a normative component to the utility function but treat personal norms $y$ as constant (26–28). Few existing models consider the joint dynamics of actions ($x$) and personal norms ($y$). For example in Refs.(46, 48, 49), utility functions include material payoffs $\pi(x)$ as well cognitive dissonance and conformity with peers terms. Refs.(48, 49) describe the dynamics of personal norms $y$ allowing for the effects of cognitive dissonance and conformity with peers. However these authors assume that individuals know exactly the personal norms $y$ of their peers which in general is not realistic. There is also a very large number of social influence models (18–22, 24, 50) which consider the dynamics of personal attitudes and opinions $y$ as a result of the exchange of opinions between group members (using linear equations related to the second and third terms in equation 5a). The linear equations describing the changes in attitudes and beliefs are also related to those used in cognitive neuroscience (51). Focusing on dyadic interactions, Ref. (25) models how individuals update their values of $y$ and $\tilde{x}$ on the basis of payoffs received. Ref. (52) considered similar models but with addition of an external influence (described by a term analogous to the last term in equation 5a). Models of social influence neglect material factors, and explicitly assume that players know exactly the opinions of their peers. None of all these models consider second order beliefs of individuals captured by variables $\tilde{y}$ and $\tilde{x}$.

I note that the model's structure reflects the facts that human behavior and beliefs are complex phenomena and that real people differ in their psychology and behavior. As I show below, in spite of its apparent complexity, the model's behavior is quite tractable, its parameters combine into a small number of effective measures controlling the equilibria, and individual parameters can be estimated using behavioral economics' methods or surveys.

# Results

## Long-term behavior

Equations (4-5) describe the joint dynamics of actions ($x$), attitudes ($y$), and beliefs ($\tilde{y}, \tilde{x}$). Numerical iterations of these equations show convergence to a stochastic equilibrium (see Figure 2 for an example to be considered in detail below). In the SI I find an approximation for this equilibrium. Here I summarize what happens in several important special cases. For the rest of my paper, variables $x, y, \tilde{y}, \tilde{x}$, and $X$ will specify the corresponding *equilibrium* values (rather than the dynamically changing values as above).

*No external influence; no variation in material payoffs.* Assume that the external influence is absent (i.e., $A_4 = C_{i3} = \gamma_i = 0$ for all $i$) and that there is no variation in material payoffs between

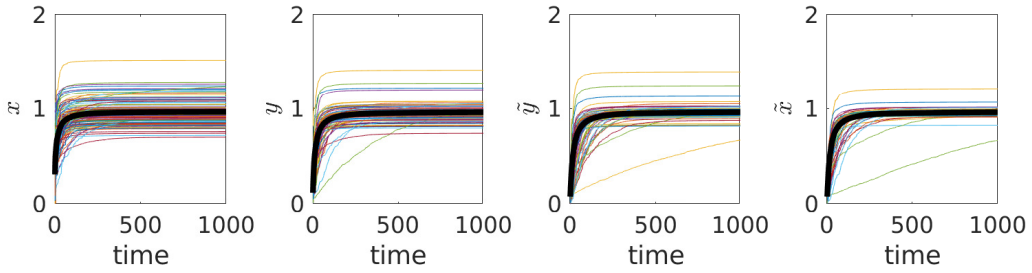



**Figure 2:** The dynamics of $x, y, \tilde{y}$ and $\tilde{x}$ of individual players in the Coordination Game with no external influence observed in a single run of agent-based simulations. The thick black lines show the group averages. Group size $n = 100$. Parameters are chosen randomly and independently from certain distributions (as described in the SI) so that the mean value of $\theta$ is equal to 1. Initial values of $y, \tilde{y}$ and $\tilde{x}$ are chosen randomly and independently from a uniform distribution on $[0, 0.1]$.

individuals (so that coefficient $\theta$ is the same for all individuals). Then the system evolves to an equilibrium at which

$$x = y = \tilde{y} = \tilde{x} = \max(0, \theta) \tag{6}$$

for all individuals. That is, with no variation in material costs and benefits, the population eventually becomes homogeneous in actions, attitudes, and beliefs independently of the differences between individuals in all other parameters (i.e., $A_i, \alpha_i, \beta_i$). The value of $x$ at equilibrium is the one maximizing the material payoff.

*External influence only.* If there are no material payoffs in the utility function (i.e., if all $D_i = 0$) while the external authority promotes action $G$, then at a long-term equilibrium

$$x = y = \tilde{y} = \tilde{x} = G \tag{7}$$

for each individual. That is, the population's actions, attitudes, and beliefs are completely determined by the external influence and there is no variation between individuals.

*No external influence; variation in material payoffs.* With variation in material benefits and costs between individuals (which is present in any realistic situation), one finds that the system evolves to an equilibrium state at which the average action

$$X \approx \overline{\theta}. \tag{8a}$$

[Here and below the bar means the average over the whole population.] That is, at equilibrium the average action is the average of individual $\theta$'s which depend only on material payoffs. I also find that at equilibrium for each individual

$$x \approx X + \eta\ (\theta - \overline{\theta}), \tag{8b}$$

$$y \approx X + \alpha_1 \eta\ (\theta - \overline{\theta}), \tag{8c}$$

$$\tilde{y} \approx X + \alpha_1 \alpha_2 \eta\ (\theta - \overline{\theta}), \tag{8d}$$

$$\tilde{x} \approx X + \alpha_1 \alpha_2 \alpha_3 \eta\ (\theta - \overline{\theta}). \tag{8e}$$

A composite parameter $\eta$, which depends on $B$'s and $\alpha$'s, is defined by equation (S4c) in the SI. Parameters $\theta, \alpha_1, \alpha_2, \alpha_3$ and $\eta$ are individual-specific while $X$ and $\overline{\theta}$ are the same for all individuals.

With no cognitive dissonance (i.e. if $\alpha_1 = 0$), $y = \tilde{y} = \tilde{x} = X$, so that, the society becomes homogeneous in attitudes and beliefs while still exhibiting variation in actions $x$. Without the

"theory of mind" (i.e. if $\alpha_2 = 0$), $\tilde{y} = \tilde{x} = X$, so that, the society becomes homogeneous in beliefs while still exhibiting variation in actions $x$ and attitudes $y$. Without logic constraints (i.e. if $\alpha_3 = 0$), $\tilde{x} = X$, so that there will be no variation in second order $\tilde{x}$ beliefs about actions. Note that if the correlation between $\theta, \eta$ and the strength of cognitive factors $\alpha_1, \alpha_2, \alpha_3$ are low, the mean values of $x, y, \tilde{y}$ and $\tilde{x}$ are all approximately equal to $\overline{\theta}$. That is, on average individual preferences and beliefs align with actions.

One can also approximate the corresponding variances. These approximations show (see the SI) that at equilibrium

$$\mathrm{var}(x) > \mathrm{var}(y) > \mathrm{var}(\tilde{y}) > \mathrm{var}(\tilde{x}). \tag{9}$$

That is, the model predicts that the variation in actions will be the largest, followed by the variation in personal norms, followed by the variation in beliefs about norms of others, followed by the variation in beliefs about the action of others. Similarly, the correlation with material benefits (characterized by parameter $\theta$) will be the highest for individual actions $x$, followed by personal beliefs $y$, followed by normative expectations $\tilde{y}$, and empirical expectations $\tilde{x}$ (see the SI). These are testable predictions. I will illustrate these results below when considering specific social interactions.

## Examples

Next I illustrate my results using several games which have been extensively studied using methods of classical game theory, evolutionary game theory, and behavioral economics. In experimental studies, the subjects are usually identical in terms of the expected costs and benefits of their actions. In contrast, in real life there is usually a lot of variation between individuals in these factors. Consequently, I will consider a group of $n$ individuals who differ in various relevant characteristics such as their costs, benefits and/or valuation of the resource produced. (See Ref.(53) for a review of models of collective action in heterogeneous groups.) I will also allow for differences between individuals in parameters characterizing the effects of immaterial factors.

In agent-based simulations, I will assign parameters $D_i, A_i$ of the utility function $u$ and parameters $C_{ij}$ specifying the dynamics of attitudes and beliefs randomly and independently from certain distributions. In my graphs, I will use an additional parameter $\varepsilon$ which will vary from 0 to 1. I will scale parameters $A_1, \ldots, A_4$ by multiplying them by $\varepsilon$. For example, with $\varepsilon = 0$ any normative effect in the utility function will be absent and individuals will behave according to standard evolutionary game theory assumptions. In contrast with $\varepsilon = 1$, the expected weight of each term in the utility function will be the same. Individuals will revise their actions and beliefs with probability 50% per individual per time step. I will also introduce small random errors during the update processes. I will compute the means and standard deviations of my variables at a long-term equilibrium, the Kendall rank correlation between them and $\theta$, and the half-time $\tau$ of convergence to an equilibrium (defined as the time to reduce the distance to an equilibrium value by one half). My main focus will be on games with quadratic payoffs functions. However in the SI, I also consider several models with linear and quasi-linear payoff functions and a more complex example of a nonlinear payoff function. Table S1 in the SI summarizes the games I consider.

**Coordination Game**. Let individuals interact in randomly formed groups. Adapting the model in Ref. (48) (see also Ref. (54)), assume that each individual has a preferred action $\theta_i$ and pays a cost proportional to the square of the deviation from $\theta_i$. Each player also pays a cost if his action deviates from the average action of the group. The corresponding (subjective) payoff function for individual $i$ is

$$\pi(x_i, \tilde{x}_i) = b_i - 0.5c_i(x_i - \theta_i)^2 - 0.5d_i(x_i - \tilde{x}_i)^2, \tag{10}$$

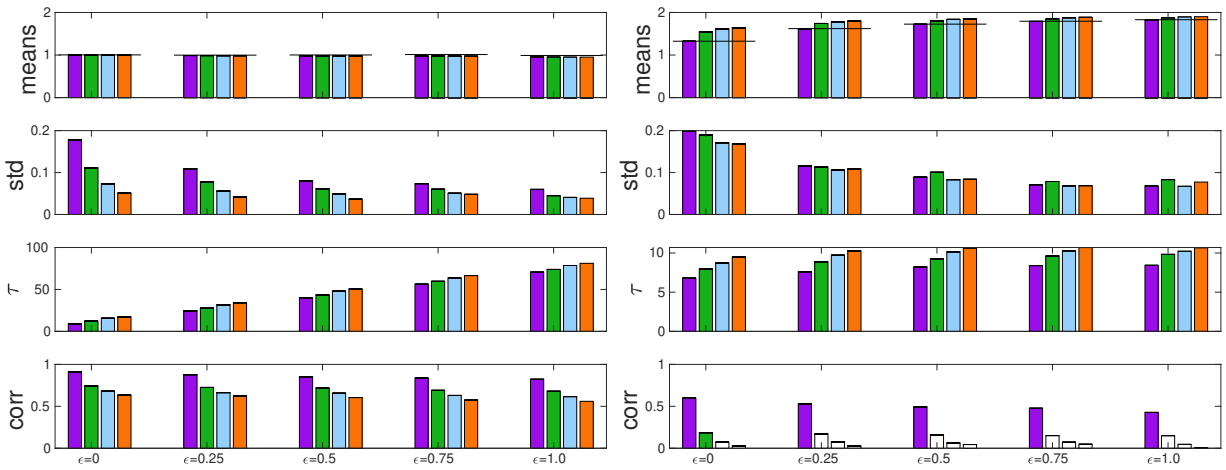



**Figure 3:** Properties of equilibria in the Coordination Game. (a) No external influence. (b) With external influence ($G = 2$). From top to bottom: mean, standard deviation, half-time of convergence to an equilibrium $\tau$, and Kendall rank correlation with $\theta$ for $x$ (purple), $y$ (green), $\tilde{y}$ (blue) and $\tilde{x}$ (orange), respectively. Bars with no color mean the corresponding correlations are statistically insignificant (at 0.05). The thin black lines show the theoretical predictions for $x$ (given by equation S6 in the SI). Notice the difference between $y$-axis scales in graphs for $\tau$.

Parameter $\varepsilon$ measures the weight of each normative factor relative to material payoffs in the utility function. Group size $n = 100$. Parameters $\theta_i, c_i, d_i$ are drawn from lognormal distributions with mean 1 and standard deviation 0.1, so that $\overline{\theta} \approx 1$. Statistics are calculated over 100 last time steps over 40 independent runs each of length 1,000 time steps.

where parameter $b_i$ is the maximum benefit, and $c_i$ and $d_i$ are parameters measuring the costs of deviation from the personally preferred action and from the mismatch with the partners' actions, respectively. Here parameter $\theta_i$ defined by equation (3) is exactly $\theta_i$ of the payoff function (10).

*EGT analysis.* In the Evolutionary Game Theory (EGT) version of this game using the best response strategy revision, the term $\tilde{x}_i$ is replaced by the average action of peers in the previous time step, $(\sum_{j \neq i} x_{j,prev})/(n-1)$. Let $r_i = \frac{d_i}{c_i + d_i}$ be the relative strength of conformity pressure for individual $i$. Assume that parameters $\theta_i$ and $r_i$ are chosen randomly and independently from certain distributions. Then the Nash equilibrium effort for individual $i$ can be approximated as $x_i^* = \theta_i + r_i(\overline{\theta} - \theta_i)$, and the average effort of the group $\overline{x}^* = \overline{\theta}$ (see the SI).

*General case.* The average action predicted by my approach is the same: $\overline{\theta}$. However the predictions for individual values $x_i^*$ will differ between the two approaches (because $\eta$ in equation 8b is different from $r$). Obviously, besides $\overline{x}^*$ and $x_i^*$, my model makes predictions for the expectations and variances of $y_i, \tilde{y}_i$ and $\tilde{x}_i$.

Figure 3 illustrates the equilibria in this model found using agent-based simulations. The EGT predictions correspond to purple bars for $\varepsilon = 0$. The case of no external influence was modeled by setting all coefficients $A_4$ and $C_{i3}$ to zero. Figure 3a shows that with no external influence,

- The mean values of $x, y, \tilde{y}$ and $\tilde{x}$ are close to $\overline{\theta}$ as predicted.
- Although with $\varepsilon = 0$ (leftmost set of bars), normative factors are absent from the utility function, variables $y, \tilde{y}$ and $\tilde{x}$ still evolve towards $\overline{\theta}$ due to the psychological processes modeled.
- The standard deviations and correlations with $\theta$ are in the order predicted: from the largest for $x$ to the smallest for $\tilde{x}$.

- Increasing the strength $\varepsilon$ of normative factors decreases within-group variation in all traits and delays convergence to an equilibrium.

Figure 3b shows that with external influence (with $G = 2$, so that the authority effectively asks individuals to double their efforts):

- Individuals respond to external influence by increasing their efforts, attitudes, and beliefs towards $G$ as $\varepsilon$ increases with the mean of $\tilde{x}$ getting the closest to $G$ and the mean of $x$ lagging the most.
- Only $x$ and, for $\varepsilon = 0$, $y$ significantly correlate with $\theta$.
- The time to convergence to the equilibrium is shorter than that without an external influence and does not depend much on $\varepsilon$.
- Even though with $\varepsilon = 0$ normative effects do not affect the utility function, mean actions are increased relative to the case of no external influence. This happens because the presence of an external influence increases individual beliefs $\tilde{x}_i$ about the actions of others which in turn pushes them to increase their action $x_i$ in order to coordinate better with groupmates.

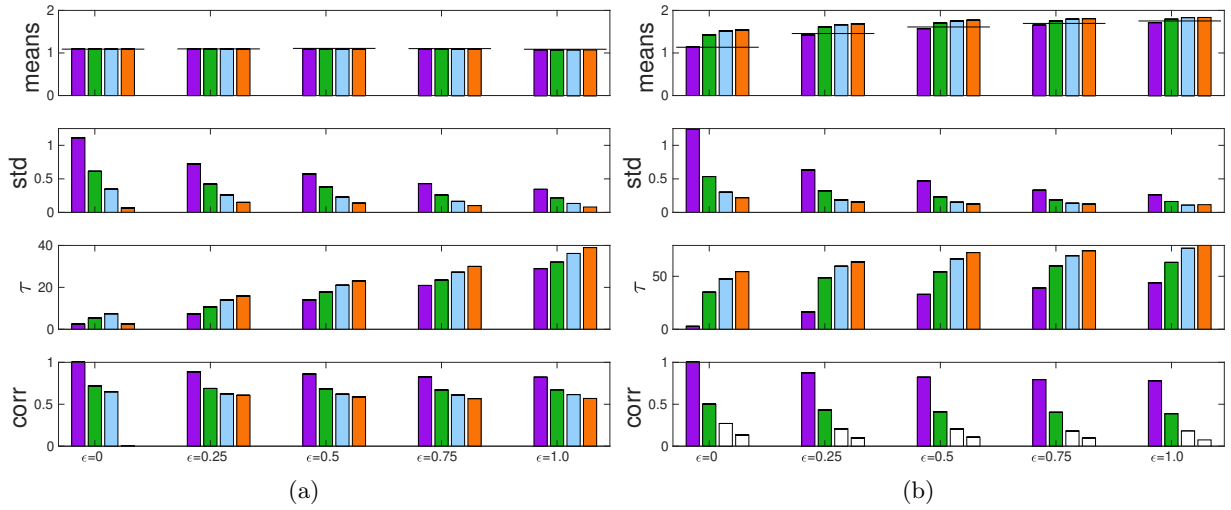



**Figure 4:** Properties of equilibria in the Public Goods game with quadratic costs. (a) No external influence. (b) With external influence promoting increased effort ($G = 2$). From top to bottom: equilibrium means, standard deviations, correlation with $\theta$, and half-time of convergence for $x, y, \tilde{y}$ and $\tilde{x}$, respectively. The thin black lines show the theoretical predictions for $x$ (given by equation S6 in the SI). Parameter $\varepsilon$ measures the importance of each of the normative factors relative to material payoffs. Group size $n = 40$. Parameters: $b_i = 40$ for each $i$, parameters $c_i$ are drawn from a lognormal distribution with mean 1 and standard deviation 0.1, parameters $v_i$ are drawn from a broken stick distributions, so that $\bar{\theta} \approx 1$. Statistics are calculated over 100 last time steps over 40 independent runs each of length 1,000 time steps.

**Public Goods Game with quadratic personal costs**. In this game, individuals make costly contributions to a total group effort $Z$ the value of which is then multiplied by a constant factor $b$. The resulting amount $P = bZ$ is then distributed back to the group members with $i$th individual getting value $v_i P$, where $v_i$ is a constant individual-specific parameter. Following Refs.(55, 56, 53, 49), assume that the cost to an individual is quadratic in their effort. In my

framework, individual $i$ making effort $x_i$ predicts that his group effort will be $Z_i = x_i + (n-1)\tilde{x}_i$. Then the estimated material payoff of individual $i$ is

$$\pi(x_i, \tilde{x_i}) = v_i b Z_i - 0.5 c_i x_i^2, \tag{11}$$

where $c_i$ is an individual cost coefficient. Straightforward calculation then shows that $\theta_i = v_i b / c_i$ which is just the benefit to cost ratio.

*EGT analysis.* In the EGT version of this model, the term $(n-1)\tilde{x}$ in the equation for $Z_i$ above is substituted by the sum of efforts of groupmates at the previous time step, $\sum_{j \neq i} x_{j,prev}$. Then the best response and the Nash equilibrium for the individual effort are equal to $\theta_i$ defined above.

*General analysis.* Figure 4 illustrates the properties of equilibria in this model which are very similar to those in the Coordination game.

**Common Pool Resource Game**. In this game (57, 58), the production function shows a diminishing return in the group effort: $P = bZ - 0.5dZ^2$, where $b$ and $d$ are constant parameters, the individual cost is linear in effort $x_i$, and the share $v_i$ is the resource going to individual $i$ is proportional to their effort: $v_i = x_i/Z$ as in the Tullock contest (59), and $Z$ is the same as defined above. The individual payoff is

$$\pi_i = v_i P - c_i x_i. \tag{12}$$

In this model, $\theta_i = \frac{2(b-c_i)}{d(n+1)}$.

*EGT analysis.* Proceeding as above, one finds that the best response action is $x_{i,\text{BR}} = \max\left(0, \frac{n+1}{2}\ \theta_i - 0.5 \sum_{j \neq i} x_{j,prev}\right)$ while and the corresponding Nash equilibria are $x_{i,\text{NE}} = \theta_i + n(\theta_i - \overline{\theta})$. If all individuals have identical coefficients $c_i = c$ and $b > c$, then the Nash equilibrium is $x_{\text{NE}} = \theta$, while the individual effort maximizing the total group payoff is $x_{\text{opt}} = (b-c)/d$, that is, $2n/(n+1)$ times smaller.

*General analysis.* Figure S6a shows that with no external influence and positive $\varepsilon$, the general equilibrium patterns are similar to those in the two others games except that with $\varepsilon = 0$ the observed values exceed the predictions. This happens because of the non-equilibrium occasionally observed in this case (see the SI). The time to convergence is very short. With positive $\varepsilon$, all individual characteristics strongly correlate with the measure $\theta$ of material benefits.

With an external authority promoting a socially optimal individual effort $G = x_{\text{opt}}$, group members actually increase rather than decrease their efforts (Figure S6b). In this game, the term $D_1$ is proportional to the group size $n$ which makes individual estimates of the expected payoff $\pi(x, \tilde{x})$ and, correspondingly, their best response $x$ very sensitive to changes in $\tilde{x}$ (see equations 2 and 4). If external authority promotes low efforts, individuals develop decreased expectations for $\tilde{x}$ about the effort of others which in turn make them to believe that opportunistically increasing their own effort will be beneficial.

**Other games**. In the SI, I consider a number of other games. A Tragedy of the Commons game with diminishing return, a game of the trade-offs between public and private production (60, 61, 56), and an "Us vs. nature" game (53, 27) show behavior similar to that of the Public Goods game with quadratic costs (illustrated in Figure 4). In particular, in these four games individuals change their action in the direction promoted by an external authority. A Public Goods game with diminishing return (62, 58) and a Tragedy of the Commons game with quadratic costs are similar to the Common Pool Resource game (illustrated in Figure S6). In particular, in these three games individuals can change their actions in the direction opposite to that promoted by an external authority. (In these games, the term $D_1$ is linearly proportional to the group size $n$.)

I also consider several games with linear payoff functions (in which $D_1 = D_2 = 0$): the classical Dictator game and the Linear Public Goods as well as the Give-or-Take game (37) and the Rule

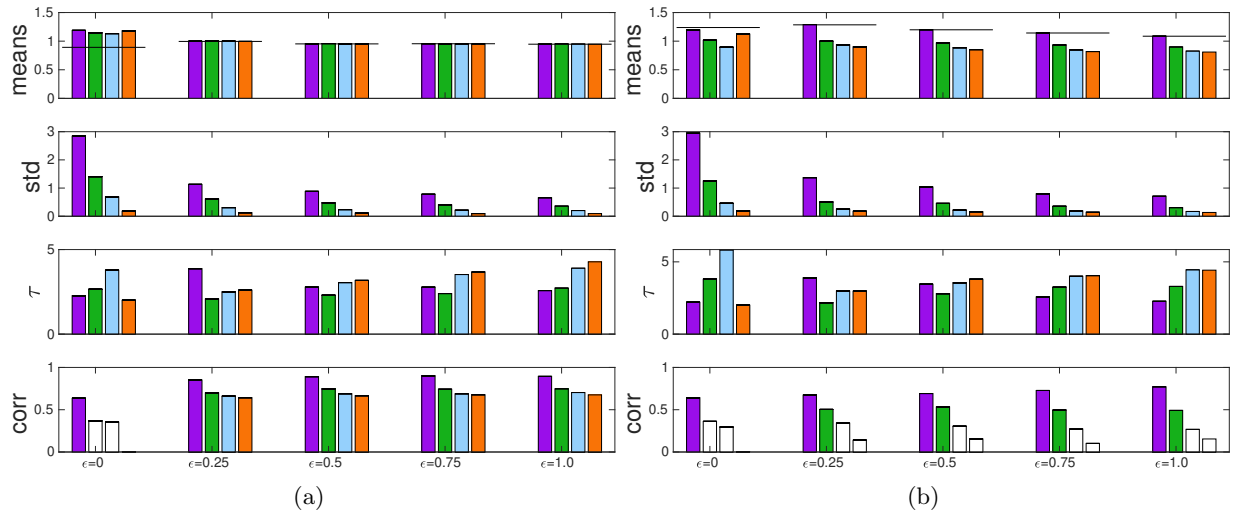



**Figure 5:** Properties of equilibria in the Common Pool Resources game. (a) No external influence. (b) With external influence promoting decreased, socially optimal effort $G = 0.5$. From top to bottom: equilibrium means, standard deviations, correlation with $\theta$, and half-time of convergence for $x, y, \tilde{y}$ and $\tilde{x}$, respectively. The thin black horizontal lines show the theoretical predictions for $x$ (given by equation S6 in the SI). Parameter $\varepsilon$ measures the importance of each of the normative factors relative to material payoffs. Group size $n = 20$. Parameters: $b_i = 10$ for each $i$ while $c_i$ and $d_i$ are drawn from lognormal distributions with mean 1 and standard deviation 0.1 so that $\bar{\theta} \approx 1$. Initial values of $y, \tilde{y}$ and $\tilde{x}$ were chosen randomly and independently from a uniform distribution on $[0, 0.1]$. Statistics are calculated over 100 last time steps over 40 independent runs each of length 1,000 time steps.

Following game (43). In the EGT versions of these games, the Nash equilibrium effort is zero but the presence of an external influence can lead to positive efforts. Similar behavior is exhibited by a continuous Prisoner's Dilemma game (63) in which the payoff function is quasi-linear (i.e., $D_2 = 0$ but both $D_0$ and $D_2$ are different from zero).

## Discussion

Here I have developed a theoretical approach for modeling the dynamics of social interactions in situations where individuals' personal norms and beliefs about others affect their own actions which in turn cause subsequent adjustments in norms and beliefs. My approach combines evolutionary game theory models focusing on material costs and benefits (13, 14) and an adaptation of social influence models focusing on the dynamics of publicly expressed opinions (19–21) with novel modeling components capturing the dynamics of beliefs about others. In my approach, the publicly observable variables are individual actions while individual attitudes and beliefs are private and can only be guessed by others. Besides predicting individual and group behavior, my models shed light on two other types of questions: which norms get internalized and which factors control beliefs about others.

**Individual characteristics**. My models predict that individual actions in social interactions, their attitudes (i.e. personal norms), and beliefs about others coevolve in a particular way. Specifically, the two most important factors in long-term dynamics are material payoffs and the influence of external authorities. In the absence of the latter, individual behavior tends to evolve towards actions maximizing their material payoffs while personal norms (attitudes) and beliefs about others exhibit coherence with individual actions. On longer time-scales, variation in normative beliefs between individuals largely reflects variation in their material benefits and costs. My models thus

predict that people have a tendency to internalize the ideas and beliefs that are most beneficial for their material well-being. In a sense, these modeling conclusions align with Marx' postulate that "...material life determines the social, political and intellectual life process in general" (64).

At the same time, as stressed already by Aristotle, human nature is deeply social and political. Culture, social learning, and conformity have played crucial roles since the origin of our species (65–68). Therefore our actions, attitudes, and beliefs are strongly affected by those of our peers as well as by external authorities (cultural, religious, political, administrative, etc). While peer influence largely works towards reducing variation between individuals, an external influence (or propaganda) can directionally shift actions, attitude and beliefs. This is a fact well known to politicians, religious leaders, cultural models, educators, marketing professionals, and social media influencers. The resulting effects can be very positive or extremely negative both from individual or societal perspectives. My models predict how individual actions are dispersed around or shifted away from those maximizing their personal material payoffs.

Under some conditions the effort of an authority to promote certain behavior can backfire and cause an opposite effect. For example, the authority's messaging about the importance of participation in a collective action can develop higher expectations about the level of contributions of peers which then will lead individuals to opportunistically decrease their own costly effort. Or the authority's messaging about the need to reduce the consumption of a common resource, can cause individuals to opportunistically increase their consumption. This is similar to situations captured by the Volunteer's Dilemma (69) when individuals fail to perform an action they would would benefit from because they expect others to volunteer.

In some of the models I considered, an external authority can cause individuals not only to perform actions detrimental for their material well-being but also to internalize preferences for such acts. My models can potentially be used to better understand obedience to authority (such as studied in Milgram's and Zimbardo's experiments, (70, 71)) or the effects of expected supernatural punishment for violating moral norm in moralizing religions (72). My results may also be useful for better understanding of the causal effects of "institutional signals" in developing better policies for social change, e.g. those stimulating pro-environment behavior (73).

**Differences with evolutionary game theory (EGT) predictions.** Standard EGT models aim to predict human behavior solely from the expected material payoff. However, the growing understanding in behavioral economics is that certain normative factors must be considered to explain observed behavior (38, 54, 74–77) (which is a fact well appreciated in social psychology). My approach not only offers a general theoretical way for doing this but also describes explicitly how normative factors (i.e., attitudes and beliefs) change as social interactions unfold.

My results show that in some cases the EGT predictions about the average behavior at a long-term equilibrium are robust to inclusion of normative factors (see also Ref.(49)) which gives some additional confidence in the robustness of the some results/conclusions of the EGT. However on short time-scales and in the presence of an external authority, the two approaches will give very different predictions. Moreover even on long-time scales, individual efforts can be smaller or larger than the EGT predictions and the distribution of individual efforts can be qualitatively different. For example, while some EGT models of collective action predict that only a single individual with the largest benefit-to-cost ratio will contribute to the group's effort (53), my models predict there will be a large number of different contributors. The dynamics of attitudes and first- and second-order beliefs, which are at the core of my approach here, are outside of the scope of the EGT.

**Groups**. My models allow for scaling up individual behavior to group characteristics. In particular, within-group variation is predicted to be the largest for individual actions, followed by individual attitudes, followed by beliefs about attitude and actions. I also predict that a newly

formed group (or a group encountering a new social situation) will go through a process of continuous reduction in these variances towards an equilibrium. This process can be interpreted as tightening of personal norms and normative and empirical expectations and can be studied experimentally (38). Convergence to an equilibrium can be fast, although, of course, the actual time-scale depends on parameters.

My variables $y, \tilde{y}$ and $\tilde{x}$ are closely related to the notion of personal, descriptive, and injunctive (prescriptive) social norms (39, 40, 36). In particular, variable $y$ gives the personal norm of an individual. The average of $\tilde{x}$ specifying the expected average behavior of others defines the descriptive norm in the group. The average of $\tilde{y}$ specifying the average belief of individuals about what others expect from them defines the injunctive norm (28). My models shows how these norms become dynamically aligned as social interactions unfold. This process is a subject of recent experimental studies (78–80).

My results show that pinning down theoretically the importance of each individual model component is hardly possible. For example, the average individual effort in the group at equilibrium depends on the weighted average of different types of individual parameters (e.g., see eq. S8 in the SI). However this is expected given the complexity of social dynamics. Similar problems emerge and are successfully dealt with in other fields, e.g. statistical physics or genetics. I note that which forces and phenomena are most important in social behavior is ultimately an empirical question.

**Tight and loose cultures**. My theoretical results can be applied to cultural differences between different human groups. Empirical research shows that human cultures vary from very "tight" to quite "loose" in the degree to which they emphasize social norms and compliance with them (81). The tight-loose (TL) differences can exist not only between different countries (82) but also within the same country, e.g. between 50 states in the US (83) and between 31 provinces in China (84). The variation on the TL scale is also observed in non-industrial societies (85). Refs.(82, 83, 85, 86) show with data that the TL variation can be explained in terms of the history of threats (e.g., environmental, internal and external warfare, etc.) faced by societies and the need to better coordinate collective actions under conditions of threat. Ref.(84) confirm this interpretation but show that cultural tightness also correlates with tighter government control of areas of urbanization and economic growth, with the strength of religious practices, and the extent of traditionality and group collectivism. Ref.(87) provided evidence that historically rice-farming societies have tighter social norms worldwide. They explained it by the facts that rice's production was very labor intensive and required farmers to coordinate water use and developed strong norms for labor exchange. Using data on small-scale societies, Ref.(85) showed the importance of two additional factors: societal complexity (88) and kinship heterogeneity. Less complex societies and patrilocal societies (in which wives settle near their husband's parents) are more tight.

All these analyses are correlational and therefore cannot claim that the factors discussed there cause cultural tightness. However theoretical studies can provide support for causality. Ref.(86) modeled cooperation in collective actions and showed that increasing the relative benefit of cooperation (which they interpreted as related to the level of the threat faced by the society) leads to a higher frequency of cooperative actions. The latter can be viewed is a measure of the strength of the (descriptive) cooperative norm.

Extending this work, my general approach allows one to study not only the effects of different factors on behavior but also on individual attitudes and beliefs, both the average values and their distributions and correlations in the group. Next I discuss these effects within the context of the TL culture scale. In my model, the variation on this scale can be measured by the variances and coefficients of variation of $x, y, \tilde{y}$ and $\tilde{x}$.

*Social heterogeneity.* My results show that in the absence of external influences, the most important factor in maintaining variation in actions, personal norms and beliefs is the variation in

parameter $\theta$ measuring individual material costs and benefits (equation 3). Variation in $\theta$ is high if individuals differ in the roles they play in the society, their abilities, compensation/valuation of the material benefit produced, and in the individual costs paid. This variation is directly related to social complexity of the society with simpler societies being expected to have less variation and, thus, more strict norms than more complex societies. My conclusion is thus in line with with the observations that urbanized areas have looser norms than rural areas (83, 87) and that more complex and heterogeneous societies have looser norms (85).

*Societal threat.* Behavioral response to a threat can often be just a rational change in the actions taken. For example, if cooperation becomes more profitable, its frequency is expected to increase as modeled in Ref.(86). Societal threat will however also affect attitudes and beliefs potentially making them more uniform and tightening culture (38). There are several ways to introduce the effects of an environmental or social threat in my models. One is via a change in the payoff function $\pi$. In the Coordination Game, a threat can be modeled as an increase in the individual cost $d_i$ of mismatch of the individual's action with the average action of peers. This would increase parameters $r_i$ making actions chosen more similar and consequently making all attitudes and beliefs more homogeneous. In other games with quadratic payoff function and in the Continuous PD game, a societal threat can be modeled as a change in parameters $\theta_i$ measuring individual benefit-to-cost ratio. Although such a change will change the means and variances of actions, attitudes and beliefs, the corresponding coefficients of variation will not be affected. Societal threat can also increase the perceived cost of disapproval by peers $A_2$, of nonconformity with peers' action $A_3$, and non-conformity with authority $A_4$. Increasing these parameters will decrease $\eta$ reducing variation in action, attitudes and beliefs, so that the society becomes more uniform.

*Propaganda effort.* Societies also vary in the strength of the effort of political, religious, intellectual, and other leaders and role models to promote certain types of behavior. As discussed above, increasing the perceived cost $A_4$ of nonconformity with authority will make the society more homogeneous. Similar effects can be achieved if the action $G$ promoted by authority significantly deviates from $\overline{\theta}$ which can be viewed as a "natural" optimum behavior for the population. With sufficiently large values of $A_4$, individual actions can shift towards $G$ "dragging" individual attitudes and beliefs along and making them more uniform. For example, in China the strength of governmental control of provinces predicts norm tightness (84).

*Cultural variation.* Data show significant cultural variation in conformity (89), cognitive dissonance (90, 91), and certain aspects of the Theory of Mind (92–94). Collectivistic cultures put special emphasis on conformity. In my model, such cultures would be characterized by increasing costs of non-conformity $A_3$ and $A_4$ and in increasing parameters $\beta$ and $\gamma$ measuring the strength of social influence on attitudes and beliefs. Such increases will cause the society to become more uniform. Similar effects will be achieved by a decrease in the strength of cognitive dissonance ($\alpha_1$) and a reduced perception of logic constraints ($\alpha_3$) which would increase the ability to "doublethink".

*Population size.* In my models of collective action, I consider a single group the size of which enters explicitly only via parameter $D_1$ and only in some models. Increasing the group size $n$ increases $D_1$ which will always decrease $\theta$ and the level of cooperation in the model because of increased free-riding. However the group size also enters implicitly because the perceived costs of cost of disapproval by peers $A_2$ and of nonconformity with peers' action $A_3$ are expected to increase with $n$. Therefore increasing population size is expected to make the culture tighter.

*Differences in the subsistence style.* Societies may differ in the types of social interactions their members most often involved. For example, coordination and reciprocal exchange of labor was very important in rice production which has contributed to tighter cultures in rice-producing regions of the world relative to wheat-producing regions (87). As discussed above, the higher the cost of mis-coordination, the tighter the society if predicted to be. Subsistence style also affects the extent

to which people rely on social learning (95).

Overall, my analysis provides a theoretical support for a causal relationship between the factors just discussed and the extent of cultural tightness/looseness.

**Possible generalizations**. My conclusions have important caveats though. First, they concern the expected average behavior of the population. In any realistic situation one may expect the presence of individuals who will not be affected by certain factors included in my model. (Mathematically for such individuals, some of the corresponding coefficients $A_i, D_i, C_{ij}$ will be equal to zero.) Second, my predictions mostly focus on long-term equilibria under the assumption of repeated interactions occurring according to a fixed set of rules. Predicting transient dynamics on short time scales is much more challenging. Third, my derivations assume that social interactions happen within a single constant group. An important future generalization would be to consider interactions on a (dynamically changing) social network or in randomly formed groups. Also important is to consider the dynamics of beliefs represented by discrete rather than continuous variables (because it is know that their equilibria can be rather different (96)). Additional potential generalizations include multidimensional extensions of the model (20, 24, 97), more realistic models of learning (e.g., Bayesian learning (98)) and strategy revision, equity concerns, and learning from others' performance. It would be interesting to use my models for studying political polarization (99) as well as the processes through which people change their social identity (100).

**Model validation**. My models can be validated using data from experiments or surveys. For example, the methods of experimental economics can be used to elicit beliefs about the actions and attitudes of others (38, 76, 77, 101, 102). For example, Ref. (76) measured subjects' actions and beliefs corresponding to my variables $x, y, \tilde{x}$ and $\tilde{y}$ in a single round of the Dictator game while Ref. (38) did the same for a group of subjects repeatedly playing a collective risk game (8). Compliance with authority was studied in a Public Goods game (103) and in the Joy of Destruction games (104). Importantly, because my main equations (e.g., eq. 5) are linear, estimating the distributions of relevant parameters using (e.g., multilevel) regressions should be relatively straightforward. In experimental economics studies of social dilemmas it is common to classify subjects into different types such as altruists, free-riders, conditional cooperators, etc (75, 105). Similar approaches can be used to study differences between individuals in their tendencies to change their personal norms and beliefs. In principle, it may be possible to compare quantitatively the relative strengths of cognitive factors ($\alpha$'s in my models), of learning from others ($\beta$'s), and of complying with authority ($\gamma$'s). Existing surveys that correlate different characteristics of societies with tightness-looseness of their norms (82–85) as well as studies of how values and social preferences change over time (106–108) offer additional opportunities to test my models.

People's attitudes and beliefs are important not only in social dilemmas as considered here but also in many other aspects of our life. They change dynamically throughout a person's life as a result of experiences (both personal and shared) and other psychological processes. They must be considered when scholars, practitioners, or policymakers try to understand or predict social processes happening at different levels of our societies. The models developed here offer a way of doing it from the theoretical point of view. The challenge will be to integrate these models with empirical work.

## Acknowledgements

I thank Yu. N. Gavrilets, A. Sánchez, D. Tverskoi, Y. Rosokha, G. Andrighetto, and reviewers for comments and suggestions. Supported by the U. S. Army Research Office grants W911NF-14-1-0637 and W911NF-18-1-0138 and the Office of Naval Research grant W911NF-17-1-0150, the National Institute for Mathematical and Biological Synthesis through NSF Award #EF-0830858,

and by the University of Tennessee, Knoxville.

*Data Availability.* All data are in the manuscript.

*Code Availability.* The Matlab code for agent-based simulations is available upon request.

## References

[1] Olson, M. *Logic of collective action: Public goods and the theory of groups* (Harvard University Press, Cambridge, MA, 1965).

[2] Pecorino, P. Olson's Logic of Collective Action at fifty. *Public Choice* **162**, 243–262 (2015).

[3] Hardin, G. Tragedy of commons. *Science* **162**, 1243–1248 (1968).

[4] Ostrom, E. Collective action and the evolution of social norms. *The Journal of Economic Perspectives* **14**, 137–158 (2000).

[5] Platt, J. Social traps. *American Psychologist* **28**, 641–651 (1973).

[6] Schelling, T. *Micromotives and Macrobehavior* (Norton, New York, 1978).

[7] Molander, P. The prevalence of free riding. *Journal of Conflict Resolution* **36**, 756–771 (1992).

[8] Milinski, M., Sommerfeld, R. D., Krambeck, H.-J., Reed, F. A. & Marotzke, J. The collective-risk social dilemma and the prevention of simulated dangerous climate change. *Proceedings of the National Academy of Sciences USA* **105**, 2291–2294 (2008).

[9] Festinger, L. *A Theory of Cognitive Dissonance* (Stanford University Press, Palo Alto, CA, 1957).

[10] Premack, D. & Wodruff, G. Does the chimpanzee have a theory of mind. *Behavioral and Brain Sciences* **1**, 515–526 (1979).

[11] Apperly, I. *Mindreaders: The Cognitive Basis of Theory of Mind* (Taylor & Francis Group, Hove, 2010).

[12] Krueger, J. I. From social projection to social behaviour. *European Review of Social Psychology* **18**, 1–35 (2007).

[13] Fudenberg, D. & Tirole, J. *Game Theory* (The MIT Press, Cambride, MS, 1992).

[14] Sandholm, W. H. *Population Games and Evolutionary Dynamics* (MIT Press, Cambridge, Massachusetts, 2010).

[15] Tembine, H. Mean-field-type games. *AIMS Mathematics* **2**, 706–735 (2017).

[16] Piotrowski, E. W. & Sladkowski, J. An invitation to quantum game theory. *International Journal of Theoretical Physics,* **42**, 1089–1099 (2003).

[17] Siopsis, G., Balu, R. & Solmeyer, N. Quantum prisoners' dilemma under enhanced interrogation. *Quantum Information Processing* **18**, Article number 144 (2018).

[18] DeGroot, M. Reaching a consensus. *Journal of the American Statistical Association* **69**, 118–121 (1974).

[19] Watts, D. J. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences USA* **99**, 5766–5771 (2002).

[20] Friedkin, N. E., Proskurnikov, A. V., Tempo, R. & Parsegov, S. E. Network science on belief system dynamics under logic constraints. *Science* **354**, 321–326 (2016).

[21] Redner, S. Reality inspired voter models: A mini-review. *Comptes Rendus Physique* (2019).

[22] Galesic, M. & Stein, D. L. Statistical physics models of belief dynamics: Theory and empirical tests. *Physica A: Statistical Mechanics and its Applications* **519**, 275–294 (2019).

[23] Zino, L., Ye, M. & Cao, M. A two-layer model for coevolving opinion dynamics and collective decision-making in complex social systems. *Chaos* **20**, 083107 (2020).

[24] Kashima, Y., Perfors, A., Ferdinand, V. & Pattenden, E. Ideology, communication and polarization. *Philosophical Transactions of the Royal Society London B* **376**, 20200133 (2021).

[25] Golman, R., Bhatia, S. & Kane, P. B. The dual accumulator model of strategic deliberation and decision making. *Psychological Review* **127**, 477–454 (2020).

[26] Azar, O. What sustains social norms and how they evolve? The case of tipping. *Journal of Economic Behavior & Organization* **54**, 49–64 (2004).

[27] Gavrilets, S. & Richerson, P. J. Collective action and the evolution of social norm internalization. *Proceedings of the National Academy of Sciences USA* **114**, 6068–6073 (2017).

[28] Gavrilets, S. The dynamics of injunctive social norms. *Evolutionary Human Sciences* **2**, e60 (2020).

[29] Perry, L., Shrestha, M. D., Vose, M. D. & Gavrilets, S. Collective action problem in heterogeneous groups with punishment and foresight. *Journal of Statistical Physics* **172**, 293–312 (2018).

[30] Perry, L. & Gavrilets, S. Foresight in a game of leadership. *Scientific Reports* **10**, 2251 (2020).

[31] Thomas, W. I. *The Child in America: Behavior Problems and Programs* (Alfred A. Knopf, New York, 1928).

[32] Wood, W. Attitude change: Persuasion and social influence. *Annu. Rev. Psychol.* **51**, 539–570 (2000).

[33] Albarracin, D. & Shavitt, S. Attitudes and attitude change. *Annual Review of Psychology* **69**, 299–327 (2017).

[34] Hedden, T. & Zhang, J. What do you think I think you think? strategic reasoning in matrix games. *Cognition* **85**, 1–36 (2002).

[35] Goodie, A. S., Doshi, P. & Young, D. L. Levels of theory-of-mind reasoning in competitive games. *Journal of Behavioral Decision Making* **25**, 95–108 (2012).

[36] Bicchieri, C. *The Grammar of Society. The Nature and Dynamics of Social Norms* (Cambridge University Press, Cambridge, 2006).

[37] Bicchieri, C., Dimant, E., Gachter, S. & Nosenzo, D. Observability, social proximity, and the erosion of norm compliance. *papers.ssrn.com/sol3/papers.cfm?abstractid=3355028* (2020).

[38] Szekely, A. *et al.* Collective risks change social norms and promote cooperation: Evidence from a long-term experiment. *Nature Communications* **x**, x (2021).

[39] Cialdini, R. L., Reno, R. R. & Kallgren, C. A. A focus theory of normative conduct:701recycling the concept of norms to reduce littering in public places. *Personality and Social Psychology* **58**, 1015–1026 (1990).

[40] Cialdini, R. B. & Goldstein, N. J. Social influence: Compliance and conformity. *Annu. Rev. Psychol.* **55**, 591–621 (2004).

[41] Song, G., Ma, Q., Wu, F. & Li, L. The psychological explanation of conformity. *Social Behavior and Personality* **40**, 1365–1372 (2012).

[42] French, J. & Raven, B. The bases of social power. In Cartwright, D. (ed.) *Studies in Social Power*, 150–167 (Inst. Soc. Res., Ann Arbor, MI, 1959).

[43] Kimbrough, E. O. & Vostroknutov, A. Norms make preferences social. *Journal of the European Economic Association* **14**, 608–638 (2016).

[44] Kimbrough, E. O. & Vostroknutov, A. A portable method of eliciting respect for social norms. *Economics Letters* **168**, 147–150 (2018).

[45] Kashima, Y. *et al.* Social transmission of cultural practices and implicit attitudes. *Organ. Behav. Hum. Decis. Process.* **127**, 113–125 (2015).

[46] Rabin, M. Cognitive dissonance and social change. *Journal of Economic Behavior and Organization* **24**, 177–194 (1994).

[47] Gomes, D. A. & Saúde, J. Mean field games models - a brief survey. *Dynamic Games and Applications* **4**, 110–154 (2014).

[48] Kuran, T. & Sandholm, W. H. Cultural integration and its discontents. *Review of Economic Studies* **75**, 201–228 (2008).

[49] Calabuig, V., Olcina, G. & Panebianco, F. Culture and team production. *Journal of Economic Behavior and Organization* **149**, 32–45 (2018).

[50] Centola, D., Willer, R. & Macy, M. The emperor's dilemma: a computational model of self-enforcing norms. *American Journal of Sociology* **110**, 1009–1040 (2005).

[51] Olsson, A., Knapska, E. & Lindström, B. The neural and computational systems of social learning. *Nature Review Neuroscience* **21**, 197–212 (2020).

[52] Gavrilets, Y. N. Stochastic modeling of between-group social interactions. *Economics and Mathematical Methods (in Russian)* **39**, 106–116 (2003).

[53] Gavrilets, S. Collective action problem in heterogeneous groups. *Proceedings of the Royal Society London B* **370**, 20150016 (2015).

[54] Andreoni, J., Nikiforakis, N. & Siegenthaler, S. Predicting social tipping and norm change in controlled experiments. *https://www.nber.org/papers/w27310* (2020).

[55] Esteban, J. & Ray, D. Collective action and the group size paradox. *American Political Science Review* **95**, 663–672 (2001).

[56] McGinty, M. & Milam, G. Public goods provision by asymmetric agents: experimental evidence. *Soc Choice Welf* **40**, 1159–1177 (2013).

[57] Walker, J. M., Gardner, R. & Ostrom, E. Rent dissipation in a limited-access common-pool resource: Experimental evidence. *Journal of Environmental Economics and Management* **19**, 203–211 (1990).

[58] Apesteguia, J. & Maier-Rigaud, F. P. The tole of rivalry: Public goods versus common-pool resources. *Journal of Conflict Resolution* **50**, 646–663 (2006).

[59] Konrad, K. *Strategy and Dynamics in Contests* (Oxford University Press, Oxford, 2009).

[60] Willinger, M. & Ziegelmeyer, A. Framing and cooperation in public good games: an experiment with an interior solution. *Economics Letters* **65**, 323–328 (1999).

[61] Willinger, M. & Ziegelmeyer, A. Association strength of the social dilemma in a public goods experiment: An exploration of the error hypothesis. *Experimental Economics* **4**, 131–144 (2001).

[62] Anderson, S. P., Goeree, J. K. & Hol, C. A. A theoretical analysis of altruism and decision error in public goods games. *Journal of Public Economics* **70**, 297–323 (1998).

[63] Verhoeff, T. The trader's dilemma: A continuous version of the prisoner's dilemma. Tech. Rep., Faculty of Mathematics and Computing Science, Technische Universiteit Eindhoven, The Netherlands (1998).

[64] Marx, K. *A Contribution to the Critique of Political Economy* (Charles H Kerr and Company, Chicago, 1959).

[65] Darwin, C. *The Descent of Man, and Selection in Relation to Sex* (John Murray, London, 1871).

[66] Richerson, P. J. & Boyd, R. *Not by genes alone. How culture transformed human evolution* (University of Chicago Press, Chicago, 2005).

[67] Henrich, J. *The Secret of Our Success* (Princeton University Press, Princeton, NJ, 2015).

[68] Richerson, P. J., Gavrilets, S. & de Waal, F. B. M. Modern theories of human evolution foreshadowed by Darwin's *Descent of Man*. *Science* **371**, 00–00 (2021).

[69] Diekmann, A. Volunteer's dilemma. *Journal of Conflict Resolution* **29**, 605–610 (1985).

[70] Milgram, S. Behavioral study of obedience. *Journal of Abnormal and Social Psychology* **67**, 371–378 (1963).

[71] Haney, C., Banks, C. & Zimbardo, P. Interpersonal dynamics in a simulated prison. *International Journal of Criminology and Penology* **1**, 69–97 (1973).

[72] Willard, A. K., Baimel, A., Turpin, H., Jong, J. & Whitehouse, H. Rewarding the good and punishing the bad: The role of karma and afterlife beliefs in shaping moral norms. *Evolution and Human Behavior* **41**, 385–396 (2020).

[73] Tankard, M. E. & Paluck, E. L. Norm perception as a vehicle for social change. *Social Issues and Policy Review* **10**, 181–211 (2016).

[74] Loewenstein, G. & Molnar, A. The renaissance of belief-based utility in economics. *Nature Human Behavior* **2**, 166–167 (2018).

[75] Fehr, E. & Schurtenberger, I. Normative foundations of human cooperation. *Nature Human Behaviour* **2**, 458–468 (2018).

[76] d'Adda, G., Dufwenberg, M., Passarelli, F. & Tabellin, G. Social norms with private values: Theory and experiments. *Games and Economic Behavior* **124**, 288–304 (2020).

[77] Górges, L. & Nosenzo, D. Measuring social norms in economics: Why it is important and how it is done. *Analyse & Kritik* **42**, 285–311 (2020).

[78] Eriksson, K., Strimling, P. & Coultas, J. C. Bidirectional associations between descriptive and injunctive norms. *Organizational Behavior and Human Decision Processes* **129**, 59–69 (2015).

[79] Tworek, C. M. & Cimpian, A. Why do people tend to infer "ought" from "is"? The role of biases in explanation. *Psychological Science* **27**, 1109–1122 (2016).

[80] Lindstrom, B., Jangard, S., Selbing, I. & Olsson, A. The role of a "common is moral" heuristic in the stability and change of moral norms. *Journal of Experimental Psychology - General* **147**, 228–242 (2018).

[81] Pelto, P. J. The differences between 'tight' and 'loose' societies. *Trans-action* **5**, 37–40 (1968).

[82] Gelfand, M. J. *et al.* Differences between tight and loose cultures: A 33-nation study. *Science* **332**, 1100–1104 (2011).

[83] Harrington, J. R. & Gelfand, M. J. Tightness-looseness across the 50 united states. *Proceedings of the National Academy of Sciences USA* **111**, 7990–7995 (2014).

[84] Chua, R. Y. J., Huang, K. G. & Jin, M. Mapping cultural tightness and its links to innovation, urbanization, and happiness across 31 provinces in China. *Proceedings of the National Academy of Sciences USA* **116**, 6720–6725 (2019).

[85] Jackson, J. C., Gelfand, M. & Ember, C. R. A global analysis of cultural tightness in non-industrial societies. *Proceedings of the Royal Society London B* **287**, 20201036 (2020).

[86] Roos, P., Gelfand, M., Nau, D. & Lun, J. Societal threat and cultural variation in the strength of social norms: An evolutionary basis. *Organizational Behavior and Human Decision Processes* **129**, 14–23 (2015).

[87] Talhelm, T. & English, A. S. Historically rice-farming societies have tighter social norms in china and worldwide (2020).

[88] Murdock, G. P. & Provost, C. Measurement of cultural complexity. *Ethnology* **12**, 379–392 (1973).

[89] Bond, R. & Smith, P. B. Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological Bulletin* **119**, 111–137 (1996).

[90] Heine, S. J. & Lehman, D. R. Culture, dissonance, and self-affirmation. *Personality and Social Psychology Bulletin* **23**, 389–400 (1997).

[91] Hoshino-Browne, E., Zanna, A. S., Spencer, S. J., Zanna, M. P. & Kitayama, S. e. a. On the cultural guises of cognitive dissonance: The case of Easterners and Westerners. *Personality and Social Psychology* **89**, 294–310 (2005).

[92] Lillard, A. Ethnopsychologies: Cultural variations in theories of mind. *Psychological Bulletin* **123**, 3–32 (1998).

[93] Lecce, S. & Hughes, C. 'The Italian job?': Comparing theory of mind performance in British and Italian children. *Journal of Developmental Psychology* **28**, 747–776 (2010).

[94] Heyes, C. M. & Frith, C. D. The cultural evolution of mind reading. *Science* **344**, 1243091 (2014).

[95] Glowacki, L. & Molleman, L. Subsistence styles shape human social learning strategies. *Nature Human Behavior* **1**, 0098 (2017).

[96] Zhong, W., Kokubo, S. & Tanimoto, J. How is the equilibrium of continuous strategy game different from that of discrete strategy game? *Biosystems* **107**, 88–94 (12).

[97] Converse, P. E. The nature of belief systems in mass publics. In D, A. (ed.) *Ideology and discontent*, 206–261 (Free Press, New York, NY, 1964).

[98] Khalvati, K. *et al.* Modeling other minds: Bayesian inference explains human choices in group decision-making. *Science Advances* **5**, eaax8783 (2019).

[99] Lees, J. & Cikara, M. Understanding and combating misperceived polarization. *Philosophical Transactions of the Royal Society London B* **376**, 20200143 (2021).

[100] Green, E. The politics of ethnic identity in Sub-Saharan Africa. *Comparative Political Studies* **53**, https://doi.org/10.1177/0010414020970223 (2020).

[101] Gill, D. & Rosokha, Y. Beliefs, learning, and personality in the indefinitely repeated prisoner's dilemma. Tech. Rep., http://dx.doi.org/10.2139/ssrn.3652318 (2020).

[102] Andreozzi, L., Ploner, M. & Saral, A. S. The stability of conditional cooperation: beliefs alone cannot explain the decline of cooperation in social dilemmas. *Scientific Reports* **10**, 13610 (2020).

[103] Silverman, D., Slemrod, J. & Uler, N. Distinguishing the role of authority "in" and authority "to". *Journal of Public Economics* **113**, 32–42 (2014).

[104] Karakostas, A. & Zizzo, D. J. Compliance and the power of authority. *Journal of Economic Behavior & Organization* **124**, 67–80 (2016).

[105] Fehr, E. & Fischbacher, U. Third-party punishment and social norms. *Evolution and Human Behavior* **25**, 63–87 (2004).

[106] Tormos, R. *The Rhythm of Modernization. How Values Change Over Time* (Brill, Leiden, 2020).

[107] Kiley, K. & Vaisey, S. Measuring stability and change in personal culture using panel data. *SocArXiv* 25 Mar. (2020).

[108] Böhm, R., Fleis, J. & Rybnicek, R. On the stabilityof social preferencesin inter-group conflict:a lab-in-the-field panel study. *Journal of Conflict Resolution* (2021).

# Supplementary Information

**Coevolution of actions, personal norms, and beliefs about others in social dilemmas**

**Sergey Gavrilets**

## Contents

**Best response action** **S1**

**A single individual joining a large and stable social system** **S2**

**Equilibrium in a general case** **S3**
Quadratic payoff function with no external influence . . . . . S4
No variation in material costs and benefits . . . . . S4
Variation in material costs and benefits parameters . . . . . S4
External influence only . . . . . S5
Linear payoff function with an exogenous influence . . . . . S5

**Games** **S6**
Numerical procedure . . . . . S6
Coordination Game . . . . . S7
Public Goods Game with quadratic personal costs . . . . . S9
Public Goods Game with diminishing returns . . . . . S10
Common Pool Resources Game . . . . . S13
Tragedy of the Commons Game with quadratic costs . . . . . S15
Tragedy of the Commons game with diminishing returns . . . . . S17
Trade-offs between public and private production . . . . . S19
Games with linear payoff functions: Dictator, Take-or-Give, Rule Obedience, and Public Goods . . . . . S21
Continuous Prisoner's Dilemma game . . . . . S23
"Us v. nature" game . . . . . S24

## Best response action

The action $x$ maximizing the utility function $u$ can be found by computing the derivative of the utility function (1):

$$\begin{aligned}\frac{\partial u}{\partial x} =& D_0 - D_1\tilde{x} - D_2 x - A_1(x-y) - A_2(x-\tilde{y}) - A_3(x-\tilde{x}) - A_4(x-G),\\ =& (D_0 - D_1\tilde{x} + A_1 y + A_2\tilde{y} + A_3\tilde{x} + A_4 G) - (D_2 + A_1 + A_2 + A_3 + A_4)x.\end{aligned}$$

Solving the above equation for $x$ gives us the best response action given a certain attitude $y$ and beliefs $\tilde{x}$ and $\tilde{y}$. I will write it as

$$x_{\text{BR}} = \max(0, B_0 + B_1 y + B_2\tilde{y} + B_3\tilde{x} + B_4 G), \tag{S1a}$$

where

$$B_0 = \frac{D_0}{S}, B_1 = \frac{A_1}{S}, B_2 = \frac{A_2}{S}, B_3 = \frac{A_3 - D_1}{S}, B_4 = \frac{A_4}{S} \tag{S1b}$$

are re-scaled individual-specific parameters measuring the effects of material and immaterial forces on individual actions ($i = 0, 1, 2$ and 3), and

$$S = D_2 + \sum_{i=1}^{4} A_i \tag{S1c}$$

The above equation for $x_{\text{BR}}$ naturally assumes that $S \neq 0$. In analogous evolutionary game theory models (in which all coefficients $A_i$ are zero), $S$ will be zero if $D_2 = 0$. In this case, the best response $x_{\text{BR}}$ will be equal to a maximum (if $D_0 - D_1\tilde{x} > 0$) or 0 (if $D_0 - D_1\tilde{x} < 0$) possible value of $x$.

As an example, if one disregards all other forces involved in decision-making and focus only on material cost-benefit considerations (i.e. if all $A_i = 0$), the best response action will be

$$x_{\text{BR}} = \frac{D_0 - D_1\tilde{x}}{D_2}. \tag{S2a}$$

If the individual believes that the average action of their social partners will always match their own action (i.e., $\tilde{x} = x$),

$$x_{\text{BR}} = \frac{D_0}{D_1 + D_2}. \tag{S2b}$$

which is the definition of parameter $\theta$ (equation 3 of the main text).

Note that in standard evolutionary game theory (EGT) models using myopic best response, variable $\tilde{x}$ is replaced by the average action $\sum_{j\neq i} x_{j,prev}/(n-1)$ of their social partners which individuals know exactly.

## A single individual joining a large and stable social system

Assume that an individual joins a society where the actions, attitudes and beliefs have already evolved to a certain stable distribution. Let the society be large enough so that the impact of a single additional individual on it is negligible. This will allow us to treat the average action of social partners $X$ as constant. I am interested in how the individual's characteristics will change after joining the society. [Note that this model can be used for describing the subject's behavior when embedded in a group with bots acting according to a certain pre-programmed pattern.]

The attitude and beliefs of the focal individual will change according to recurrence equations (S3). Assume that they converge to an equilibrium $(x^*, y^*, \tilde{y}^*, \tilde{x}^*)$ at which $x^* > 0$. From equations (5) and using the fact that $\beta_i = 1 - \alpha_i - \gamma_i$ for all $i$, at this equilibrium:

$$y^* = X + \alpha_1(x^* - X) + \gamma_1(G - X) \tag{S3a}$$

$$\tilde{y}^* = X + \alpha_2(y^* - X) + \gamma_2(G - X) = X + \alpha_1\alpha_2(x^* - X) + (\alpha_2\gamma_1 + \gamma_2)(G - X), \tag{S3b}$$

$$\tilde{x}^* = X + \alpha_3(\tilde{y}^* - X) + \gamma_3(G - X) = X + \alpha_1\alpha_2\alpha_3(x^* - X) + (\alpha_3(\alpha_2\gamma_1 + \gamma_2) + \gamma_3)(G - X), \tag{S3c}$$

Substituting these into the best response equation (S1a) and solving for $x$,

$$\begin{aligned} x =& \frac{B_0 + B_4 G + (B_1 + B_2 + B_3 - B_1\alpha_1 - B_2\alpha_1\alpha_2 - B_3\alpha_1\alpha_2\alpha_3)X \dots}{1 - (B_1\alpha_1 + B_2\alpha_1\alpha_2 + B_3\alpha_1\alpha_2\alpha_3)} \\ & \underline{\dots + (G - X)\left[B_1\gamma_1 + B_2(\gamma_1\alpha_2 + \gamma_2) + B_3(\gamma_1\alpha_2\alpha_3 + \gamma_2\alpha_3 + \gamma_3)\right]} \\ & \dots \\ =& \frac{B_0 + B_4 G}{1 - (B_1\alpha_1 + B_2\alpha_1\alpha_2 + B_3\alpha_1\alpha_2\alpha_3)} \\ & + \left(\frac{B_1 + B_2 + B_3 - 1}{1 - (B_1\alpha_1 + B_2\alpha_1\alpha_2 + B_3\alpha_1\alpha_2\alpha_3)} + 1\right) X \dots \\ & + \frac{B_1\gamma_1 + B_2(\gamma_1\alpha_2 + \gamma_2) + B_3(\gamma_1\alpha_2\alpha_3 + \gamma_2\alpha_3 + \gamma_3)}{1 - (B_1\alpha_1 + B_2\alpha_1\alpha_2 + B_3\alpha_1\alpha_2\alpha_3)}(G - X). \end{aligned}$$

I can rewrite the last equation as

$$x^* = \delta + (1 - \eta)X + \xi(G - X), \tag{S4a}$$

where

$$\delta = \frac{B_0 + B_4 G}{1 - (B_1\alpha_1 + B_2\alpha_1\alpha_2 + B_3\alpha_1\alpha_2\alpha_3)}, \tag{S4b}$$

$$\eta = \frac{1 - B_1 - B_2 - B_3}{1 - (B_1\alpha_1 + B_2\alpha_1\alpha_2 + B_3\alpha_1\alpha_2\alpha_3)}, \tag{S4c}$$

$$\xi = \frac{B_1\gamma_1 + B_2(\gamma_1\alpha_2 + \gamma_2) + B_3(\gamma_1\alpha_2\alpha_3 + \gamma_2\alpha_3 + \gamma_3)}{1 - (B_1\alpha_1 + B_2\alpha_1\alpha_2 + B_3\alpha_1\alpha_2\alpha_3)} \tag{S4d}$$

Note that $\xi = 0$, if propaganda by the external authority does not affect individual attitude and beliefs (i.e. all $\gamma_i = 0$). Correspondingly, the deviation of the equilibrium value of $x$ for a focal individual from $X^*$ can be written as

$$x^* - X^* = \delta - \eta X^* + \xi(G - X^*). \tag{S5}$$

From here it is straightforward to find equilibrium values of $y, \tilde{y}$ and $\tilde{x}$ using equations (S3). Note that only non-negative values of $x^*, y^*, \tilde{x}^*, \tilde{y}^*$ and $X^*$ make sense within my framework.

## Equilibrium in a general case

In the general case, all $n$ individuals will be updating their attitudes and beliefs and the average efforts of peers $X$ will be changing in time. However one still can use equation (S4a) to approximate the equilibrium. Specifically, summing up across all individuals and equating the average values of $x$ and $X$, one finds that at equilibrium

$$X^* = \frac{\overline{\delta} + G\overline{\xi}}{\bar{\eta} + \overline{\xi}}, \tag{S6a}$$

where the bar means the average over the group.

In some of the models I consider in the main text, only a subset of individuals, typically

with the largest benefit-to-cost ratios and/or most affected by external influences, make contribution at equilibrium, while others free-ride. In such situations, to predict the average group effort one needs to sum up equations (S4a) only over a subset $L$ of individuals making positive efforts. Equation (S6a) then takes a form

$$X^* = \frac{\sum_L \delta + G\sum_L \xi}{n - l + \sum_L \eta + \sum_L \xi}, \tag{S6b}$$

where $l$ is the number of contributing individuals. In principle, one can find the individuals who make positive contributions at equilibrium using an iterative procedure similar to that in ref.(1). I leave this for future work.

Knowing $X^*$ allows us to find the equilibrium values of $x, y, \tilde{y}, \tilde{x}$ for each individual. Next I consider some special cases.

## Quadratic payoff function with no external influence

Assume that external influence is absent so that $B_4 = \gamma_1 = \gamma_2 = \gamma_3 = 0$. Then $\xi = 0$. Note that in this case, the numerator in the equation for $\delta$ is $D_0/S$ while that in the equation for $\eta$ is $1 - (B_1 + B_2 + B_3) = (D_1 + D_2)/S$. Both equations have the same denominator. This implies that $\delta(D_1 + D_2) = \eta D_0$.

### No variation in material costs and benefits

Assume that there is no variation in coefficients $D_0, D_1$ and $D_2$ between individuals. Then I find that

$$X^* = \frac{\overline{\delta}}{\overline{\eta}} = \frac{D_0}{D_1 + D_2} = \theta. \tag{S7}$$

That is, the average action is the action predicted if immaterial forces are neglected (see equation S2a). Therefore, $\delta - \eta X^* = 0$, so that $x^* = X^*$. From equations (S3), one concludes that $y^* = \tilde{y} = \tilde{x} = X^*$ for all individuals. That is, with no variation in material costs and benefits, the group will converge to a state with identical actions, attitudes, and beliefs.

### Variation in material costs and benefits parameters

Allowing for variation in $D_0, D_1$ and $D_2$ and approximating the ratio of expectations $\frac{\overline{\delta}}{\overline{\eta}}$ by the expectation of ratio $\overline{\beta/\eta}$ I find that

$$X^* = \frac{\overline{\delta}}{\overline{\eta}} \approx \overline{\delta/\eta} = \overline{\frac{D_0}{D_1 + D_2}} = \overline{\theta}. \tag{S8a}$$

That is, the average action is approximately the average of actions predicted if immaterial forces are neglected (see equation S2a).

Using equations (S5) and (S3), I find that at equilibrium for each individual

$$x^* \approx X^* + \eta\ (\theta - \overline{\theta}), \tag{S8b}$$

$$y^* \approx X^* + \alpha_1\eta\ (\theta - \overline{\theta}), \tag{S8c}$$

$$\tilde{y}^* \approx X^* + \alpha_1\alpha_2\eta\ (\theta - \overline{\theta}), \tag{S8d}$$

$$\tilde{x}^* \approx X^* + \alpha_1\alpha_2\alpha_3\eta\ (\theta - \overline{\theta}). \tag{S8e}$$

With no cognitive dissonance (i.e. if $\alpha_1 = 0$), $y^* = \tilde{y}^* = \tilde{x}^* = X^*$. Without the "theory of mind" (i.e. if $\alpha_2 = 0$), $\tilde{y}^* = \tilde{x}^* = X^*$. Without beliefs dissonance (i.e. if $\alpha_3 = 0$), $\tilde{x}^* = X^*$. Note that mean values of $x^*, y^*, \tilde{y}^*$ and $\tilde{x}^*$ are all approximately equal to $X^*$ if the correlation between $r, \eta$ and the strength of cognitive factors $\alpha_1, \alpha_2, \alpha_3$ is low. Assuming independence of $\theta$ and $\eta$, the corresponding variances are approximately

$$\mathrm{var}(x) = \mathrm{var}(\theta)\ \overline{\eta^2}, \tag{S9a}$$

$$\mathrm{var}(y) = \mathrm{var}(\theta)\ \overline{(\alpha_1\eta)^2}, \tag{S9b}$$

$$\mathrm{var}(\tilde{y}) = \mathrm{var}(\theta)\ \overline{(\alpha_1\alpha_2\eta)^2}, \tag{S9c}$$

$$\mathrm{var}(\tilde{x}) = \mathrm{var}(\theta)\ \overline{(\alpha_1\alpha_2\alpha_3\eta)^2}, \tag{S9d}$$

where $\mathrm{var}(\theta)$ is the variance of $\theta$'s in the group. Because $\alpha_1, \alpha_2, \alpha_3 < 1$, all this implies that

$$\mathrm{var}(x) > \mathrm{var}(y) > \mathrm{var}(\tilde{y}) > \mathrm{var}(\tilde{x}). \tag{S10}$$

That is, my model predicts that the variation in actions (and deviation from the mean) will be the largest, followed by the variation in personal norms, followed by the variation in beliefs about norms of others, followed by the variation in beliefs about the action of others. Similarly, the correlation with material benefits (characterized by parameter $\theta$) will be the highest for personal beliefs $y$, followed by normative expectations $\tilde{y}$, and empirical expectations $\tilde{x}$. These are testable predictions.

**External influence only**

If there are no material payoffs in the utility function, i.e. if $D_0 = D_1 = D_2 = 0$, straightforward calculation shows that $\sum B_i = 1$ and that $\eta - \delta/G = 0$ for each individual. Therefore, using equations (S3)-(S4a) for each individual,

$$x^* = y^* = \tilde{y} = \tilde{x} = G. \tag{S11}$$

That is, the population's actions, attitudes, and beliefs at long-term equilibrium are completely determined by the the external influence. There will be no variation between individuals.

**Linear payoff function with an exogenous influence**

Here I assume that $D_1 = D_2 = 0$ for all individuals while there is variation between individuals in $D_0$. The corresponding game-theoretic models neglecting immaterial factors predict a simple behavior: individuals will make the maximum possible effort (if $D_0 > 0$) or no effort

(if $D_0 < 0$). I will assume that $D_0 < 0$ which is the case in several games I consider below. With immaterial factors added but still with no external influence (i.e. $A_4 = 0$), individuals' actions, attitudes and beliefs will converge to 0.

Assume there is an external authority promotes a positive effort $G$. In this case, using equations (S1a) and (S1c) I find that $S = \sum_{i=1}^{4} A_i, \sum_{i=1}^{4} B_i = 1$, and that $\delta = (\kappa + G)\eta$, where

$$\kappa = D_0/A_4 \tag{S12}$$

is a measure of the strength of material forces relative to that of external influence. Then the average effort at equilibrium can be approximated as

$$X = G - \frac{\overline{\kappa\eta}}{\bar{\eta} + \bar{\xi}}. \tag{S13}$$

where $\kappa = |D_0|/A_4$ is a measure of the strength of material forces relative to that of external influences and the composite parameters $\xi$ (defined in the SI) is non-negative. If the effect $A_4$ of an external authority is large enough, $\kappa$ is small and the average group effort will be close to $G$.

One can now find the equilibrium values of $y, \tilde{x}$ and $\tilde{y}$ from equations (S3).

## Games

### Numerical procedure

In numerical simulations used to illustrate my results, I used the following procedure for generating parameter values. I start by assigning parameters $D_0, D_1, D_2$ by drawing numbers randomly and independently from certain distributions (specified below). Then I assign parameters $A_1, \ldots, A_4$ of the utility function (1) using a two-step procedure. At the first step, I choose them randomly and independently from a "broken stick distribution" on a unit interval (2). Then I multiply these numbers by a parameter $\varepsilon$ which will vary from 0 to 1. With $\varepsilon = 0$, any normative effect in the utility function will be absent and individuals will behave according to standard evolutionary game theory assumptions. In contrast with $\varepsilon = 1$, the expected values of each of parameters $A_i$ will be the same as that of $D_2$ (in models with $D_2 \neq 0$) or $D_0$ (in models with $D_2 = 0$). That is, with $\varepsilon = 1$, the expected weight of each term in the utility function (1) will be the same. Finally I draw parameters $C_{ij}$ randomly and independently from a broken stick distribution on interval $[0, 0.1]]$. Initial values of $x, y, \tilde{y}$ and $\tilde{x}$ are drawn randomly and independently form a uniform distribution on $[0, 0.1]$. At each round, each individual revises its effort with probability 0.5. [In some simulations, after each update I have perturbed the dynamic variables by small random errors drawn from a uniform distribution on $[-\sigma, \sigma]$. The effects of such random noise are intuitive. Therefore for clarity, I removed it from the simulations illustrated in all figures.]

Table 1 summarizes the games I will consider. For each game, I will: a) define the payoff function $\pi(x, \tilde{x})$ and identify the corresponding $\theta$ value, b) identify the Nash equilibrium in the corresponding evolutionary game theory (EGT) model, and c) show results of agent-based simulations illustrating individual and group characteristics and compare them with my approximations and EGT predictions. In the EGT versions of these games, individuals will use best response to maximize their payoff. The corresponding payoff functions will be

**Table S1:** Production function $P$ (or $p_i$) and expected payoffs $\pi(x_i, \tilde{x}_i)$ in different games. Games with quadratic payoff functions: Coordination, Public Goods Game (PGG) with quadratic costs, PGG with diminishing return, Common Pool Resource (CPR), Tragedy of the Commons (TC) with quadratic costs, and TC with diminishing return. Games with linear payoff functions: Dictator, Give-or-Take, Rule Following, and Linear PGG. Game with quasi-linear payoff function: continuous Prisoner's Dilemma. Nonlinear game: "us vs. nature" game. In all collective action games, the expected group effort is $Z = x + (n-1)\tilde{x}_i$. An empty entry in the table means that in the corresponding game the corresponding function or parameter is not defined. Parameters with subscript $i$ (e.g., $b_i, c_i, d_i, r_i, v_i$) are specific for individuals. Parameters without subscripts (e.g.., $b, d, R$) are the same for all individuals. $R$ and $r_i$ are the endowments. Note that variable $P$ in collective action games is the production function while in the Prisoner's Dilemma game, $P_i$ is the punishment payoff.

| Game | Production $P/p_i$ | Expected payoff $\pi_i$ |
|---|---|---|
| Coordination | | $b_i - 0.5c_i(x_i - \theta_i)^2 - 0.5d_i(x_i - \tilde{x}_i)^2$ |
| PGG w/ quadratic costs | $P = bZ$ | $v_iP - 0.5c_ix_i^2$ |
| PGG w/ diminishing return | $P = bZ - 0.5dZ^2$ | $v_iP - c_ix_i$ |
| CPR | $P = bZ - 0.5dZ^2$ | $\frac{x_i}{Z}\ P - c_ix_i$ |
| TC w/ quadratic costs | $p_i = b_ix_i$ | $p_i - 0.5c_iZ^2$ |
| TC w/ diminishing return | $p_i = b_ix_i - 0.5d_ix_i^2$ | $p_i - c_iZ$ |
| Public vs. private production | $P = bZ - 0.5dZ^2$ | |
| | $p_i = b_iz_i - 0.5d_iz_i^2$ | $v_iP + p_i$, where $z_i + x_i = r_i$ |
| Dictator | | $R - x_i$ |
| Give-or-Take | | $R - x_i$ |
| Rule Following | | $R - x_i$ |
| Linear PGG | $P = bZ$ | $v_iP - c_ix_i$ |
| Continuous PD | | $x_i\tilde{x}_iR_i + x_i(1-\tilde{x}_i)S_i \ldots$ |
| | | $+(1-x_i)\tilde{x}_iT_i + (1-x_i)(1-\tilde{x}_i)P_i$ |
| "Us vs. nature" | $P = b\frac{Z}{Z+Z_0}$ | $v_iP - c_ix_i$ |

the same as specified in Table 1 except that the term $\tilde{x}_i$ (empirical expectation of peers' action) will be replaced by the average action $\overline{x}_{i,prev}$ of groupmates at the previous time step as is usually done in best-response modeling.

## Coordination Game

I assume that individuals interact randomly in groups. Each individual has a preferred action but each player also pays a cost if his action deviates from the average action of the group (3, 4). The corresponding (subjective) expected payoff function can be written as

$$\pi(x_i, \tilde{x}_i) = b_i - 0.5c_i(x_i - \theta_i)^2 - 0.5d_i(x_i - \tilde{x}_i)^2, \tag{S14}$$

where $\theta_i \geq 0$ is the preferred action of individual $i$, $\tilde{x}_i$ is the expected average action, $b_i$ is the maximum benefit, and $c_i$ and $d_i$ are parameters measuring the costs of deviation from the personally preferred action and from the mismatch with the partners' actions. For this

game, $d\pi/dx_i = -c_i(x_i - \theta_i) - d_i(x_i - \tilde{x}_i)$. Therefore $D_0 = c_i\theta_i, D_1 = -d_i, D_2 = c_i + d_i$. Parameter $\theta_i$ defined in equation (3) is exactly $\theta_i$ of the payoff function $\pi$.

*EGT analysis.* In the EGT version of this game, individuals will aim to maximizes the payoff function (S14) in which the term $\tilde{x}_i$ is substituted by the average action $\overline{x}_{i,prev}$ of groupmates at the previous time step. I will simplify my analysis by assuming that the groups size is large enough so that the effect of any single individual on the average is negligible. In this case, all $\overline{x}_{i,prev}$ values will approximately be the same, so that I can drop the subscript $i$.

Computing the derivative $\partial\pi/\partial x_i$, I find that the best response action for individual $i$ is

$$x_{i,\mathrm{BR}} = (1 - r_i)\theta_i + r_i\overline{x}_{prev}, \tag{S15a}$$

where

$$r_i = \frac{d_i}{c_i + d_i}$$

is the relative strength of conformity pressure.

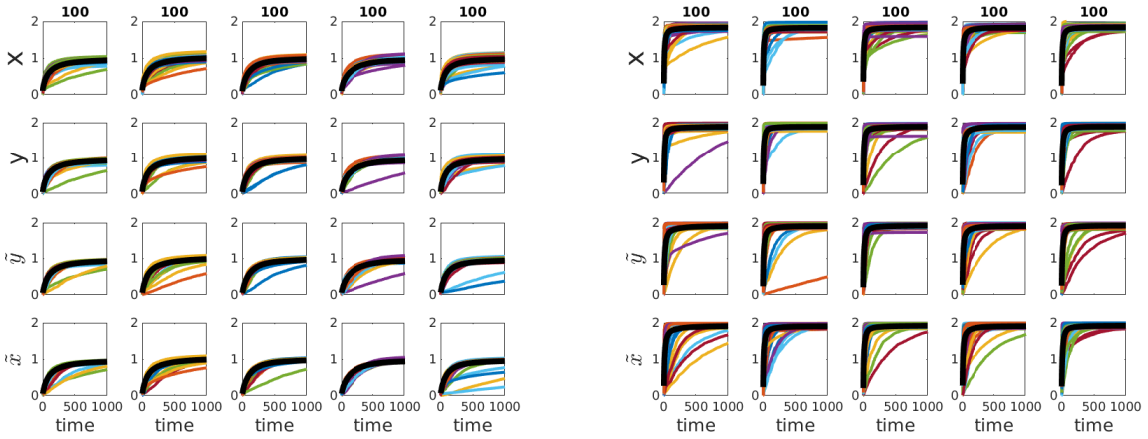



**Figure S1:** Examples of coevolutionary dynamics in the Coordination Game corresponding to Figure 3 of the main text. (a) Five runs with no external influence. (b) Five runs with external influence with $G = 2$. Different colors show different individuals. The thick black lines show the group averages. Group size $n = 100, \varepsilon = 1$. The numbers on top show the number of contributing individuals at the last time step. Other parameters: individual values of $\theta_i, c_i, d_i$ are drawn randomly and independently from lognormal distributions with mean 1 and standard deviation 0.1. Initial values of $y, \tilde{y}$ and $\tilde{x}$ were chosen randomly and independently from a uniform distribution on $[0, 0.1]$.

Assume that parameters $\theta_i$ and $r_i$ are distributed in the group independently. Then the average individual effort at (Nash) equilibrium is

$$\overline{x}^* = \overline{\theta}, \tag{S15b}$$

while the equilibrium effort for individual $i$ can be written as

$$x_i^* = \theta_i + r_i(\overline{\theta} - \theta_i). \tag{S15c}$$

Here and below bars mean the average over the group.

*General analysis.* Figure 3 in the main text summarizes my results for this model. Figure S1 shows sample trajectories corresponding to $\varepsilon = 1$ (i.e. when all components of the utility function (1) are of similar order).

### Public Goods Game with quadratic personal costs

In this game, individuals make costly contributions $x_i$ to a common group effort $Z$ the value of which is then multiplied by a constant factor $b$. The resulting amount $P = bZ$ is then distributed back to the group members with $i$th individual getting value $v_i P$. Following (5–8), assume that the cost to an individual is quadratic in their effort. Then the expected material payoff of individual $i$ making effort $x_i$ given the expectation that the groupmates will make an average effort $\tilde{x}_i$ is

$$\pi(x, \tilde{x}) = v_i b Z - 0.5 c_i x_i^2, \tag{S16}$$

where $c_i$ is the individual cost coefficient and the expected total group effort $Z = x_i + (n-1)\tilde{x}$.

One finds that $d\pi_i/dx_i = bv_i - c_i x_i$ so that $D_{0,i} = bv_i, D_{1,i} = 0, D_{2,i} = c_i$, and

$$\theta_i = \frac{bv_i}{c_i}. \tag{S17}$$

is just the benefit to cost ratio.

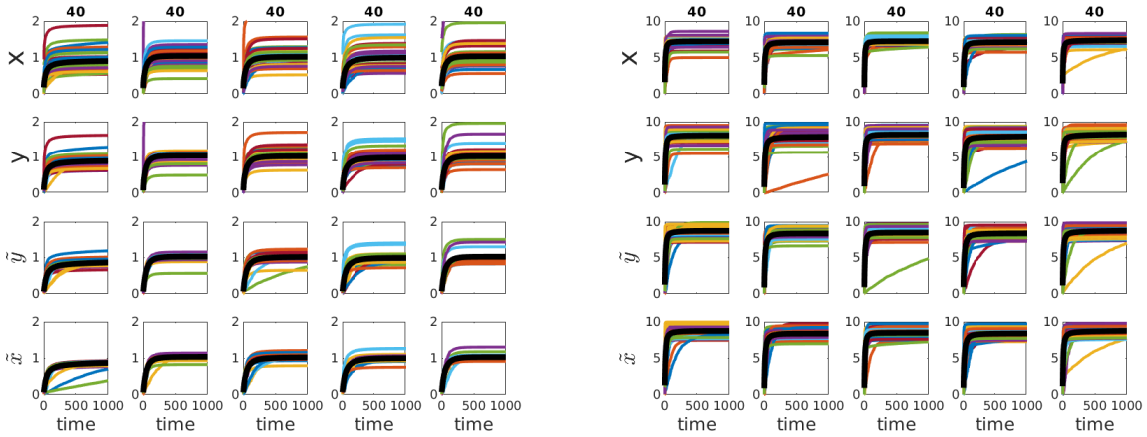



**Figure S2:** Examples of coevolutionary dynamics in the Public Goods Game with quadratic costs. (a) Five runs with no external influence. (b) Five runs with external influence with $G = 10$. Different colors show different individuals. The thick black lines show the group averages. Group size $n = 20, \varepsilon = 1$. The numbers on top show the number of contributing individuals at the last time step. Other parameters: $b = n$, individual values of $c_i$ are drawn randomly and independently from lognormal distributions with mean 1 and standard deviation 0.1, values of $v_i$ are drawn from a broken stick distribution on interval $[0, 1]$. Initial values of $y, \tilde{y}$ and $\tilde{x}$ were chosen randomly and independently from a uniform distribution on $[0, 0.1]$.

*EGT analysis.* In the EGT version of this model, the term $(n-1)\tilde{x}$ in the expression for $Z$ will be substituted by the sum of efforts of groupmates at the previous time step, $Z^-_{i,prev} = \sum_{j \neq i} x_{j,prev}$. Then, the best response and the Nash equilibrium for individual effort

is

$$x_{i,\mathrm{BR}} = x_{i,\mathrm{NE}} = \theta_i.$$

If all individuals have identical coefficients $v_i = 1/n$ and $c_i = c$, then the Nash equilibrium is

$$x_{\mathrm{NE}} = \frac{b}{nc}$$

while the effort maximizing the total group payoff is

$$x_{\mathrm{opt}} = \frac{b}{c},$$

that is, $n$ times bigger.

*General analysis.* Figure 4 in the main text illustrates the general patters in this model. Figure S2 shows sample trajectories of the general model corresponding to Figure 4 with $\varepsilon = 1$.

### Public Goods Game with diminishing returns

In this game (9, 10), the production function shows a diminishing return in the group effort $X$:

$$P = bZ - 0.5dZ^2. \tag{S18a}$$

and the individual payoff is

$$\pi_i = v_i P - c_i x_i. \tag{S18b}$$

Here, $d\pi_i/dx_i = v_i(b - d(x_i + (n-1)\tilde{x}) - c_i$. Therefore $D_{0,i} = v_i b - c_i, D_1 = v_i d(n-1), D_2 = v_i d$, and

$$\theta_i = \frac{b - c_i/v_i}{dn}. \tag{S18c}$$

*EGT analysis.* In the EGT version of this game, the term $(n-1)\tilde{x}$ in the expression for $d\pi_i/dx_i$ will be substituted by the previous total effort $Z^-_{prev,i}$ of $i$th individual's peers. The best response effort is

$$x_{\mathrm{BR},i} = \max\left(0, n\theta_i - Z^-_{prev.i}\right).$$

In the symmetric version of this game when all coefficients are the same (i.e. $c_i = c$ and $v_i = 1/n$ so that all $\theta_i$ are the same), the Nash equilibrium for the total group effort is

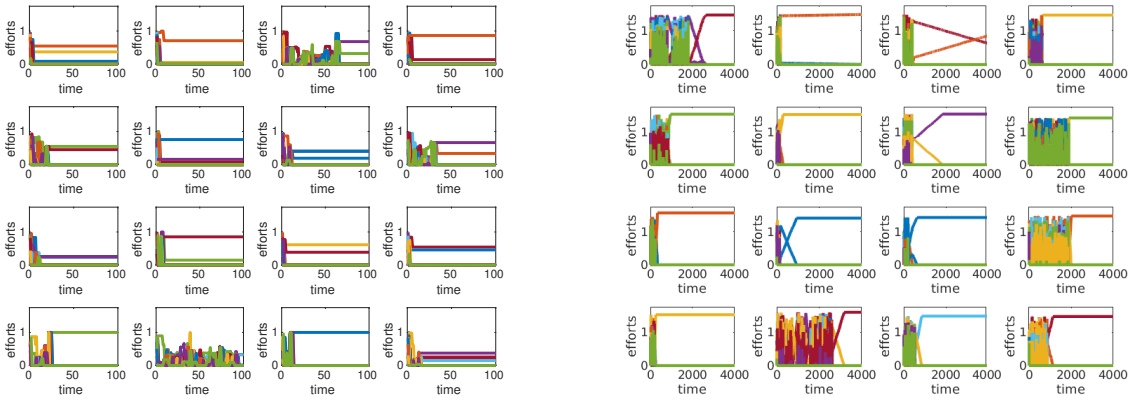
$$Z_{\mathrm{NE},sym} = n\theta = \frac{b - cn}{d}.$$

In contrast, the total group effort maximizing the total group payoff is

$$Z_{\mathrm{opt}} = \frac{b - c}{d}.$$

Individual contributions can take any values as long as they sum up to $Z_{\mathrm{NE},sym}$. Numerical agent-based simulations using myopic best response show that the system converges to this equilibrium, sometimes in a non-monotonic way; at this equilibrium the effort $Z_{\mathrm{NE},sym}$ is supplied by one or few individuals (see Figure S3a).

In the asymmetric case, when values of $v_i$ and $c_i$ differ between individuals, the system



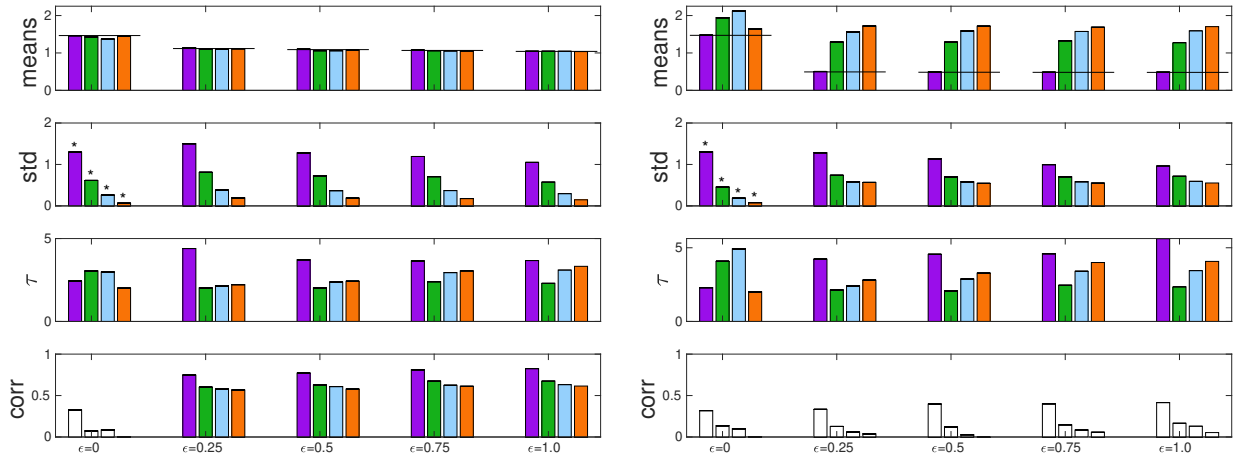

(a) (b)

**Figure S3:** Examples of best response dynamics in the Public Goods Game with quadratic costs. (a) Five runs in the symmetric model. (b) Five runs in the asymmetric model. $n = 20$. The numbers on top show the number of contributing individuals at the last time step. The appearance of nonlinear dynamics in a linear system may appear strange. However because of the truncation used to avoid negative values of my variables, the system effectively becomes nonlinear.

evolves to an equilibrium at which only a single individual with the smallest value of $c_i/v_i$ will make an effort (see Figure S3b). This effort, which is also the total group effort, is

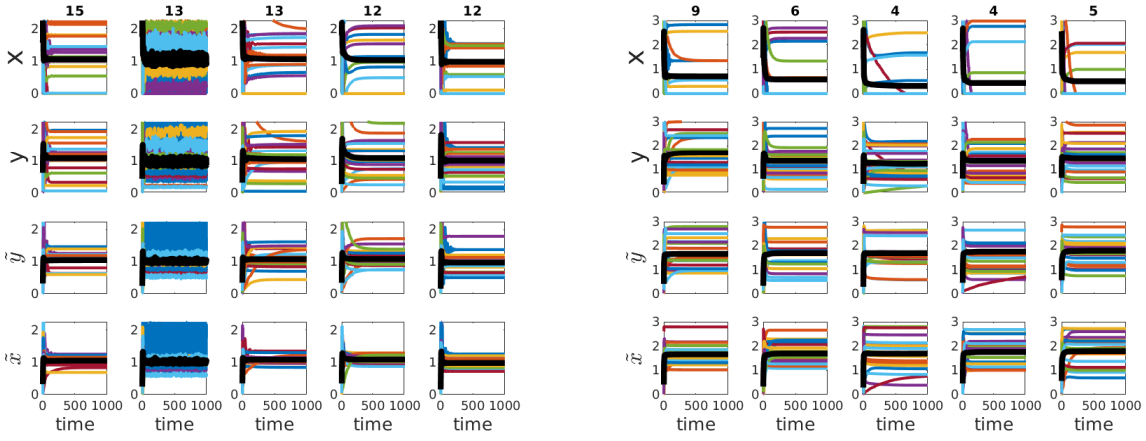
$$Z_{\mathrm{NE},asym} = \max(\theta_i).$$

*General analysis.* Figure S4 summarizes the properties of this model. The dynamics observed in agent-based simulations are often non-equilibrium (see Fig. S5). What happens is that a small number of individuals with sufficiently large values of $\theta_i$ are making large efforts while the rest of the population free-ride. For some individuals, contributions change in a cyclical or chaotic manner. The appearance of nonlinear dynamics in a linear system may appear strange. However because of the truncation used to avoid negative values of my variables, the system effectively becomes nonlinear.




(a) (b)

**Figure S4:** Properties of equilibria in the Public Goods game with diminishing return. (a) No external influence. (b) With external influence with $G = 3$. From top to bottom: equilibrium means, standard deviations, correlation with $\theta$, and half-time of convergence for $x, y, \tilde{y}$ and $\tilde{x}$, respectively. Bars with no color mean the corresponding correlations are statistically insignificant (at 0.05). The thin black horizontal lines show the theoretical predictions for $x$ (given by equation S6 in the SM). Parameter $\varepsilon$ measures the importance of each of the normative factors relative to material payoffs. Group size $n = 20$. Other parameters $b_i = 2n, c_i = d_i = 1$ for all $i$ while parameters $v_i$ are drawn from a broken stick distribution. Initial values of $y, \tilde{y}$ and $\tilde{x}$ were chosen randomly and independently from a uniform distribution on $[0, 0.1]$. The stars on top of the bars for $\varepsilon = 0$ mean that the actual values of standard deviations are 5 times larger than shown. Statistics are calculated over 100 last time steps over 40 independent runs each of length 1,000 time steps.




(a) (b)

**Figure S5:** Examples of coevolutionary dynamics in the Public Goods Game with diminishing returns corresponding to Figure S4. (a) Five runs with no external influence. (b) Five runs with external influence with $G = 3$. $\varepsilon = 1$. Different colors show different individuals. The thick black lines show the group averages. The numbers on top show the number of contributing individuals at the last time step.

**Common Pool Resources Game**

In this game (11, 10), the production function shows a diminishing return in the group effort $Z = x_i + (n-1)\tilde{x}_i$:

$$P = bZ - 0.5dZ^2. \tag{S19a}$$

the individual cost is linear in effort $x_i$, and the individual payoff is

$$\pi_i = v_i P - c_i x_i \tag{S19b}$$

as in the Public Goods Game with diminishing returns considered above. However here valuation $v_i$ is not a constant but rather depends on the individual's effort:

$$v_i = \frac{x_i}{Z} \tag{S19c}$$

as in the Tullock contest (12, 13).

One finds that $d\pi_i/dx_i = b - 0.5d[(n-1)\tilde{x} + 2x] - c_i$. Therefore $D_{0,i} = b - c_i, D_1 = (n-1)d/2, D_2 = d$, and

$$\theta_i = \frac{2(b-c_i)}{d(n+1)}.$$

*EGT analysis.* Replacing, as before, the term $(n-1)\tilde{x}$ in the payoff function by $X^-_{i,prev}$, I find that the best response action and the corresponding Nash equilibria are

$$x_{i,\mathrm{BR}} = \max\left(0, \frac{n+1}{2}\ \theta_i - 0.5X^-_{i,prev}\right), \tag{S20a}$$

$$Z_{\mathrm{NE}} = n\ \overline{\theta}, \tag{S20b}$$

$$x^*_i = \theta_i + n(\theta_i - \overline{\theta}). \tag{S20c}$$

Note that for $x^*_i$ to be non-negative, it is required that the minimum $\min(\theta_i) > \frac{n}{n+1}\overline{\theta}$, i.e. variation in $\theta_i$ should quickly decrease with $n$. Once negative values of $x^*_i$ appear, the best response dynamics can become non-equilibrium.

If all individuals have identical coefficients $c_i = c$, then the Nash equilibrium for the total group effort is

$$X_{\mathrm{NE}} = \max\left(0, n\theta\right),$$

while the group effort $X_{\mathrm{opt}}$ maximizing the total group payoff is

$$X_{\mathrm{opt}} = \max\left(0, \frac{b-c}{d}\right),$$

that is, $2n/(n+1)$ times smaller.

*General analysis.* Figure S6 of the main text summarizes the properties of this model. Figure S7 gives examples of the coevolutionary dynamics.

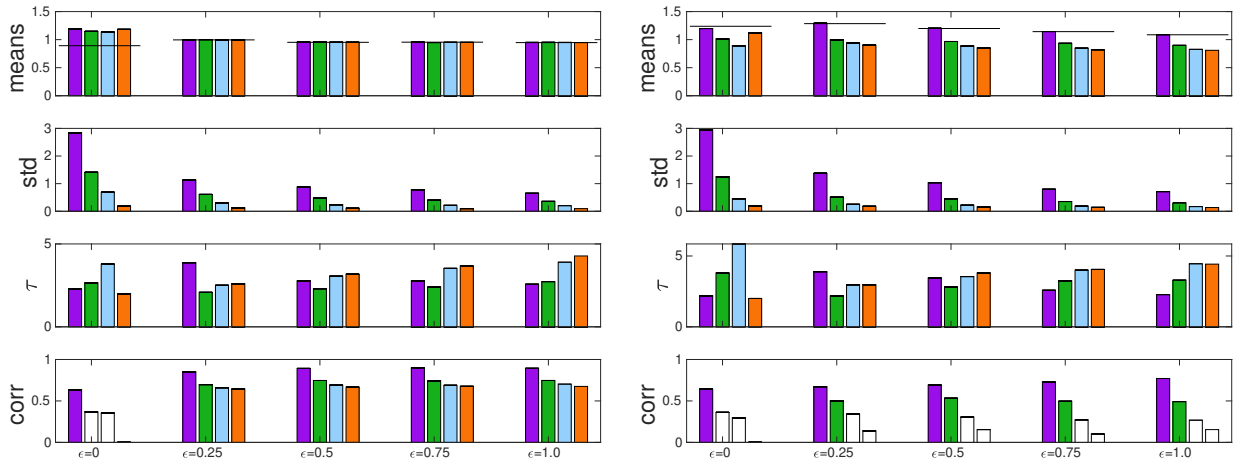



**Figure S6:** Properties of equilibria in the Common Pool Resources game. (a) No external influence. (b) With external influence with $G = 0.5$. From top to bottom: equilibrium means, standard deviations, correlation with $\theta$, and half-time of convergence for $x, y, \tilde{y}$ and $\tilde{x}$, respectively. The thin black horizontal lines show the theoretical predictions for $x$ (given by equation S6). Parameter $\varepsilon$ measures the importance of each of the normative factors relative to material payoffs. Group size $n = 20$. Parameters: $b_i = 10$ for each $i$ while $c_i$ and $d_i$ are drawn from lognormal distributions with mean 1 and standard deviation 0.1. Initial values of $y, \tilde{y}$ and $\tilde{x}$ were chosen randomly and independently from a uniform distribution on $[0, 0.1]$. Statistics are calculated over 100 last time steps over 40 independent runs each of length 1,000 time steps.

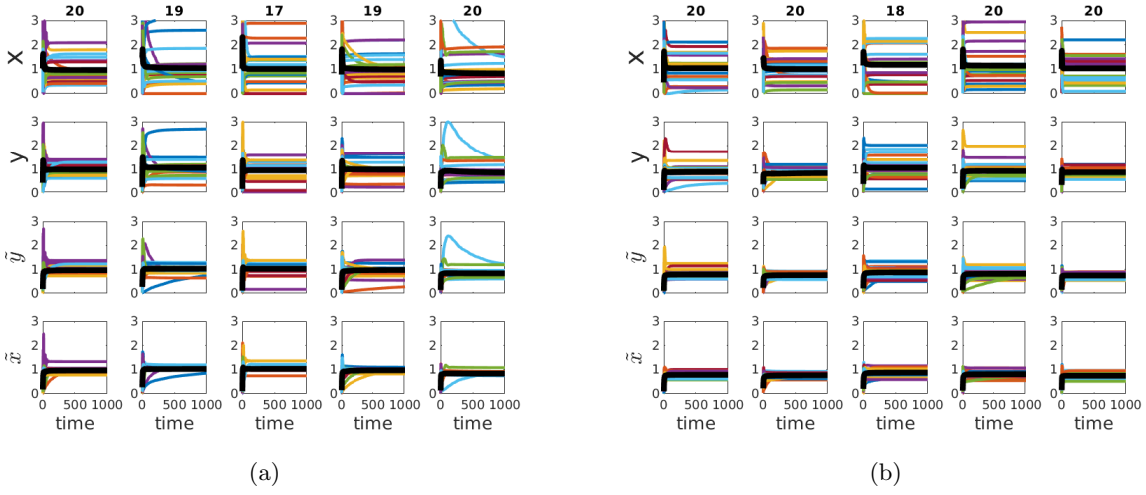



**Figure S7:** Examples of coevolutionary dynamics in the Common Pool Resource game corresponding to Figure S6. (a) Five runs with no external influence. (b) Five runs with external influence with $G = 0.5$. $\varepsilon = 1$. Different colors show different individuals. The thick black lines show the group averages. The numbers on top show the number of contributing individuals at the last time step.

**Tragedy of the Commons Game with quadratic costs**

In the Tragedy of the Commons games, individuals exploit a resource getting a benefit which increases with individual effort $x_i$ but sharing a cost of its exploitation which increases with the total group effort $X$ (14).

Assume that individual benefit is linear in individual effort $x_i$ but the cost is quadratic in group effort $X$:

$$\pi = b_i x_i - 0.5 c_i Z^2.$$

In this model, $d\pi/dx_i = b_i - c_i[(n-1)\tilde{x} + x_i]$ so that $D_0 = b_i, D_1 = c_i(n-1), D_2 = c_i$ and

$$\theta_i = b_i/(c_i n).$$

*EGT analysis.* Here the best response action is $x_{\text{BR},i} = \max(0, n\theta_i - Z^-_{i,prev})$. In the symmetric version of this model when all $\theta_i$ values are the same, the Nash equilibrium for the total group effort is

$$Z_{\text{NE}} = n\theta = b/c.$$

The total group effort maximizing the total group payoff is

$$Z_{\text{opt}} = b/(cn),$$

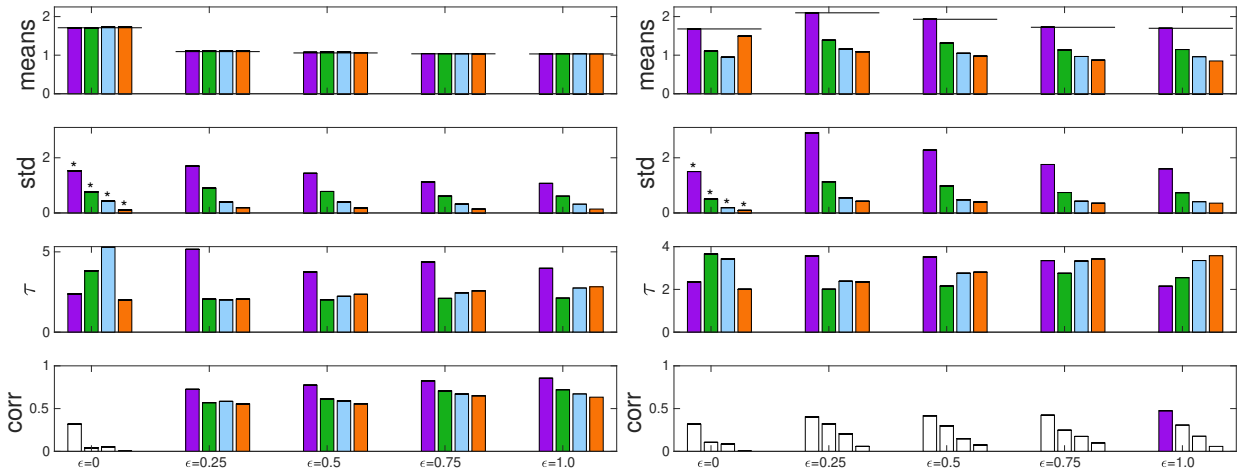
that is, $n$ times smaller. Individual contributions can take any values as long as they sum up to $Z_{\text{NE},sym}$.

In the asymmetric case, when benefit-to-cost ratios $b_i/c_i$ are different, only an individual with the largest value of $\theta_i$ will make an effort $\theta_i$ so that the group effort is

$$Z^* = \max(\theta_i).$$

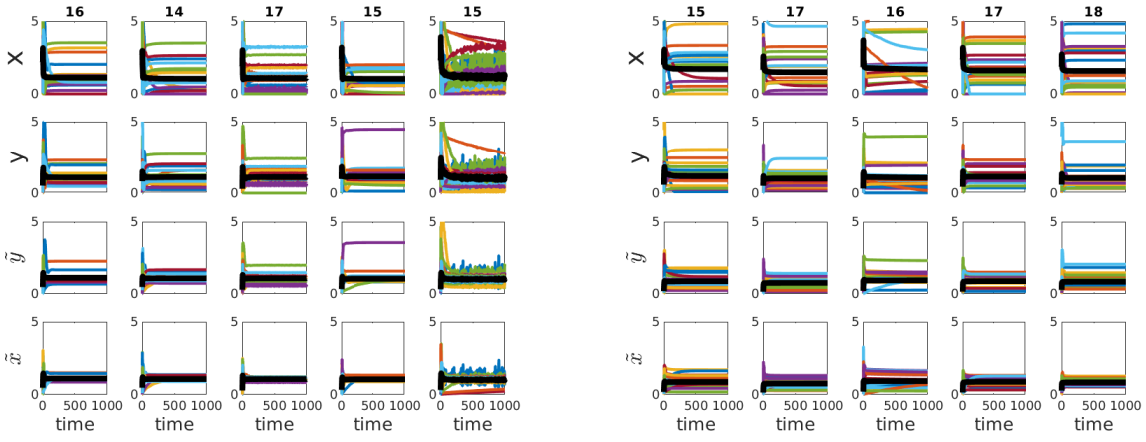
*General analysis.* Figure S8 summarizes the properties of this model.

Figure S9 show sample trajectories corresponding to Figure S8 with $\varepsilon = 1$.




(a) (b)

**Figure S8:** Properties of equilibria in the Tragedy of the Commons game with quadratic costs. (a) No external influence. (b) With external influence promoting decreased effort ($G = 1/n$). From top to bottom: equilibrium means, standard deviations, correlation with $\theta$, and half-time of convergence for $x, y, \tilde{y}$ and $\tilde{x}$, respectively. Parameter $\varepsilon$ measures the importance of each of the normative factors relative to material payoffs. Parameters: $n = 20$, $b_i$ are drawn from a lognormal distribution with mean 1 and variance 0.1, $c_i = 1/n$. Statistics are calculated over 40 independent runs. Note the difference in convergence time-scales between a and b.




(a) (b)

**Figure S9:** Examples of coevolutionary dynamics in the Tragedy of the Commons game with quadratic costs corresponding to Figure **??**. (a) Five runs with no external influence. (b) Five runs with external influence promoting decreased effort at $G = 1/n$. $\varepsilon = 1$. Different colors show different individuals. The thick black lines show the group averages. The numbers on top show the number of contributing individuals at the last time step.

**Tragedy of the Commons game with diminishing returns**

Alternatively assume that individual benefit shows a diminishing return while the cost term is linear in group effort $Z$. Then the payoff function is

$$\pi = (b_i x_i - 0.5 d_i x_i^2) - c_i Z,$$

where $b_i, d_i$ and $c_i$ are individual benefit and cost parameters.

In this game, $d\pi/dx_i = b_i - c_i - d_i x_i$. Therefore $D_0 = b_i - c_i, D_1 = 0, D_2 = d_i$ and

$$\theta_i = \frac{b_i - c_i}{d_i}.$$

*EGT analysis.* Here the best response action is

$$x_{\mathrm{BR},i} = \max(0, \theta_i),$$

which is also the Nash equilibrium.

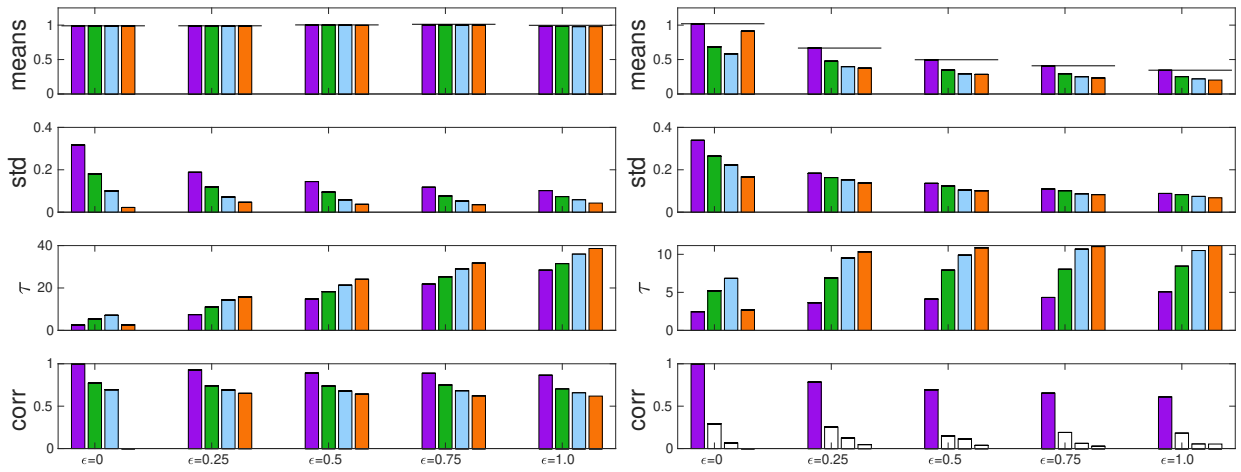
In the symmetric version of this game when all coefficients are the same (i.e. $c_i = c, b_i = b, d_i = d$, the Nash equilibrium for the individuals effort is

$$x_{\mathrm{NE},sym} = \frac{b - c}{d}.$$

The individual effort maximizing the total group payoff is

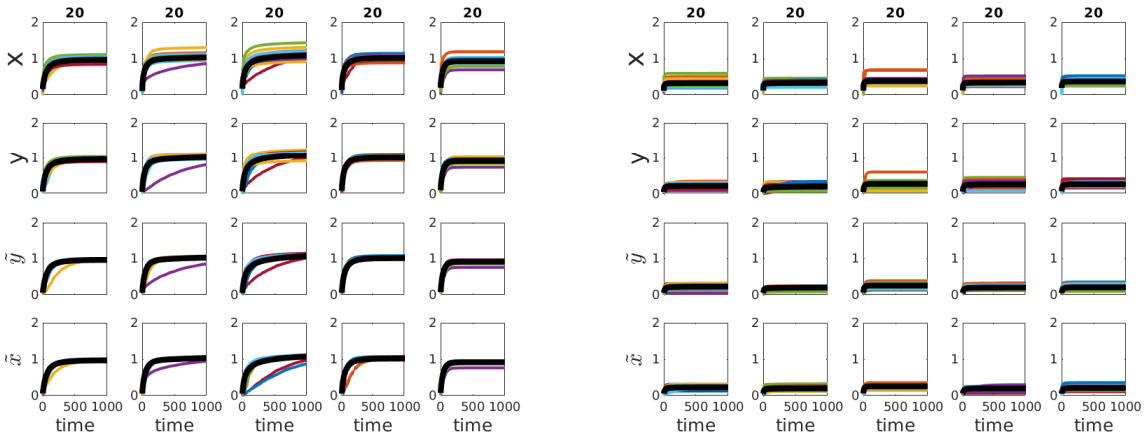
$$x_{\mathrm{opt}} = \frac{b - cn}{d}.$$

*General analysis.* Figure S10 summarizes the behavior of this model. With no external authority, parameter $\varepsilon$ has not effect on average behavior. In contrast to the previous case, individuals respond to the authority and decrease their efforts. The larger $\varepsilon$, the bigger the response. Figure S11 show sample trajectories of the general model corresponding to Figure S10 with $\varepsilon = 1$.




(a) (b)

**Figure S10:** Properties of equilibria in the Tragedy of the Commons game with diminishing returns. (a) No external influence. (b) With external influence promoting decreased effort at $G = 1/n$. From top to bottom: equilibrium means, standard deviations, correlation with $\theta$, and half-time of convergence for $x, y, \tilde{y}$ and $\tilde{x}$, respectively. The thin black horizontal lines show the theoretical predictions for $x$ (given by equation S6). Parameter $\varepsilon$ measures the importance of each of the normative factors relative to material payoffs. Parameters: group size $n = 20$, parameters $b_i$ are drawn from a lognormal distribution with mean $n + 1$ and standard deviation $0.1 \times \sqrt{n+1}$, $c_i = 1, d_i = n$.




(a) (b)

**Figure S11:** Examples of coevolutionary dynamics in the Tragedy of the Commons game with diminishing returns. (a) No external influence. (b) Five runs with external influence promoting decreased effort at $G = 1/n$. $\varepsilon = 1$. Different colors show different individuals. The thick black lines show the group averages. The numbers on top show the number of contributing individuals at the last time step.

**Trade-offs between public and private production**

In this game (15–17, 6) the payoff function is a sum of two components coming from public and private production efforts:

$$\pi_i(x_i) = v_i \underbrace{(BZ - 0.5DZ^2)}_{\text{collective production}} + \underbrace{b_i y_i - 0.5 d_i y_i^2}_{\text{private payoff}},$$

where $Z = \sum x_j$, $x_i$ is the contribution to collective production and $v_i$ is the share/valuation of this production for individuals $i$ . The effort not invested in public production, $y_i = R_i - x_i$, where $R_i$ is a constant endowment of individual $i$, is invested in private production; $b_i$ and $d_i$ are the corresponding benefit and cost coefficients.

Following (6), assume egalitarian division of pubic goods (i.e. $v_i = 1/n$) and that $b_i/d_i$ is a constant.

Then

$$\frac{d\pi_i}{dx_i} = \underbrace{v_i B - b_i + d_i R_i}_{D_0} - \underbrace{v_i D(n-1)}_{D_1}\ \tilde{x} - \underbrace{(d_i + v_i D)}_{D_2}\ x.$$

In this model,

$$\theta_i = \frac{v_i B - b_i + d_i R_i}{d_i + v_i D n} = \frac{R_i + \frac{v_i B - b_i}{d_i}}{1 + n\frac{v_i D}{d_i}} \equiv \frac{\lambda_i}{1 + n\kappa_i}.$$

with the obvious meaning of $\lambda_i$ and $\kappa_i$.

*EGT analysis.* The best response action for individual $i$ is

$$x_{\text{BR},i} = \frac{\lambda_i - \kappa_i X_i^-}{1 + \kappa_i}. \tag{S21a}$$

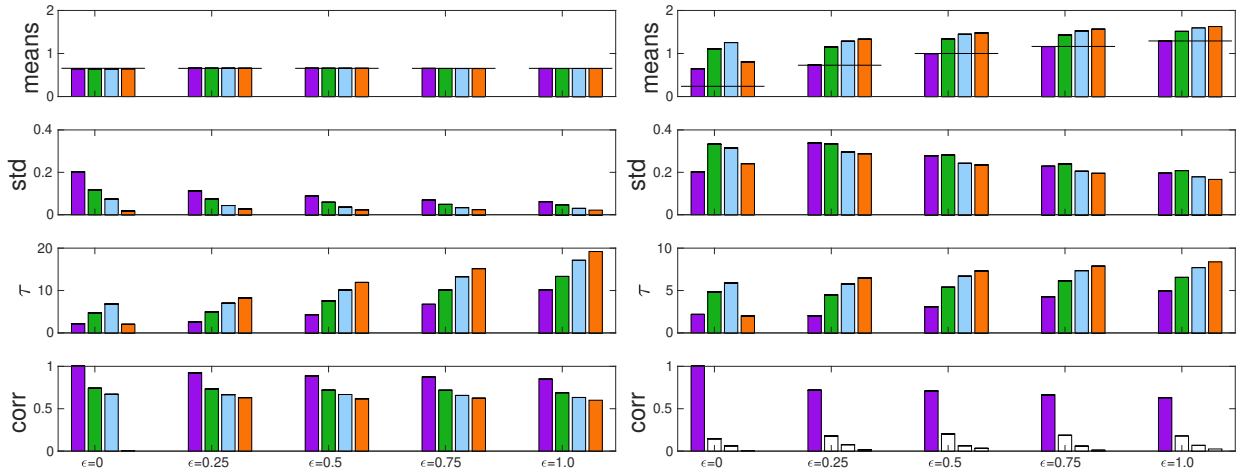
From its meaning, $x$ must stay between 0 and $R$. Generalizing (6) approach under the assumption that all $x_{\text{BR},i} > 0$, I find that the total group effort at the Nash equilibrium can be written as

$$Z_{\text{NE}} = \frac{\sum \lambda_i}{1 + \sum \kappa_i}. \tag{S21b}$$

while the individual effort is

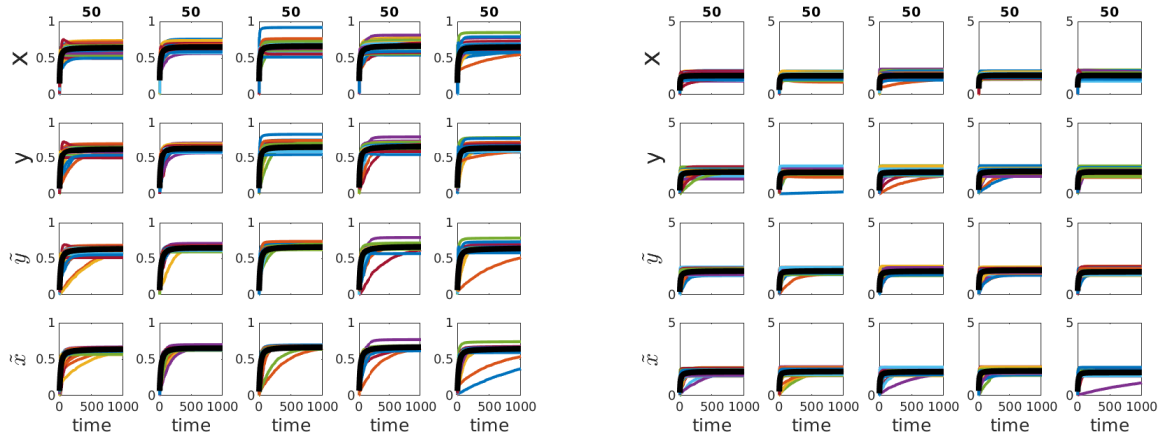
$$x_{i,\text{NE}} = \lambda_i - \kappa_i Z_{\text{NE}}. \tag{S21c}$$

*General analysis.* Figure S12 summarizes the properties of this model. Figure S13 shows sample trajectories of the general model corresponding to Figure S12 with $\varepsilon = 1$.




(a) (b)

**Figure S12:** Properties of equilibria in the Trade-offs game. (a) No external influence. (b) With external influence promoting decreased contribution to private production at $G = 1/n$. From top to bottom: equilibrium means, standard deviations, correlation with $\theta$, and half-time of convergence for $x, y, \tilde{y}$ and $\tilde{x}$, respectively. The thin black horizontal lines show the theoretical predictions for $x$ (given by equation S6). Parameter $\varepsilon$ measures the importance of each of the normative factors relative to material payoffs. Parameters: $n = 50$, $b_i$ are drawn from a lognormal distribution with mean 1 and variance 0.1, $B_m = 1, D = 1, d = 1, R = 2$ while parameters $v_i$ are drawn from a broken stick distribution on $[0, 1]$.




(a) (b)

**Figure S13:** Examples of coevolutionary dynamics in the Trade-offs Game corresponding to Figure S12. (a) No external influence. (b) Five runs with external influence promoting decreased contribution to private production at $G = 1/n$. $n = 20, \varepsilon = 1$. Different colors show different individuals. The thick black lines show the group averages. The numbers on top show the number of contributing individuals at the last time step.

**Games with linear payoff functions: Dictator, Take-or-Give, Rule Obedience, and Public Goods**

Linear payoff functions emerge in a number of simple games commonly used in experimental economics research. Some examples are given next.

*Dictator game.* Here an individual with an endowment $R$ decides on how much to give to another individual. If $x_i$ is the donation, then the payoff function is $\pi(x_i, \tilde{x}_i) = R - x_i$, so that $d\pi_i/dx_i = -1$ and $D_{0,i} = -1$.

*Take-or-Give game.* In this game (18), each individual decided on whether to contribute to a pool of money marked to be given to a charity ($x_i > 0$) or take the money from this pool for personal use ($x_i > 0$). One can write the payoff as $\pi(x_i, \tilde{x}_i) = R - x_i$, so that $d\pi_i/dx_i = -1$ and $D_{0,i} = -1$.

*Rule Obedience game.* In this game designed and studied by (19, 20) individuals can follow verbal instructions (such as "wait for a crosswalk light to turn green") and earn a certain reward or ignore instructions and get a higher reward. Let $x_i$ be the waiting time. Then the payoff function in this game can be written as $\pi(x_i, \tilde{x}_i) = R - x_i$, so that $d\pi_i/dx_i = -1$ and $D_{0,i} = -1$.

*Linear Public Goods game.* In this classical game, the payoff function is

$$\pi(x_i, \tilde{x}_i) = v_i bZ - c_i x_i, \tag{S22}$$

where $b, v_i$ and $c_i$ are constant parameters. Then $d\pi_i/dx_i = bv_i - c_i$. A standard assumption in behavioral studies is that $D_{0,i} = bv_i - c_i < 0$.

In all these games, I predict that in the absence of additional forces, contributions $x_i$ and attitudes $y_i$ and beliefs $\tilde{y}_i, \tilde{x}_i$ will evolve to the minimum values, i.e. zero. However in the presence of an external influence, the equilibrium contribution can be positive.

Figure S14a illustrates the properties of this model when $G = 2$. Figure S15 gives examples of corresponding trajectories.

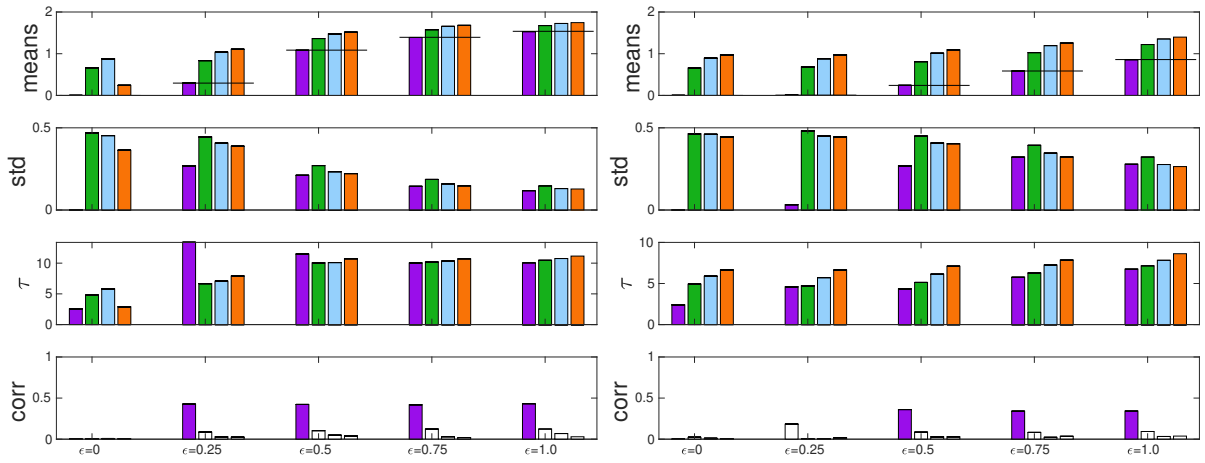



**Figure S14:** (a) Linear Public Goods game with external influence with $G = 2$. Parameters: group size $n = 100, D_1 = D_2 = 0$ for all $i$ while $D_0$ are drawn from a uniform distribution on $[-2, 0]$. (b) Continuous Prisoner's Dilemma game with external influence with $G = 2$. Parameters: group size $n = 50, D_2$ for all $i$ while parameters $D_0, D_1$ are drawn from lognormal distributions with mean $-1$ and 1, respectively, and standard deviation 0.1. These expectations arise if the expectations of $S, P, R$ and $T$ and $0, 1, 3$ and 5, respectively. From top to bottom: equilibrium means, standard deviations, correlations with $D_0$, and the half-time of convergence for $x, y, \tilde{y}$ and $\tilde{x}$, respectively. The thin black horizontal lines show the theoretical predictions for $x$ (given by equation S6). Parameter $\varepsilon$ measures the importance of each of the normative factors relative to material payoffs. Statistics are calculated over 100 last time steps over 40 independent runs each of length 1,000 time steps.

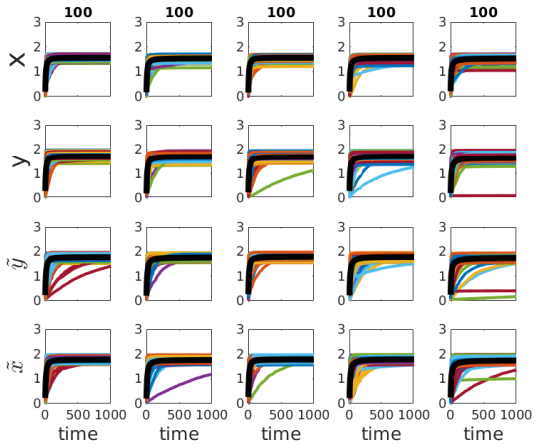



**Figure S15:** Examples of coevolutionary dynamics in the Linear Public Goods Game with external influence at $G = 2$ corresponding to Figure S14a. $n = 100, \varepsilon = 1$. Different colors show different individuals. The thick black lines show the group averages. The numbers on top show the number of contributing individuals at the last time step. Parameters $D_0$ are chosen from a uniform distribution on $[-2, 0]$. $D_1 = D_2 = 0$ for all $i$.

### Continuous Prisoner's Dilemma game

Ref.(21) introduced a continuous variant of the Prisoner's Dilemma which he called the Trader's Dilemma. In this game, each of the two players chooses an effort $x$ within a unit interval $[0, 1]$. The payoff to the player $A$ who makes effort $x_A$ against player $B$ who makes effort $x_B$ can be written as

$$\pi(x_A, x_B) = x_A x_B R + x_A(1 - x_B)S + (1 - x_A)x_B T + (1 - x_A)(1 - x_B)P.$$

Parameters $R, S, T, P$ correspond to the reward, sucker's pay, temptation and punishment payoffs in the standard Prisoner's Dilemma (with $T > R > P > S$). One interpretation of this game is that the players are trade partners. One of them bring a box of rice, the other a box of beans. An action consists of exchanging boxes filled with a certain amount of merchandise. Complete cooperation corresponds to bringing a box completely filled with the promised merchandise. Complete defection corresponds to bringing an empty box.

Adopting this model to my framework, player $i$ will expect a payoff which can be written as

$$\pi(x_i, \tilde{x}_i) = x_i \tilde{x}_i R_i + x_i(1 - \tilde{x}_i)S_i + (1 - x_i)\tilde{x}_i T_i + (1 - x_i)(1 - \tilde{x}_i)P_i.$$

where I allow for heterogeneity in players payoffs. In this model,

$$D_0 = S_i - P_i < 0, D_1 = T_i - R_i - P_i + S_i, D_2 = 0.$$

In this game, $D_0 - D_1 \tilde{x}_i < 0$ for all $\tilde{x}_i$. Therefore the players will evolve to a state with zero efforts in the standard game theoretic approach. The same outcome is predicted in my model if there is no external influence. [Note that in games of partial cooperation studied by (22), $D_1 > 0$. In these games, defection dominates cooperation, but an intermediate fraction of cooperators would maximize the group payoff.]

Figure S14b of the main text illustrates the properties of this model with external influence with $G = 2$. Figure S16 gives examples of corresponding trajectories.

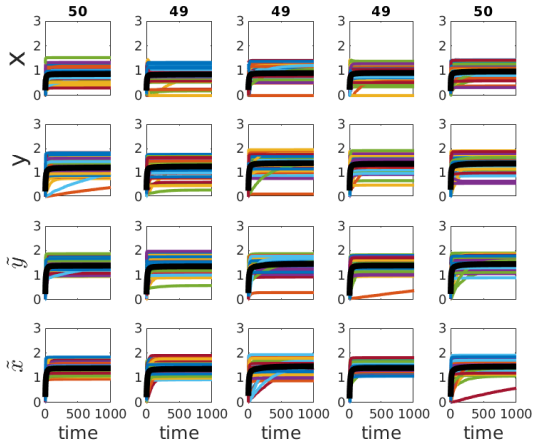



**Figure S16:** Examples of coevolutionary dynamics in the continuous Prisoner's Dilemma game with external influence at $G = 2$ corresponding to Figure S14b. $n = 50, \varepsilon = 1$. Different colors show different individuals. The thick black lines show the group averages. The numbers on top show the number of contributing individuals at the last time step.

**"Us v. nature" game**

This game (7, 23) is similar to the linear public goods game, except that the production function saturates at a constant level as the group efforts increase:

$$P = \frac{Z}{Z + Z_0},$$

where $Z_0$ is a constant half-effort parameter (at $Z = Z_0$, the group produces half of the maximum amount of resource). Because of the non-linearity of this game, my analytical results do not apply and there is no analogue of parameter $\theta$ here.

*EGT analysis.* The best response effort in this game is

$$x_i = \max\left[0, Z_0\left(\sqrt{R_i} - 1\right) - Z^-_{i,prev}\right],$$

where $R_i = b_i/(c_i Z_0)$ is the ratio of the individual benefit and the group's cost at half-effort.

In the symmetric case, the group effort at the Nash equilibrium is

$$Z^*_{sym} = Z_0(\sqrt{R} - 1).$$

In the asymmetric case, only the individual with the largest value of $R_i$ will make an effort while all other individuals will free-ride:

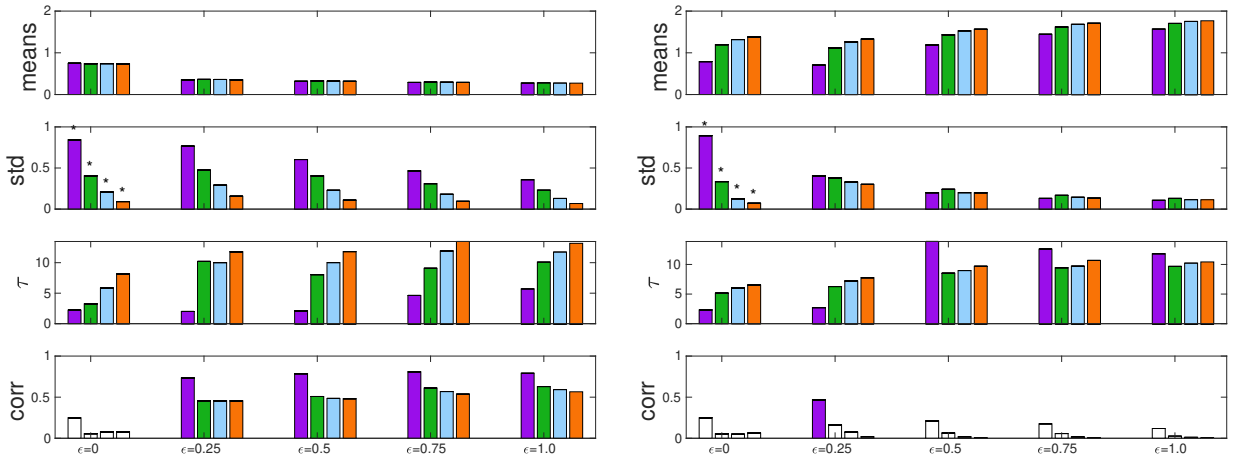
$$Z^*_{asym} = Z_0(\sqrt{\max(R_i)} - 1).$$

*General case.* With additional normative forces, finding the normalized best response effort requires one to solve the cubic:

$$bZ_0 - (c - S_1 + S_2 x)(Z_0 + x + X^-)^2 = 0,$$

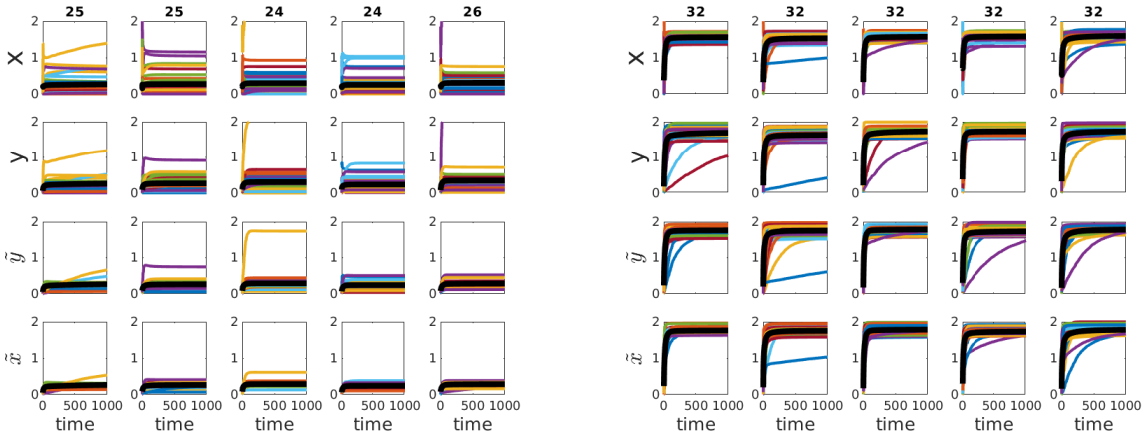
where $S_1 = A_1 y + A_2 \tilde{y} + A_3 \tilde{x} + A_4 G, S_2 = \sum_{j=1}^{4} A_j$. This can be done numerically. Note that all coefficients here except for $Z_0$ are individual-specific.

Figure S17 illustrates the properties of this model. Figure S18 gives examples of corresponding trajectories.




(a) (b)

**Figure S17:** Properties of equilibria in the "Us vs. nature" game. (a) No external influence. (b) With external influence ($G = 2$). From top to bottom: mean, standard deviation, half-time of convergence to an equilibrium $\tau$, and correlation with $\theta$ for $x$ (purple), $y$ (green), $\tilde{y}$ (blue) and $\tilde{x}$ (orange), respectively. Parameter $\varepsilon$ measures the importance of each of the normative factors relative to material payoffs in the utility function. Bars with no color mean the corresponding correlations are statistically insignificant (at 0.05). Group size $n = 32, b = 32n, c = 1, X_0 = n/4$. Parameters $v_i$ are jointly drawn from a broken stick distribution on $[0, 1]$. Statistics are calculated over 40 independent runs.




(a) (b)

**Figure S18:** Examples of coevolutionary dynamics in the "us vs. nature" game. (a) No external influence. (b) Five runs with external influence with $G = 2$. $n = 32, \varepsilon = 1$. The numbers on top show the number of contributing individuals (with $s > 0$) at the last time step.

## References


[1] Gavrilets, S. & Fortunato, L. A solution to the collective action problem in between-group conflict with within-group inequality. *Nature Communications* **5**, article 3526 (doi:10.1038/ncomms4526) (2014).

[2] MacArthur, R. On the relative abundance of bird species. *Proceedings of the National Academy of Sciences USA* **43**, 293–295 (1957).

[3] Kuran, T. & Sandholm, W. H. Cultural integration and its discontents. *Review of Economic Studies* **75**, 201–228 (2008).

[4] Andreoni, J., Nikiforakis, N. & Siegenthaler, S. Predicting social tipping and norm change in controlled experiments. *https://www.nber.org/papers/w27310* (2020).

[5] Esteban, J. & Ray, D. Collective action and the group size paradox. *American Political Science Review* **95**, 663–672 (2001).

[6] McGinty, M. & Milam, G. Public goods provision by asymmetric agents: experimental evidence. *Soc Choice Welf* **40**, 1159–1177 (2013).

[7] Gavrilets, S. Collective action problem in heterogeneous groups. *Proceedings of the Royal Society London B* **370**, 20150016 (2015).

[8] Calabuig, V., Olcina, G. & Panebianco, F. Culture and team production. *Journal of Economic Behavior and Organization* **149**, 32–45 (2018).

[9] Anderson, S. P., Goeree, J. K. & Hol, C. A. A theoretical analysis of altruism and decision error in public goods games. *Journal of Public Economics* **70**, 297–323 (1998).

[10] Apesteguia, J. & Maier-Rigaud, F. P. The tole of rivalry: Public goods versus common-pool resources. *Journal of Conflict Resolution* **50**, 646–663 (2006).

[11] Walker, J. M., Gardner, R. & Ostrom, E. Rent dissipation in a limited-access common-pool resource: Experimental evidence. *Journal of Environmental Economics and Management* **19**, 203–211 (1990).

[12] Tullock, G. Efficient rent seeking. In Buchanan, J. M., Tollison, R. D. & Tullock, G. (eds.) *Toward a theory of the rent-seeking society*, 97–112 (Texas A & M University, College Station, 1980).

[13] Konrad, K. *Strategy and Dynamics in Contests* (Oxford University Press, Oxford, 2009).

[14] Hardin, G. Tragedy of commons. *Science* **162**, 1243–1248 (1968).

[15] Willinger, M. & Ziegelmeyer, A. Framing and cooperation in public good games: an experiment with an interior solution. *Economics Letters* **65**, 323–328 (1999).

[16] Willinger, M. & Ziegelmeyer, A. Association strength of the social dilemma in a public goods experiment: An exploration of the error hypothesis. *Experimental Economics* **4**, 131–144 (2001).

[17] Laury, S. K. & Holt, C. A. Chapter 84. Voluntary provision of public goods: Experimental results with interior nash equilibria. vol. 1 of *Handbook of Experimental Economics Results*, 792 – 801 (Elsevier, 2008). URL `http://www.sciencedirect.com/science/article/pii/S1574072207000844`.

[18] Bicchieri, C., Dimant, E., Gachter, S. & Nosenzo, D. Observability, social proximity, and the erosion of norm compliance. *papers.ssrn.com/sol3/papers.cfm?abstractid=3355028* (2020).

[19] Kimbrough, E. O. & Vostroknutov, A. Norms make preferences social. *Journal of the European Economic Association* **14**, 608–638 (2016).

[20] Kimbrough, E. O. & Vostroknutov, A. A theory of injunctive norms (2019).

[21] Verhoeff, T. The trader's dilemma: A continuous version of the prisoner's dilemma. Tech. Rep., Faculty of Mathematics and Computing Science, Technische Universiteit Eindhoven, The Netherlands (1998).

[22] Stark, H.-U. Dilemmas of partial cooperation. *Evolution* **64**, 2458–2465 (2010).

[23] Gavrilets, S. & Richerson, P. J. Collective action and the evolution of social norm internalization. *Proceedings of the National Academy of Sciences USA* **114**, 6068–6073 (2017).