Evolutionary Human Sciences (2021), **3**, e44, page 1 of 22
doi:10.1017/ehs.2021.40

RESEARCH ARTICLE

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
*Corresponding author: gavrila@utk.edu

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, organisational). Also important are dynamically changing personal perceptions 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 unifying modelling framework describing the joint dynamics of actions and attitudes of individuals and their beliefs about the actions and attitudes of their groupmates. I consider which norms get internalised 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 behaviour in a certain direction can, counter-intuitively, have an opposite effect on individual behaviour. I consider how various factors can affect differences between groups and societies in the tightness/looseness of their social norms. I show that the most important factors are social heterogeneity, societal threat, effects of 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 behaviour and historical and current social processes, and in developing more efficient policies aiming to modify social behaviour.

Keywords: Cooperation; conflict; cultural evolution; social evolution; mathematical models

Social media summary: A unifying modelling framework predicts the effects of material, social and cognitive forces on human behaviour and beliefs.

Introduction

Human groups at various scales of social organisation repeatedly face situations where engaging in an individually costly collective action or refraining from an individually beneficial behaviour can help bring larger benefits or avoid certain disastrous outcomes. Examples range from cooperating in hunting or agricultural production in small-scale societies to mobilising against social injustice to modifying the collective behaviour 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 the collective action problem (Olson, 1965; Pecorino, 2015), the tragedy of the commons (Hardin, 1968, Ostrom, 2000), social traps (Platt, 1973), the many-person Prisoner's Dilemma (Schelling, 1978; Molander, 1992) and the collective risk dilemma (Milinski et al., 2008).

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, organisational). Human actions also depend on their personal perception of the situation and on beliefs about the actions and expectations of their peers (R. L. Cialdini et al., 1990; Troyer and Younts, 1997; Bicchieri, 2006). 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 behaviours conflict) can cause changes in behaviours but also in attitudes or beliefs (Festinger, 1957). To predict the intentions and beliefs of others, people may use the 'theory of mind' (Premack and Wodruff, 1979; Apperly, 2010) and social projection, which is the tendency to assume that others are similar to oneself (Krueger, 2007). Therefore changing personal attitudes can also change predictions about others.

Owing to this complexity, modelling human behaviour is notoriously difficult. Nevertheless several approaches successfully capturing certain aspects of human decision-making have been developed. These include classical (Fudenberg and Tirole, 1992), evolutionary (Sandholm, 2010), mean-field (Tembine, 2017) and quantum (Piotrowski and Sladkowski, 2003; Siopsis et al., 2018) game theories focusing on the effects of material payoffs, social influence models focusing on the dynamics of consensus formation (or fragmentation) in social networks as a result of social learning and imitation (DeGroot, 1974; Watts, 2002; Friedkin et al., 2016; Redner, 2019; Galesic and Stein, 2019; Zino et al., 2020; Kashima et al., 2021), models of strategic deliberation (Golman et al., 2020), models of normative behaviour (Azar, 2004; S. Gavrilets and Richerson, 2017; S. Gavrilets, 2020) and models of foresight (Perry et al., 2018; Perry and Gavrilets, 2020). Each of these approaches concentrates on specific forces shaping human behaviour and beliefs while neglecting many other important factors.

Here I will build on this earlier work to develop a novel theoretical approach explicitly integrating multiple material, cognitive, emotional and social forces shaping human behaviour. I posit that individuals are motivated by both material factors and 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. In my theoretical approach, the individual's actions and beliefs are influenced by their social environment as well as by certain internal psychological processes. Mathematically, these assumptions are implemented by adding several additional components besides material payoffs to the utility function and by writing down coupled equations specifying the dynamics of attitudes and beliefs about others.

My approach aims to shed theoretical light on a number of important questions: how can individuals find the right action when facing social dilemmas? Which factors (material, social, psychological) are most important in their decisions? What happens to their preferences, beliefs and behaviours as social interactions unfold dynamically? Which social norms get internalised? Which factors control individual beliefs about others? How different are the effects of peer influences from those of an external authority? What are the effects of between-individual differences in physical, social and psychological characteristics on group behaviour? How robust are game-theoretic predictions on short and long time scales in the presence of non-material influences and belief dynamics? Which psychological forces are most powerful? What are the cultural effects on individual and group behaviour? How is the tightness or looseness of social norms related to various environmental, social and psychological forces? My approach also offers a way to measure and compare the relative strengths of different factors affecting individual actions and beliefs.

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' (Thomas, 1928). 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 some psychological processes governing their thinking and emotions (Wood, 2000; Albarracin and Shavitt, 2017). The general structure of my model is illustrated in Figure 1.

https://doi.org/10.1017/ehs.2021.40 Published online by Cambridge University Press

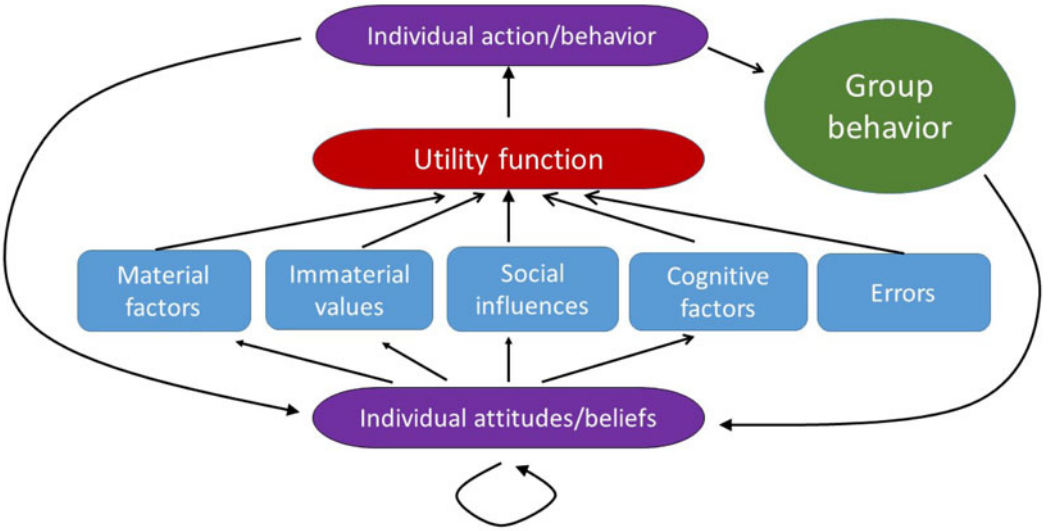

Figure 1. Model structure. The model integrates material factors, nonmaterial values, social influences (both by peers and an external authority), cognitive factors and errors (the blue boxes) into a general utility function (the red shape) which individuals attempt to maximise when making decisions (the top violet shape). Individual behavior is a part of group behavior (the green shape). Individual actions taken and observed group behavior as well as previous attitudes and beliefs feed back into updated 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 individuals' 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, the information they receive and the emotions they experience.

Below after introducing my approach and describing main results, I illustrate them by considering different types of social interactions including those stylised by Coordination, Public Goods, Tragedy of the Commons, Common Pool Resource, continuous Prisoner's Dilemma, Dictator and 'Us vs. nature' games. At the end, I discuss the implications of my results for empirical and theoretical research on the behaviour and beliefs of individuals and groups.

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 by peers (or an external authority) but they also prefer to do what they personally think is appropriate. Individuals are bounded rational (Simon, 1957). They 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 behaviours 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 characterised 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 (Hedden and Zhang, 2002; Goodie et al., 2012; Jamali et al., 2021). 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 (Bicchieri, 2006; Bicchieri et al., 2020; Szekely et al., 2021). Below I

https://doi.org/10.1017/ehs.2021.40 Published online by Cambridge University Press

will use the terms 'attitude' and 'personal norm' interchangeably. Individuals are also subject to influence by an external authority promoting a particular action G. I assume that x, y, $\tilde{x}$, $\tilde{y}$, G are non-negative. I note that recent work directly measures these variables in behavioural experiments (Szekely et al., 2021; d'Adda et al., 2020; Andreozzi et al., 2020; Basić and Verrina, 2020; Kölle and Quercia, 2021). Individuals form their beliefs about others on the basis of the 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 maximise the (subjective) utility function u. I write it as a sum of several terms:

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

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 owing to cognitive dissonance (Festinger, 1957) 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 behaviour of the focal individual (R. L. Cialdini et al., 1990; Bicchieri, 2006). The fourth term in equation (1) captures the psychic costs of non-conformity with the expected actions of others (R. B. Cialdini and Goldstein, 2004; Song et al., 2012). 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 (French and Raven, 1959; R. B. Cialdini and Goldstein, 2004). Some studies show stable variation between people in following the 'rules' (Kimbrough and Vostroknutov, 2016, 2018).

I assume that 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 behaviour. Parameters A_2 and A_3 may depend on the group size, so that individuals whose actions deviate from the expected behaviour 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

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

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.]

https://doi.org/10.1017/ehs.2021.40 Published online by Cambridge University Press

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

$$\theta = \frac{D_0}{D_1 + D_2}, \tag{3}$$

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, θ can also be viewed as a measure of the material benefit-to-cost ratio; in some games θ is the Nash equilibrium for the individual effort. As I show below, the distribution of θ 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 maximising the utility function u of the focal individual can be found by computing the derivative $\partial u/\partial x$. 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, …, B_4 are re-scaled individual-specific parameters measuring the effects of material and non-material forces on individual actions (see the Supporting Information). 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 that the next time step, X is the average action of groupmates as observed by the focal individual (so that different individuals are characterised by different X) and C_{ij} represents 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 belief that others are probably similar to themselves (Premack and Wodruff, 1979; Krueger, 2007). The 'logic constraints' term reduces a mismatch between the ego's beliefs about actions and beliefs of others (cf. Friedkin et al. 2016). The 'conformity w/ peers' and two 'learning about others' terms move the corresponding attitude and beliefs closer to the observed average behaviour X among peers (Kashima et al., 2015). The 'conformity w/ authority' terms move

https://doi.org/10.1017/ehs.2021.40 Published online by Cambridge University Press

the corresponding attitudes and beliefs closer to the promoted 'standard' G. Note that cognitive dissonance makes individuals 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. Rabin 1994). 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 behaviour. 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 that individual values X are approximately the same (and equal to the actual average action of the group).

Below I will use normalised parameters

$$\alpha_i = \frac{C_{i1}}{\sum_j C_{ij}}, \quad \beta_i = \frac{C_{i2}}{\sum_j C_{ij}}, \quad \gamma_i = \frac{C_{i3}}{\sum_j C_{ij}}$$

with $\alpha_i + \beta_i + \gamma_i = 1$ for all i. Parameter α_i characterises the relative strengths of cognitive factors (i.e. related to the cognitive dissonance, the social projection, and the logic constraint, respectively). Parameters β_i and γ_i characterise the relative strengths of two types of social factors: learning from/about peers and complying with external influences, respectively. All of 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 γ_i while strongly conformist or collectivistic communities may be characterised by higher values of parameters β_i. Parameters B_4 and γ_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 (Fudenberg and Tirole, 1992; Sandholm, 2010; Tembine, 2017; Gomes and Saúde, 2014). Note also that in contrast to standard evolutionary game theory models where individuals myopically choose the 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 (Azar, 2004; S. Gavrilets and Richerson, 2017; S. Gavrilets, 2020). Relatively few existing models consider the joint dynamics of actions (x) and personal norms (y). For example in Rabin (1994), Kuran and Sandholm (2008) and Calabuig et al. (2018), utility functions include material payoffs $\pi(x)$ as well cognitive dissonance and conformity with peers terms. Kuran and Sandholm (2008) and Calabuig et al. (2018) describe the dynamics of personal norms y allowing for the effects of cognitive dissonance and conformity with peers. Bisin and Verdier (2001) and Cheung and Wu (2018) consider the inter-generation evolution of preferences. However all of these papers 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 (DeGroot, 1974; Watts, 2002; Friedkin et al., 2016; Redner, 2019; Galesic and Stein, 2019; Kashima et al., 2021; Centola et al., 2005) 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 (Olsson et al., 2020). Focusing on dyadic interactions, Golman et al. (2020) model how individuals update their values of y and $\tilde{x}$ on the basis of payoffs received. Y. N. Gavrilets (2003) considered similar models but with the 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

https://doi.org/10.1017/ehs.2021.40 Published online by Cambridge University Press

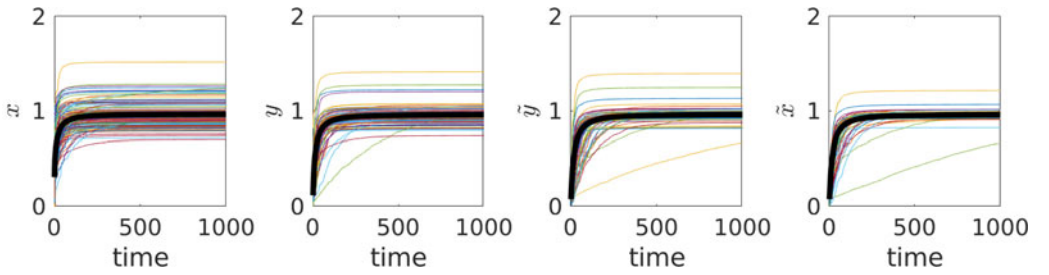

Figure 2. The dynamics of x, y, $\tilde{y}$ and $\tilde{x}$ of individual players in the Coordination Game with no external influence observed in a single run of agent-based simulations. The thick black lines show the group averages. Group size $n = 100$. Parameters are chosen randomly and independently from certain distributions (as described in the Supporting Information) so that the mean value of θ 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].

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

Results

Long-term behaviour

Equations (4) and (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 Supporting Information, I find an approximation for this equilibrium. Here I summarise what happens in several important special cases. For the rest of this 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 individuals (so that coefficient θ 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, α_i, β_i). The value of x at equilibrium is the one maximising 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

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

https://doi.org/10.1017/ehs.2021.40 Published online by Cambridge University Press

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

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

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

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

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

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

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

A composite parameter η, which depends on Bs and αs, is defined by equation (S4c) in the Supporting Information. Parameters θ, α_1, α_2, α_3 and η are individual specific while X and $\bar{\theta}$ 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 θ, η and the strength of cognitive factors α_1, α_2, α_3 are low, the mean values of x, y, $\tilde{y}$ and $\tilde{x}$ are all approximately equal to $\bar{\theta}$. That is, on average individual preferences and beliefs align with actions.

One can also approximate the corresponding variances (see the Supporting Information). These approximations show 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. A factor contributing to this pattern is that social influences act to align individual beliefs while differences in material payoffs are not affected by social influences and remain present. Similarly, the correlation with material benefits (characterised by parameter θ) 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 Supporting Information). The predictions about the properties of long-term equilibria made in this section are testable.

Examples

Next I illustrate my results using several games which have been extensively studied using methods of classical game theory, evolutionary game theory and behavioural economics. In experimental studies,

https://doi.org/10.1017/ehs.2021.40 Published online by Cambridge University Press

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 S. Gavrilets 2015 for a review of models of collective action in heterogeneous groups.) I will also allow for differences between individuals in parameters characterising the effects of non-material 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 ε which will vary from 0 to 1. I will scale parameters $A_1,\ldots, A_4$ by multiplying them by ε. 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 θ, and the half-time τ 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 Supporting Information, I also consider several models with linear and quasi-linear payoff functions and a more complex example of a non-linear payoff function. Table S1 in the Supporting Information summarises the games I consider.

Coordination game

Let individuals interact in randomly formed groups. Following Kuran and Sandholm (2008) (see also Andreoni et al., 2021), assume that each player pays a cost if his action deviates from the average action of the group. Without any additional factors, there is a line of equilibria in x that the groups can converge on. Further assume that each individual has a preferred action θ_i and pays a cost proportional to the square of the deviation from θ_i. The corresponding (subjective) payoff function for individual i is

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

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 θ_i defined by equation (3) is exactly θ_i of the payoff function (10).

Evolutionary game theory analysis. Let

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

be the relative strength of conformity pressure for individual i. Assume that parameters θ_i and r_i are chosen randomly and independently from certain distributions. Then there is a single Nash equilibrium effort for individual i which can be approximated as $x_i^* = \theta_i + r_i(\bar{\theta} - \theta_i)$, and the average effort of the group $\bar{x}^* = \bar{\theta}$ (see the Supporting Information).

General case

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

Figure 3 illustrates the equilibria in this model found using agent-based simulations. The evolutionary game theory (EGT) predictions correspond to purple bars for $\varepsilon = 0$. The case of no external

https://doi.org/10.1017/ehs.2021.40 Published online by Cambridge University Press

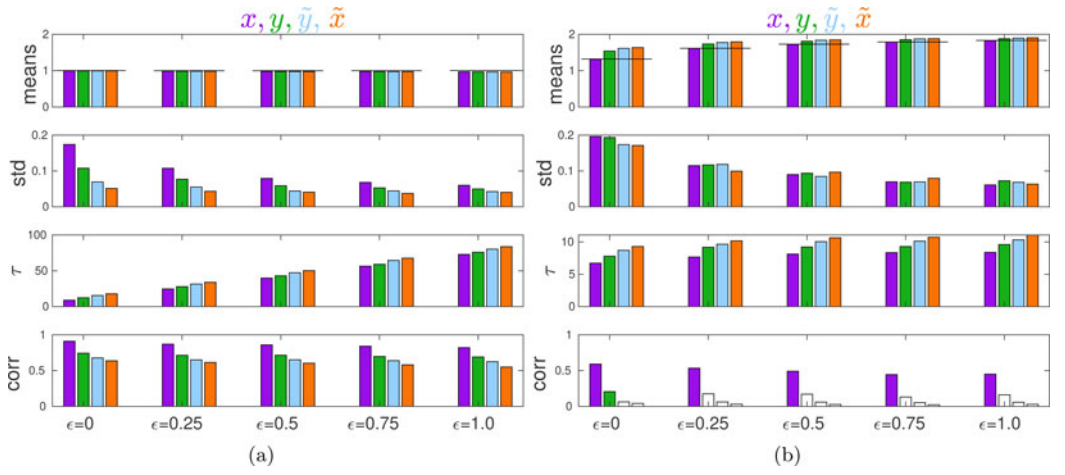

Figure 3. Properties of equilibria in the Coordination Game. (a) No external influence. (b) With external influence ($G = 2$). From top to bottom: mean, standard deviation, half-time of convergence to an equilibrium τ, and Kendall rank correlation with θ for x (purple), y (green), $\tilde{y}$ (blue) and $\tilde{x}$ (orange), respectively. Bars with no colour mean that the corresponding correlations are statistically insignificant (at 0.05). The thin black lines show the theoretical predictions for x. Notice the difference between y-axis scales on graphs for τ. Parameter ε measures the weight of each normative factor relative to material payoffs in the utility function. Group size $n = 100$. Parameters θ_i, c_i, d_i are drawn from log–normal distributions with mean 1 and standard deviation 0.1, so that $\bar{\theta} \approx 1$. Statistics are calculated over the 100 last time steps over 40 independent runs each of length 1000 time steps.

influence was modelled 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 $\bar{\theta}$ 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 $\bar{\theta}$ owing to the psychological processes modelled.
- The standard deviations and correlations with θ are in the order predicted – from the largest for x to the smallest for $\tilde{x}$.
- Increasing the strength ε 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 ε 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 θ.
- The time to convergence to the equilibrium is shorter than that without an external influence and does not depend much on ε.
- 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.

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

https://doi.org/10.1017/ehs.2021.40 Published online by Cambridge University Press

members with ith individual getting value v_iP, where v_i is a constant individual-specific parameter. For example, if each individual gets an equal share, $v_i = 1/n$. Following Calabuig et al. (2018), S. Gavrilets (2015), Esteban and Ray (2001) and McGinty and Milam (2013) 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

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

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

The best response and the Nash equilibrium for the individual effort are equal to θ_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 (Walker et al., 1990; Apesteguia and Maier-Rigaud, 2006), the production function shows a diminishing return in the group effort: $P = bZ - 0.5dZ^2$, where b and d are constant parameters, and Z is the same as defined above. The individual payoff is

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

where c_i is the individual cost coefficient and the resource share going to individual i is proportional to their effort: $v_i = x_i/Z$ as in the Tullock contest model (Konrad, 2009). In this model

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

EGT analysis

In this model, Nash equilibria are $x_{i,NE} = \theta_i + n(\theta_i - \bar{\theta})$. If all individuals have identical coefficients $c_i = c$ and $b > c$, then the Nash equilibrium is $x_{NE} = \theta$, while the individual effort maximising the total group payoff is $x_{\text{opt}} = (b-c)/d$, that is, $2n/(n+1)$ times smaller.

General analysis

Figure 5a shows that with no external influence and positive ε, 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 Supporting Information). The time to convergence is very short. With positive ε, all individual characteristics strongly correlate with the measure θ 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 5b). 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.

https://doi.org/10.1017/ehs.2021.40 Published online by Cambridge University Press

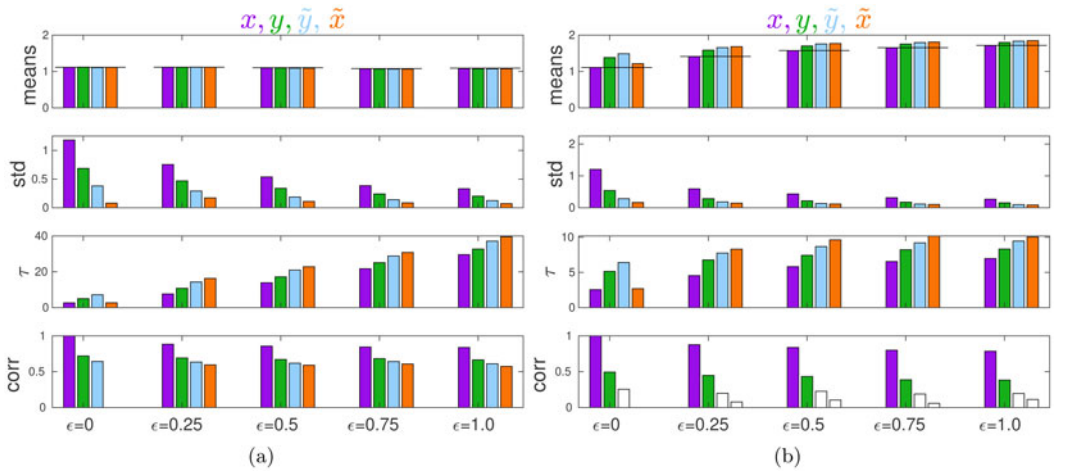

Figure 4. Properties of equilibria in the Public Goods game with quadratic costs. (a) No external influence. (b) With external influence promoting increased effort ($G = 2$). From top to bottom: equilibrium means, standard deviations, half-time of convergence to an equilibrium τ and Kendall correlation with θ for x, y, $\tilde{y}$ and $\tilde{x}$, respectively. The thin black lines show the theoretical predictions for x. Parameter ε 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 log–normal 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.

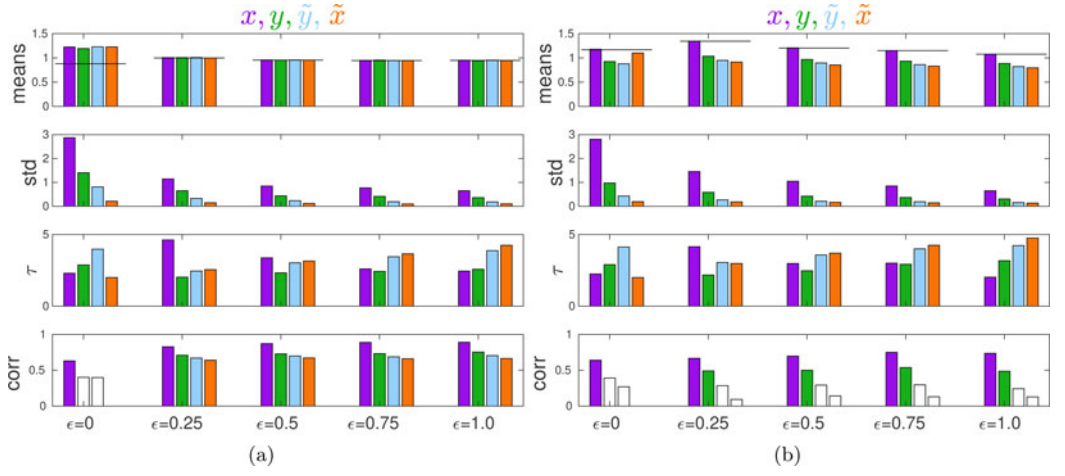

Figure 5. Properties of equilibria in the Common Pool Resources game. (a) No external influence. (b) With external influence promoting decreased, socially optimal effort $G = 0.5$. From top to bottom: equilibrium means, standard deviations, half-time of convergence to an equilibrium τ, and Kendall correlation with θ for x, y, $\tilde{y}$ and $\tilde{x}$, respectively. The thin black horizontal lines show the theoretical predictions for x. Parameter ε 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 log–normal 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 1000 time steps.

Other games

In the Supporting Information, 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 (Willinger and Ziegelmeyer, 1999, 2001; McGinty and Milam, 2013), and an 'Us vs. Nature' game (S. Gavrilets, 2015; S. Gavrilets and Richerson, 2017) show behaviour similar to that of the Public

https://doi.org/10.1017/ehs.2021.40 Published online by Cambridge University Press

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 (Anderson et al., 1998; Apesteguia and Maier-Rigaud, 2006) and a Tragedy of the Commons game with quadratic costs are similar to the Common Pool Resource game (illustrated in Figure 5). 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 (Bicchieri et al., 2020) and the Rule Following game (Kimbrough and Vostroknutov, 2016). 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 behaviour is exhibited by a continuous Prisoner's Dilemma game (Verhoeff, 1998) 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 unifying theoretical approach for modelling the dynamics of social interactions in situations where individuals' personal norms and beliefs about others affect their own actions which in turn causes subsequent adjustments in norms and beliefs. My approach combines evolutionary game theory models focusing on material costs and benefits (Fudenberg and Tirole, 1992; Sandholm, 2010) and an adaptation of social influence models focusing on the dynamics of publicly expressed opinions (Watts, 2002; Friedkin et al., 2016; Redner, 2019) with novel modelling 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 behaviour, my models shed light on two other types of questions: which norms get internalised and which factors control beliefs about others.

Individual characteristics

One of the goals of my approach was to understand the relative importance of different material, social and cognitive factors for individual behaviour and beliefs from the theoretical point of view. A major conclusion of my analysis is that some of these factors are much more important than others. 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 behaviour tends to evolve towards actions maximising 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 internalise the ideas and beliefs that are most beneficial for their material well-being. In a sense, these modelling conclusions align with Marx' postulate that 'material life determines the social, political and intellectual life process in general' (Marx, 1959).

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 (Darwin, 1871; Richerson and Boyd, 2005; Henrich, 2015; Richerson et al., 2021). 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 from both individual and societal perspectives. My models

https://doi.org/10.1017/ehs.2021.40 Published online by Cambridge University Press

predict how individual actions are dispersed around or shifted away from those maximising their personal material payoffs.

Under some conditions the effort of an authority to promote certain behaviour 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. Alternatively, 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 (Diekmann, 1985) – when individuals fail to perform an action they 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 internalise preferences for such acts. My models can potentially be used to better understand obedience to authority such as that studied in Milgram's and Zimbardo's experiments (Milgram, 1963; Haney et al., 1973) or the effects of expected supernatural punishment for violating moral norms in moralising religions (Willard et al., 2020). 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 behaviour (Tankard and Paluck, 2016).

Differences from EGT predictions

Standard EGT models aim to predict human behaviour solely from the expected material payoff. However, the growing understanding in behavioural economics is that certain normative factors must be considered to explain observed behaviour (Szekely et al., 2021; d'Adda et al., 2020; Andreoni et al., 2021; Loewenstein and Molnar, 2018; Fehr and Schurtenberger, 2018; Górges and Nosenzo, 2020; which is a fact well appreciated in social psychology). Allowing for certain normative factors in the utility function shifts the corresponding model predictions away from the Nash equilibrium based on the material payoffs. Recent work has offered empirically based ways to modify the utility function to explain apparently non-rational behaviour (i.e. deviations from Nash equilibria) of individuals in behavioural experiments (d'Adda et al., 2020; Andreozzi et al., 2020; Basić and Verrina, 2020; Kölle and Quercia, 2021).

My modelling framework allows one to treat and contrast the effects of multiple normative factors and psychological mechanisms on behaviour at once, thus generalising earlier work. Moreover one should also expect that the relative strengths of these factors and mechanisms will change as social interactions unfold. My approach offers a general theoretical way for describing these dynamics. It also allows for differences between individuals in various characteristics including psychology which are often disregarded in the EGT models.

Interestingly and importantly, allowing for changing attitudes and beliefs makes Nash equilibria of the EGT relevant again. Specifically my results show that in the absence of external authority, the average behaviour at a long-term equilibrium is exactly as predicted by the EGT (see also Calabuig et al., 2018). This gives some additional confidence in the robustness of some results/conclusions of the EGT. Besides choosing actions as predicted by the EGT on average, individuals are also predicted to develop attitudes and beliefs justifying (or matching) their behaviours.

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 (reviewed in S. Gavrilets, 2015), my models predict that 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.

https://doi.org/10.1017/ehs.2021.40 Published online by Cambridge University Press

As noticed by an anonymous reviewer, the model's prediction that neither conformity with peers nor cognitive dissonance affects long-term average equilibrium behaviour may require us to reevaluate our assumptions about the adaptive function of these mechanisms, at least under conditions modelled here.

Groups

My models allow for scaling up individual behaviour 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 (Szekely et al., 2021). 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 (R. L. Cialdini et al., 1990; R. B. Cialdini and Goldstein, 2004; Bicchieri, 2006). In particular, variable y gives the personal norm of an individual. The average of $\tilde{x}$ specifying the expected average behaviour 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 (S. Gavrilets, 2020). My models shows how these norms become dynamically aligned as social interactions unfold. This process is a subject of recent experimental studies (Eriksson et al., 2015; Tworek and Cimpian, 2016; Lindstrom et al., 2018; Szekely et al., 2021).

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 equation S8 in the Supporting Information). 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 behaviour 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 emphasise social norms and compliance with them (Pelto, 1968). The tight–loose (TL) differences can exist not only between different countries (Gelfand et al., 2011) but also within the same country, e.g. between 50 states in the US (Harrington and Gelfand, 2014) and between 31 provinces in China (Chua et al., 2019). The variation on the TL scale is also observed in non-industrial societies (Jackson et al., 2020). Gelfand et al. (2011), Harrington and Gelfand (2014), Jackson et al. (2020) and Roos et al. (2015) show with data that the TL variation can be explained in terms of the history of threats (e.g. environmental, internal and external warfare) faced by societies and the need to better coordinate collective actions under conditions of threat. Chua et al. (2019) confirm this interpretation but show that cultural tightness also correlates with tighter government control of areas of urbanisation and economic growth, with the strength of religious practices, and the extent of traditionality and group collectivism. Talhelm and English (2020) provided evidence that historically rice-farming societies have tighter social norms worldwide. They explained this by the fact that rice production was very labour intensive and required farmers to coordinate water use and develop strong norms for labour exchange. Using data on small-scale societies, Jackson et al. (2020) showed the importance of two additional factors: cultural complexity (sensu Murdock and Provost 1973) and kinship heterogeneity. Less complex societies and patrilocal societies (in which wives settle near their husband's parents) are tighter.

https://doi.org/10.1017/ehs.2021.40 Published online by Cambridge University Press

All of these analyses are correlational and therefore it cannot be claimed that the factors discussed there cause cultural tightness. However theoretical studies can provide support for causality. Roos et al. (2015) modelled 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 as a measure of the strength of a (descriptive) cooperative norm.

Extending this work, my general approach allows one to study the effects of different factors not only on behaviour 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 θ measuring individual material costs and benefits (equation 3). Variation in θ is high if individuals differ in the roles they play in the society, their abilities, the compensation/valuation of the material benefit produced and 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, stricter norms than more complex societies. My conclusion is thus in line with with the observations that urbanised areas have looser norms than rural areas (Harrington and Gelfand, 2014; Talhelm and English, 2020) and that more complex and heterogeneous societies have looser norms (Jackson et al., 2020).

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 modelled in Roos et al. (2015). Societal threat will however also affect attitudes and beliefs, potentially making them more uniform and tightening culture (Szekely et al., 2021). There are several ways to introduce the effects of an environmental or social threat into my models. One is via a change in the payoff function π. In the Coordination Game, a threat can be modelled 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 in that model, making the actions chosen more similar and consequently making all attitudes and beliefs more homogeneous. In other games with quadratic payoff function and in the Continuous Prisoner's Dilemma game, a societal threat can be modelled as a change in parameters θ_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 non-conformity with peers' action A_3 and non-conformity with authority A_4. Increasing these parameters will decrease η, reducing the 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 behaviour. As discussed above, increasing the perceived cost A_4 of non-conformity with authority will make the society more homogeneous. Similar effects can be achieved if the action G promoted by authority significantly deviates from $\bar{\theta}$, which can be viewed as a 'natural' optimum behaviour 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 (Chua et al., 2019).

https://doi.org/10.1017/ehs.2021.40 Published online by Cambridge University Press

Cultural variation

Data show significant cultural variation in conformity (Bond and Smith, 1996), cognitive dissonance (Heine and Lehman, 1997; Hoshino-Browne et al., 2005) and certain aspects of the Theory of Mind (Lillard, 1998; Lecce and Hughes, 2010; Heyes and Frith, 2014). Collectivistic cultures put special emphasis on conformity. In my model, such cultures would be characterised by increasing costs of non-conformity A_3 and A_4 and increasing parameters β and γ 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 (α_1)and a reduced perception of logic constraints (α_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 θ 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 disapproval by peers A_2 and of non-conformity with peers' action A_3 are expected to increase with n. Therefore increasing population size is expected to make the culture tighter. This conclusion may appear to contradict the fact that urban areas show looser norms (see above). However, what this means is that the model predicts alternative pathways linking population size/density with norm tightness that have opposite effects. Which one is stronger is an empirical question.

Differences in the subsistence style

Societies may differ in the types of social interactions that their members are most often involved in. For example, coordination and reciprocal exchange of labour was very important in rice production which has contributed to tighter cultures in rice-producing regions of the world relative to wheat-producing regions (Talhelm and English, 2020). As discussed above, the higher the cost of miscoordination, the tighter the society is predicted to be. Subsistence style also affects the extent to which people rely on social learning (Glowacki and Molleman, 2017).

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

Possible generalisations

My conclusions have important caveats though. First, they concern the expected average behaviour 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 generalisation 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; Zhong et al., 2012) and other types of utility function (1) allowing for multiple equilibria (Michaeli and Spiro, 2017). Additional potential generalisations include multidimensional extensions of the model (Converse, 1964; Friedkin et al., 2016; Kashima et al., 2021), more realistic models of learning (e.g. Bayesian learning, Khalvati et al. 2019) and strategy revision, equity concerns and learning from others' performance. It would be interesting to use my models for studying political

https://doi.org/10.1017/ehs.2021.40 Published online by Cambridge University Press

polarisation (Lees and Cikara, 2021) as well as the processes through which people change their social identity (Green, 2020).

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 (d'Adda et al., 2020; Górges and Nosenzo, 2020; Gill and Rosokha, 2020; Andreozzi et al., 2020; Szekely et al., 2021). For example, d'Adda et al. (2020) 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 Szekely et al. (2021) did the same for a group of subjects playing a collective risk game (Milinski et al., 2008) over 28 rounds. Compliance with authority was studied in a Public Goods game (Silverman et al., 2014) and in the Joy of Destruction game (Karakostas and Zizzo, 2016). Importantly, because my main equations (e.g. equation 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 and conditional cooperators (Fehr and Schurtenberger, 2018; Fehr and Fischbacher, 2004). 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 (αs in my models), of learning from others (βs) and of complying with authority (γs). Existing surveys that correlate different characteristics of societies with the tightness–looseness of their norms (Gelfand et al., 2011; Harrington and Gelfand, 2014; Chua et al., 2019; Jackson et al., 2020) as well as studies of how values and social preferences change over time (Tormos, 2020; Kiley and Vaisey, 2020; Böhm et al., 2021) 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, L. Barrett and the reviewers for comments and suggestions.

Author contributions. SG designed and performed the research and wrote the paper.

Financial support. Supported by the US Army Research Office grants W911NF-14-1-0637 and W911NF-18-1-0138, the Office of Naval Research grant W911NF-17-1-0150, the Air Force Office of Scientific Research grant FA9550-21-1-0217, the National Institute for Mathematical and Biological Synthesis through NSF Award no. EF-0830858, and by the University of Tennessee, Knoxville.

Conflict of interest. None declared.

Research transparency and reproducibility. All data are in the manuscript. The Matlab code used is available upon request.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/ehs.2021.40.

References

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

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

Andreoni, J., Nikiforakis, N., & Siegenthaler, S. (2021). Predicting social tipping and norm change in controlled experiments. *Proceedings of the National Academy of Sciences USA*, *118*, e2014893118.

Andreozzi, L., Ploner, M., & Saral, A. S. (2020). The stability of conditional cooperation: Beliefs alone cannot explain the decline of cooperation in social dilemmas. *Scientific Reports*, *10*, 13610.
Apesteguia, J. and Maier-Rigaud, F. P. (2006). The tole of rivalry: Public goods versus common-pool resources. *Journal of Conflict Resolution*, *50*, 646–663.
Apperly, I. (2010). *Mindreaders: The cognitive basis of theory of mind*. Taylor & Francis.
Azar, O. (2004). What sustains social norms and how they evolve? The case of tipping. *Journal of Economic Behavior & Organization*, *54*, 49–64.
Basić, Z. and Verrina, E. (2020). Personal norms – and not only social norms – shape economic behavior. Technical report, Max Planck Institute for Research on Collective Goods.
Bicchieri, C. (2006). *The grammar of society. The nature and dynamics of social norms*. Cambridge University Press.
Bicchieri, C., Dimant, E., Gachter, S., & Nosenzo, D. (2020). Observability, social proximity, and the erosion of norm compliance. https://dx.doi.org/10.2139/ssrn.3355028.
Bisin, A. and Verdier, T. (2001). The economics of cultural transmission and the dynamics of preferences. *Journal of Economic Theory*, *97*, 298–319.
Böhm, R., Fleis, J., & Rybnicek, R. (2021). On the stability of social preferencesin inter-group conflict: A lab-in-the-field panel study. *Journal of Conflict Resolution*, *65*, 1215–1248.
Bond, R. and Smith, P. B. (1996). Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological Bulletin*, *119*, 111–137.
Calabuig, V., Olcina, G., & Panebianco, F. (2018). Culture and team production. *Journal of Economic Behavior and Organization*, *149*, 32–45.
Centola, D., Willer, R., & Macy, M. (2005). The emperor's dilemma: A computational model of self-enforcing norms. *American Journal of Sociology*, *110*, 1009–1040.
Cheung, M.-W. and Wu, J. (2018). On the probabilistic transmission of continuous cultural traits. *Journal of Economic Theory*, *174*, 300–323.
Chua, R. Y. J., Huang, K. G., & Jin, M. (2019). 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.
Cialdini, R. B. and Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Reviews in Psychology*, *55*, 591–621.
Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Personality and Social Psychology*, *58*, 1015–1026.
Converse, P. E. (1964). The nature of belief systems in mass publics. In D. E. Apter (Ed.), *Ideology and discontent* (pp. 206–261). Free Press.
d'Adda, G., Dufwenberg, M., Passarelli, F., & Tabellin, G. (2020). Social norms with private values: Theory and experiments. *Games and Economic Behavior*, *124*, 288–304.
Darwin, C. (1871). *The descent of man, and selection in relation to sex*. John Murray.
DeGroot, M. (1974). Reaching a consensus. *Journal of the American Statistical Association*, *69*, 118–121.
Diekmann, A. (1985). Volunteer's dilemma. *Journal of Conflict Resolution*, *29*, 605–610.
Eriksson, K., Strimling, P., & Coultas, J. C. (2015). Bidirectional associations between descriptive and injunctive norms. *Organizational Behavior and Human Decision Processes*, *129*(SI), 59–69.
Esteban, J. and Ray, D. (2001). Collective action and the group size paradox. *American Political Science Review*, *95*, 663–672.
Fehr, E. and Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*, 63–87.
Fehr, E. and Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, *2*, 458–468.
Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.
French, J. and Raven, B. (1959). The bases of social power. In D. Cartwright (Ed.), *Studies in social power* (pp. 150–167). Institute of Social Research.
Friedkin, N. E., Proskurnikov, A. V., Tempo, R., & Parsegov, S. E. (2016). Network science on belief system dynamics under logic constraints. *Science*, *354*, 321–326.
Fudenberg, D. and Tirole, J. (1992). *Game Theory*. The MIT Press.
Galesic, M. and Stein, D. L. (2019). Statistical physics models of belief dynamics: Theory and empirical tests. *Physica A: Statistical Mechanics and its Applications*, *519*, 275–294.
Gavrilets, S. (2015). Collective action problem in heterogeneous groups. *Proceedings of the Royal Society London B*, *370*, 20150016.
Gavrilets, S. (2020). The dynamics of injunctive social norms. *Evolutionary Human Sciences*, *2*, e60.
Gavrilets, S. and Richerson, P. J. (2017). Collective action and the evolution of social norm internalization. *Proceedings of the National Academy of Sciences USA*, *114*, 6068–6073.
Gavrilets, Y. N. (2003). Stochastic modeling of between-group social interactions. *Economics and Mathematical Methods*, *39*, 106–116 (in Russian).
Gelfand, M. J. *et al.* (2011). Differences between tight and loose cultures: A 33-nation study. *Science*, *332*, 1100–1104.

https://doi.org/10.1017/ehs.2021.40 Published online by Cambridge University Press

Gill, D. and Rosokha, Y. (2020). Beliefs, learning, and personality in the indefinitely repeated prisoner's dilemma. Technical report; http://dx.doi.org/10.2139/ssrn.3652318.
Glowacki, L. and Molleman, L. (2017). Subsistence styles shape human social learning strategies. *Nature Human Behavior*, *1*, 0098.
Golman, R., Bhatia, S., & Kane, P. B. (2020). The dual accumulator model of strategic deliberation and decision making. *Psychological Review*, *127*, 477–454.
Gomes, D. A. and Saúde, J. (2014). Mean field games models – A brief survey. *Dynamic Games and Applications*, *4*, 110–154.
Goodie, A. S., Doshi, P., & Young, D. L. (2012). Levels of theory-of-mind reasoning in competitive games. *Journal of Behavioral Decision Making*, *25*(1), 95–108.
Górges, L. and Nosenzo, D. (2020). Measuring social norms in economics: Why it is important and how it is done. *Analyse & Kritik*, *42*, 285–311.
Green, E. (2020). The politics of ethnic identity in Sub-Saharan Africa. *Comparative Political Studies*, *53*; https://doi.org/10.1177/0010414020970223.
Haney, C., Banks, C., & Zimbardo, P. (1973). Interpersonal dynamics in a simulated prison. *International Journal of Criminology and Penology*, *1*, 69–97.
Hardin, G. (1968). Tragedy of commons. *Science*, *162*(3859), 1243–1248.
Harrington, J. R. and Gelfand, M. J. (2014). Tightness-looseness across the 50 United States. *Proceedings of the National Academy of Sciences USA*, *111*, 7990–7995.
Hedden, T. and Zhang, J. (2002). What do you think I think you think? Strategic reasoning in matrix games. *Cognition*, *85*(1), 1–36.
Heine, S. J. and Lehman, D. R. (1997). Culture, dissonance, and self-affirmation. *Personality and Social Psychology Bulletin*, *23*, 389–400.
Henrich, J. (2015). *The Secret of Our Success*. Princeton University Press.
Heyes, C. M. and Frith, C. D. (2014). The cultural evolution of mind reading. *Science*, *344*, 1243091.
Hoshino-Browne, E., Zanna, A. S., Spencer, S. J., Zanna, M. P., & Kitayama, S. *et al.* (2005). On the cultural guises of cognitive dissonance: The case of Easterners and Westerners. *Personality and Social Psychology*, *89*, 294–310.
Jackson, J. C., Gelfand, M., & Ember, C. R. (2020). A global analysis of cultural tightness in non-industrial societies. *Proceedings of the Royal Society London B*, *287*, 20201036.
Jamali, M., Grannan, B. L., Fedorenko, E., Saxe, R., Báez-Mendoza, R., & Williams, Z. M. (2021). Single-neuronal predictions of others' beliefs in humans. *Nature*, *591*, 610–614.
Karakostas, A. and Zizzo, D. J. (2016). Compliance and the power of authority. *Journal of Economic Behavior & Organization*, *124*, 67–80.
Kashima, Y., Laham, S. M., Dix, J., Levis, B., Wong, D., & Wheeler, M. (2015). Social transmission of cultural practices and implicit attitudes. *Organisational Behavior and Human Decision Processes*, *127*, 113–125.
Kashima, Y., Perfors, A., Ferdinand, V., & Pattenden, E. (2021). Ideology, communication and polarization. *Philosophical Transactions of the Royal Society London B*, *376*, 20200133.
Khalvati, K., Park, S. A., Mirbagheri, S., Philippe, R., Sestito, M., & Rao, J.-C. D. R. P. N. (2019). Modeling other minds: Bayesian inference explains human choices in group decision-making. *Science Advances*, *5*, eaax8783.
Kiley, K. and Vaisey, S. (2020). Measuring stability and change in personal culture using panel data. *American Sociological Review, 85*, 477–506.
Kimbrough, E. O. and Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, *14*, 608–638.
Kimbrough, E. O. and Vostroknutov, A. (2018). A portable method of eliciting respect for social norms. *Economics Letters*, *168*, 147–150.
Kölle, F. and Quercia, S. (2021). The influence of empirical and normative expectations on cooperation. www.econtribute.de.
Konrad, K. (2009). *Strategy and dynamics in contests*. Oxford University Press.
Krueger, J. I. (2007). From social projection to social behaviour. *European Review of Social Psychology*, *18*, 1–35.
Kuran, T. and Sandholm, W. H. (2008). Cultural integration and its discontents. *Review of Economic Studies*, *75*(1), 201–228.
Lecce, S. and Hughes, C. (2010). 'The Italian job?': Comparing theory of mind performance in British and Italian children. *Journal of Developmental Psychology*, *28*, 747–776.
Lees, J. and Cikara, M. (2021). Understanding and combating misperceived polarization. *Philosophical Transactions of the Royal Society London B*, *376*, 20200143.
Lillard, A. (1998). Ethnopsychologies: Cultural variations in theories of mind. *Psychological Bulletin*, *123*, 3–32.
Lindstrom, B., Jangard, S., Selbing, I., & Olsson, A. (2018). The role of a 'common is moral' heuristic in the stability and change of moral norms. *Journal of Experimental Psychology – General*, *147*(2), 228–242.
Loewenstein, G. and Molnar, A. (2018). The renaissance of belief-based utility in economics. *Nature Human Behavior*, *2*, 166–167.
Marx, K. (1959). *A Contribution to the Critique of Political Economy*. Charles H Kerr.

https://doi.org/10.1017/ehs.2021.40 Published online by Cambridge University Press

McGinty, M. and Milam, G. (2013). Public goods provision by asymmetric agents: Experimental evidence. *Social Choice and Welfare*, *40*, 1159–1177.
Michaeli, M. and Spiro, D. (2017). From peer pressure to biased norms. *American Economic Journal: Microeconomics*, *9*, 152–216.
Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, *67*, 371–378.
Milinski, M., Sommerfeld, R. D., Krambeck, H.-J., Reed, F. A., & Marotzke, J. (2008). The collective-risk social dilemma and the prevention of simulated dangerous climate change. *Proceedings of the National Academy of Sciences USA*, *105*, 2291–2294.
Molander, P. (1992). The prevalence of free riding. *Journal of Conflict Resolution*, *36*, 756–771.
Murdock, G. P. and Provost, C. (1973). Measurement of cultural complexity. *Ethnology*, *12*, 379–392.
Olson, M. (1965). *Logic of collective action: Public goods and the theory of groups*. Harvard University Press.
Olsson, A., Knapska, E., & Lindström, B. (2020). The neural and computational systems of social learning. *Nature Review Neuroscience*, *21*, 197–212.
Ostrom, E. (2000). Collective action and the evolution of social norms. *The Journal of Economic Perspectives*, *14*, 137–158.
Pecorino, P. (2015). Olson's logic of collective action at fifty. *Public Choice*, *162*, 243–262.
Pelto, P. J. (1968). The differences between 'tight' and 'loose' societies. *Trans-action*, *5*, 37–40.
Perry, L., Duwal Shrestha, M., Vose, M. D., & Gavrilets, S. (2018). Collective action problem in heterogeneous groups with punishment and foresight. *Journal of Statistical Physics*, *172*, 293–312.
Perry, L. and Gavrilets, S. (2020). Foresight in a game of leadership. *Scientific Reports*, *10*, 2251.
Piotrowski, E. W. and Sladkowski, J. (2003). An invitation to quantum game theory. *International Journal of Theoretical Physics*,, *42*, 1089–1099.
Platt, J. (1973). Social traps. *American Psychologist*, *28*, 641–651.
Premack, D. and Wodruff, G. (1979). Does the chimpanzee have a theory of mind. *Behavioral and Brain Sciences*, *1*, 515–526.
Rabin, M. (1994). Cognitive dissonance and social change. *Journal of Economic Behavior and Organization*, *24*, 177–194.
Redner, S. (2019). Reality inspired voter models: A mini-review. *Comptes Rendus Physique*, *20*, 275–292.
Richerson, P. J. and Boyd, R. (2005). *Not by genes alone. How culture transformed human evolution*. University of Chicago Press.
Richerson, P. J., Gavrilets, S., & de Waal, F. B. M. (2021). Modern theories of human evolution foreshadowed by Darwin's *Descent of Man*. *Science*, *372*, eaba3776.
Roos, P., Gelfand, M., Nau, D., & Lun, J. (2015). Societal threat and cultural variation in the strength of social norms: An evolutionary basis. *Organizational Behavior and Human Decision Processes*, *129*, 14–23.
Sandholm, W. H. (2010). *Population games and evolutionary dynamics*. MIT Press.
Schelling, T. (1978). *Micromotives and macrobehavior*. Norton.
Silverman, D., Slemrod, J., & Uler, N. (2014). Distinguishing the role of authority 'in' and authority 'to'. *Journal of Public Economics*, *113*, 32–42.
Simon, H. A. (1957). *Models of man*. John Wiley.
Siopsis, G., Balu, R., & Solmeyer, N. (2018). Quantum prisoners' dilemma under enhanced interrogation. *Quantum Information Processing*, *18*, article number 144.
Song, G., Ma, Q., Wu, F., & Li, L. (2012). The psychological explanation of conformity. *Social Behavior and Personality*, *40*, 1365–1372.
Szekely, A., Lipari, F., Antonioni, A., Paolucci, M., Sánchez, A., Tummolini, L., & Andrighetto, G. (2021). Collective risks change social norms and promote cooperation: Evidence from a long-term experiment. *Nature Communications,* x, x.
Talhelm, T. and English, A. S. (2020). Historically rice-farming societies have tighter social norms in China and worldwide. *Proceedings of the National Academy of Sciences*, *117*, 19816–19824.
Tankard, M. E. and Paluck, E. L. (2016). Norm perception as a vehicle for social change. *Social Issues and Policy Review*, *10*, 181–211.
Tembine, H. (2017). Mean-field-type games. *AIMS Mathematics*, *2*, 706–735.
Thomas, W. I. (1928). *The child in America: Behavior problems and programs*. Alfred A. Knopf.
Tormos, R. (2020). *The rhythm of modernization. How values change over time*. Brill.
Troyer, L. and Younts, C. W. (1997). Whose expectations matter? the relative power of first- and second-order expectations in determining social influence. *American Journal of Sociology*, *103*, 692–732.
Tworek, C. M. and Cimpian, A. (2016). Why do people tend to infer 'ought' from 'is'? The role of biases in explanation. *Psychological Science*, *27*(8), 1109–1122.
Verhoeff, T. (1998). The trader's dilemma: A continuous version of the prisoner's dilemma. Technical report, Faculty of Mathematics and Computing Science, Technische Universiteit Eindhoven, The Netherlands.
Walker, J. M., Gardner, R., & Ostrom, E. (1990). Rent dissipation in a limited-access common-pool resource: Experimental evidence. *Journal of Environmental Economics and Management*, *19*, 203–211.
Watts, D. J. (2002). A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences USA*, *99*, 5766–5771.

https://doi.org/10.1017/ehs.2021.40 Published online by Cambridge University Press

Willard, A. K., Baimel, A., Turpin, H., Jong, J., & Whitehouse, H. (2020). 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.

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

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

Wood, W. (2000). Attitude change: Persuasion and social influence. *Annual Reviews in Psychology*, *51*, 539–570.

Zhong, W., Kokubo, S., & Tanimoto, J. (2012). How is the equilibrium of continuous strategy game different from that of discrete strategy game? *Biosystems*, *107*, 88–94.

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

Cite this article: Gavrilets S (2021). Coevolution of actions, personal norms and beliefs about others in social dilemmas. *Evolutionary Human Sciences* **3**, e44, 1–22. https://doi.org/10.1017/ehs.2021.40