PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Review

Cite this article: Gavrilets S, Tverskoi D, Sánchez A. 2024 Modelling social norms: an integration of the norm-utility approach with beliefs dynamics. *Phil. Trans. R. Soc. B* **379**: 20230027.
https://doi.org/10.1098/rstb.2023.0027

Received: 13 July 2023
Accepted: 9 September 2023

One contribution of 15 to a theme issue 'Social norm change: drivers and consequences'.

Subject Areas:
behaviour, theoretical biology

Keywords:
behaviour, beliefs, mathematical modelling, game theory, social evolution, cultural evolution

Author for correspondence:
Sergey Gavrilets
e-mail: gavrila@utk.edu

Modelling social norms: an integration of the norm-utility approach with beliefs dynamics

Sergey Gavrilets[1,2,3], Denis Tverskoi[2,3] and Angel Sánchez[4,5]

[1]Department of Ecology and Evolutionary Biology, [2]Department of Mathematics, and [3]Center for the Dynamics of Social Complexity, University of Tennessee, Knoxville, TN 37996, USA
[4]Grupo Interdisciplinar de Sistemas Complejos, Departamento de Matemáticas Universidad Carlos III de Madrid, Leganés, Madrid 28911, Spain
[5]Instituto de Biocomputación y Física de Sistemas Complejos, Universidad de Zaragoza, Zaragoza 50018, Spain

SG, 0000-0003-1581-4018; DT, 0000-0001-5324-4533; AS, 0000-0003-1874-2881

We review theoretical approaches for modelling the origin, persistence and change of social norms. The most comprehensive models describe the coevolution of behaviours, personal, descriptive and injunctive norms while considering influences of various authorities and accounting for cognitive processes and between-individual differences. Models show that social norms can improve individual and group well-being. Under some conditions though, deleterious norms can persist in the population through conformity, preference falsification and pluralistic ignorance. Polarization in behaviour and beliefs can be maintained, even when societal advantages of particular behaviours or belief systems over alternatives are clear. Attempts to change social norms can backfire through cognitive processes including cognitive dissonance and psychological reactance. Under some conditions social norms can change rapidly via tipping point dynamics. Norms can be highly susceptible to manipulation, and network structure influences their propagation. Future models should incorporate network structure more thoroughly, explicitly study online norms, consider cultural variations and be applied to real-world processes.

This article is part of the theme issue 'Social norm change: drivers and consequences'.

1. Background

In social sciences, most definitions of social norms involve beliefs about what others do and about what they should or should not do. The former are called descriptive norms [1], empirical expectations [2] or folkways (emerging out of routines, such as waiting in line). The latter are termed injunctive norms [1], normative expectations [2], mores (specifying what is moral or unethical), taboos (prohibition of behaviours so strict it results in disgust), prescriptive norms (encouraging positive behaviour), and proscriptive norms (discouraging negative behaviour) [3]. Such norms exist because of the collective belief in their existence, something akin to self-fulfilling prophecies [4]. Norms vary among families, cultural, ethnic or religious groups, regions and countries, and are influenced by exposure to different situations, leading to different degrees of adherence often described in terms of societal tightness–looseness [5–8]. Specifically, 'tight' cultures display strong norms, low tolerance for deviance, resistance to innovations and uniform social conduct, while 'loose' cultures demonstrate more relaxed norms, are more tolerant, and exhibit more diverse conducts. Importantly, people can incorrectly perceive others' beliefs, leading to pluralistic ignorance: people may believe their private thoughts and feelings differ from those of others when in fact they are not [9–11].

THE ROYAL SOCIETY PUBLISHING

While the two types of norms mentioned above focus on beliefs about others' actions and beliefs, personal norms (normative beliefs) describe what individuals believe they themselves should do. Personal norms can be shaped by an individual's moral values, often stemming from considerations about the welfare of others [2,12,13], or from their sense of what actions and beliefs are most appropriate [14,15]. These norms can also evolve from internalized social norms [16]. Here, we adopt a broad interpretation of personal norms, acknowledging that they can change over time. Independently of all these details, the ultimate factors explaining the origin, maintenance and diversity of norms are human susceptibility to social influence [17], pay-off differences between behaviours in different environments, and stochasticity involved in the appearance, spread and disappearance of behaviours in populations.

(a) Why people follow the norms

There are multiple reasons for people to follow social norms [2,18–22]. Social norms enable individuals to anticipate others' behaviours, thus leading to smoother social interactions. In uncertain situations, people infer latent norms via observation (when in Rome, do as the Romans do), a self-reinforcing process perpetuating these norms. Various factors such as mimicry, desire for approval and group identity contribute to norm adherence [23–25]. Individuals may also conform with others' perceived beliefs owing to perceptual and behavioural constraints or to avoid punishment of norm violators [2,19,21,26–29]. Norm internalization, where norms are adopted as personal beliefs and values, also enhances adherence [30–35]. Violation of these internalized norms can cause psychological discomfort, even when associated with material benefits [36]. Norm internalization can reduce costs related to information processing and decision-making [35], and help ensure cooperation [33,35]. While the inclination to follow norms is partly innate, specific norms are culturally influenced [37,38]. However, personal norms may be disregarded under conditions like high-compliance costs. Overall, following social norms is a multifaceted process influenced by individual cognition, group dynamics and broader societal factors.

(b) How norms change

New norms can emerge in younger generations, driven by a desire for a distinct social identity or competition for resources with older generations [39–44]. Changes in norms can also be triggered by fresh information about costs, benefits or others' behaviours and beliefs, and by alignment with authoritative or influential individuals. Normative beliefs can be recalibrated by correcting misperceptions about group behaviour and approval. Structural, ecological, historical, economic changes or specific policies that incentivize or regulate behaviours can impact norms and normative systems [45]. Education campaigns and communications by cultural or institutional actors can significantly influence norm changes [46–48]. Social norms can be changed by relatively small groups. As Margaret Mead said, 'never doubt that a small group of thoughtful, committed citizens can change the world,' reflecting the potential impact of trendsetters [46] and committed minorities [47,48] on norm evolution. Conversely, norms may persist despite environmental shifts, leading to a cultural mismatch [49].

Below, we discuss the forces and factors essential for modelling norm dynamics realistically. Next, we evaluate existing approaches based on how they incorporate these factors. We identify an emerging integrative approach optimal for modelling norm dynamics and review related work. We conclude by outlining general norm dynamics patterns identified by mathematical models.

2. Perspectives on modelling social norms

(a) Forces and factors to account for in models of norm dynamics

There are several crucial factors that must be accounted for in any realistic theory attempting to describe and predict norm dynamics.

(i) Decision-making and beliefs

Human beliefs are crucial in decision-making [50,51] as reflected in what is known in social psychology as the 'Thomas' theorem': 'If men define situations as real, they are real in their consequences' [52, pp. 571–572]. Regarding social norms, there are four elements of decision-making where human beliefs are very important. Firstly, empirical expectations influence our decision-making by affecting our anticipated material pay-offs. Secondly, our decisions are also influenced by the psychological well-being derived from conformity, where our behaviour is aligned with perceived norms leading to feelings of belonging and acceptance. Therefore, we are more likely to engage in a behaviour expected in our group even if it contradicts our personal beliefs or interests. Thirdly, normative expectations can significantly influence decision-making: we avoid actions we think will be disapproved to maintain our social standing. Conversely, we may behave in a way we believe will earn approval, even at personal cost. Lastly, personal norms impact our behaviour because they align with our values and self-image. Following them reinforces our self-concept as moral and good individuals, improving our psychological well-being.

(ii) Beliefs and attitudes changes

Changes in social norms occur simultaneously with changes in our beliefs about what others do, what others think and what is right or wrong in different situations [53–59]. Some changes occur gradually over generations, such as the norm regarding gender roles in many societies [60]. Other norms can change relatively quickly [61]. Norm change velocity can also be influenced by the level of consensus about a norm and the connectivity in a society or group [62]. Sometimes, the formation of our beliefs is not driven by conscious reasoning but by subconscious anticipation of their potential effects on others. These others can either reward or chastise us—occasionally promoting baseless beliefs, while sometimes penalizing justified ones [29,63]. Thus, integrating belief dynamics into theories of social norms is crucial.

(iii) Cognitive and psychological processes

Various psychological and cognitive processes influence decision-making and belief dynamics [64]. Cognitive

dissonance, the mental discomfort experienced when holding contradictory beliefs, values or attitudes, can be lessened through behaviour changes, new beliefs or selective memory [65–69]. Social projection, where individuals attribute their own thoughts and feelings to others [70–72], equips individuals with a 'theory of mind,' the ability to attribute mental states to oneself and others [73–76]. Another important process is psychological reactance, where individuals resist threats to their freedom, leading to oppositional behaviour or belief reinforcement [77–79]. Emotions also influence decision-making: fear promotes avoidance and conformity, and happiness drives behaviours with immediate rewards [80–84]. Emotions can both stem from and contribute to cognitive dissonance, and assist in understanding others' mental states.

(iv) Between-individual differences

Unique personality traits, cognitive styles, emotional reactions and social experiences can result in between-individual differences crucial in decision-making and belief dynamics. For instance, social identity theory shows that group identification can influence decisions [23,85], while cognitive dissonance theory highlights different strategies for resolving conflicting beliefs [65]. The theory of mind depends on diverse abilities to understand others' perspectives [73], and social projection underlines individual tendencies to assume shared beliefs, affecting interpretations of social norms [72]. Variations in conformity, anticonformity [86] and psychological reactance [77] affect behaviour in the context of social norms. Neglecting these differences leads to inaccurate predictions of behaviour and ineffective behaviour promotion. Cultural differences also play a significant role [87].

(b) Theoretical approaches for modelling social norms

There is a very large number of different theories of behavioural change [88] many of which have been studied using mathematical models. Here, we evaluate several approaches most fitting for modelling social norms dynamics in light of their ability to capture the factors discussed above.

(c) Classical and evolutionary game theory models

Classical non-cooperative game theory relies on utility maximization under perfect rationality of players [89]. By contrast, evolutionary game theory considers bounded rationality through processes like myopic best responses or imitation [90]. Social norms are often seen as equilibria in this context [39,45,91,92]. According to North [91, p. 821], a norm is 'an established and self-reinforcing pattern of behaviour: everyone wants to play their part given the expectation that everyone else will continue to play theirs. It is, in short, an equilibrium of a game.' Social norms emerge from interactions impacting individual pay-offs and are reinforced by reduced pay-offs for deviating behaviours, such as miscoordination costs or punishment by peers or institutions [39,90]. In evolutionary game theory, norms can undergo abrupt shifts (tipping) rather than gradual changes. Multiple equilibria are common, resulting in local populations conforming to different norms, maintaining global diversity [39]. Both classical and evolutionary game theories offer valuable frameworks for understanding how pay-off structures can influence behaviour across various scenarios. However, these theories often overlook normative considerations and psychological factors. Evolutionary game theory focuses instead on pre-programmed behavioural responses/strategies like cooperation, defection or punishment of defectors. Individuals in these models are typically assumed to either execute specific actions or imitate those with higher pay-offs, subject to occasional errors. While these approaches excel at capturing descriptive norms, they are less adept at addressing injunctive norms although some models that include punishment mechanisms for free-riders [32,93,94] can be seen as touching on injunctive norms as well.

(d) Psychological game theory models

Psychological game theory integrates beliefs, emotions and cognitive biases, improving our understanding of human behaviour in strategic situations by recognizing imperfect information and deviations from strict rationality [95–101]. Psychological game theory aims to account for the fact that what you believe others will do or think can actually make you happier or unhappier. For example, a player may experience guilt when he believes that the pay-off of his partner is lower than what the partner expected [101]. These beliefs can then influence the player's decision-making. Psychological game theory can indirectly model norms by incorporating psychological factors that can be influenced by them, capturing how social norms shape individual expectations about others and how guilt resulting from norm violations affects behaviour. However, existing models focus on anticipated behaviour rather than on normative expectations, making injunctive social norms challenging to model [101–103].

(e) Social influence models

Social influence mechanisms, such as imitation and conformity with peers, authority figures or high-status individuals can lead to convergence on shared behaviours, even without precise information about costs and benefits [104,105]. Convergence can lead to a consensus or to polarized states where multiple norms coexist within a population [48,104–108]. Norm transmission occurs through imitation and copying within the same generation or across generations. Recent research focuses on how social network structures impact norm dynamics [109–111]. Persistence, tipping, local convergence and global diversity, observed in evolutionary game theory, are present in social influence models. However, they often oversimplify by neglecting strategic behaviour and norm-adherence costs and benefits.

(f) Norm-utility models

Norm-utility models, a term not widely used (but see [103]), incorporate adherence to or deviation from social norms into individuals' utility functions [2,103,112–114]. Thus, people's decisions are not purely based on material considerations, but also on perceptions of what is appropriate or acceptable in a social group. These models usually represent social norms as rules or expectations about appropriate behaviour whose violation leads to a decrease in utility. Thus, behaviour that appears irrational in terms of material pay-offs becomes rational when the utility from norm adherence is considered. They are particularly useful

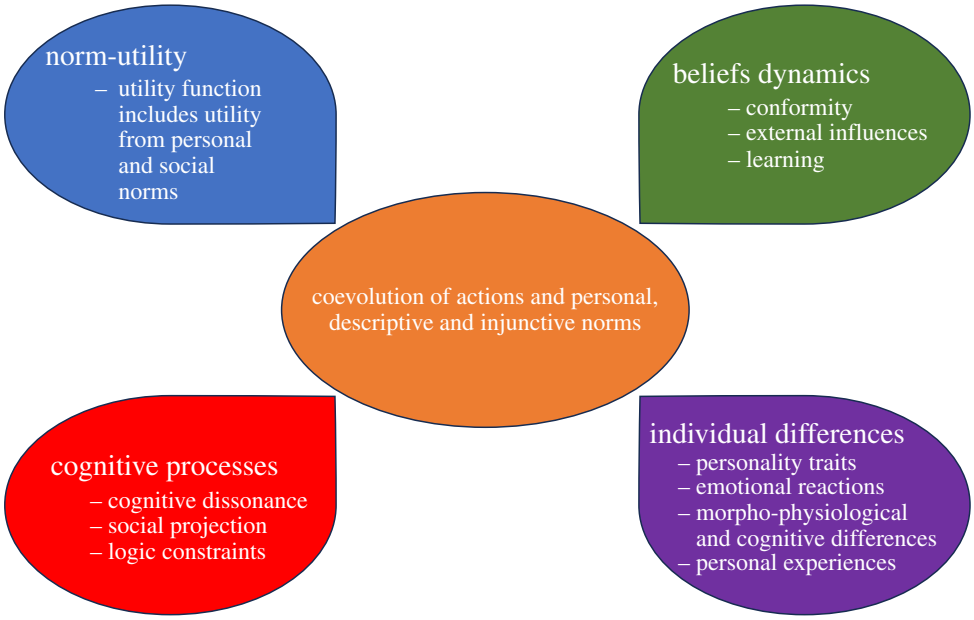

Figure 1. An integrative approach to modelling social norms. (Online version in colour.)

for analysing social dilemmas, cooperation and phenomena where social norms are relevant. However, norm-utility models usually do not consider the changing nature of personal norms or normative expectations.

(g) The role of beliefs and between-individual differences in models of norm dynamics

The above approaches differ in the role of individual beliefs. Evolutionary game theory embodies beliefs based on expectations about others' behaviour, and individuals adopt successful or expected-to-be-successful strategies without having explicit beliefs. In psychological game theory, individuals have beliefs about other players' strategies, intentions and mental states, influencing their decisions and responses. In social influence models, beliefs refer to individuals' opinions or attitudes influenced by others and are updated based on received information, leading to collective behaviours like consensus or polarization. In norm-utility models, individuals' beliefs about normative behaviours shape their utility from different actions, so changes in beliefs about norms can drive changes in norm adherence. Regarding between-individual differences, game-theoretic models often neglect them except for strategies, while social influence models ignore them except for opinions and positions in the social network. By contrast, between-individual differences are a crucial component of many norm-utility models.

Importantly, as we discussed above, beliefs coevolve with actions. Therefore, adequately modelling social norms requires considering jointly the dynamics of actions, attitudes (personal norms) and beliefs about others while accounting for cognitive processes and between-individual differences (figure 1). This can be achieved by integrating norm-utility approaches with social influence models, as we show in the next section.

3. Some specific models of norm dynamics

In this section, we look into the details of specific norm-utility models in some of which decision-making coevolves with norms. Along our discussion, we might occasionally simplify these models, omitting certain components for clarity. We will also modify and streamline notations for ease of understanding and comparison. Even if the original models did not explicitly centre on social norms, we aim to interpret their implications and conclusions in that context. In all models we consider, individuals choose the action maximizing the utility function. If belief dynamics are considered, they are usually described by simple linear equations that extend the classical DeGroot model of opinion change [107,115]. These extended models account for additional factors that influence individual beliefs (e.g. cognitive dissonance, social projection or authority's messaging), beyond just the opinions of peers. In a few cases, the changes in both actions and beliefs are found by a joint maximization of the utility function with respect to variables describing beliefs dynamics. We will organize the models based on the variables that undergo dynamic changes—whether these are merely actions or also personal norms and beliefs about others. We will discuss modelling assumptions about factors controlling decision-making (via utility function) and belief dynamics as well as main conclusions. For the purpose of our discussion, we will adopt the following notations: x for action (behaviour), y for personal norm (intrinsic preference or attitude), $\tilde{x}$ for first-order belief (empirical expectation), $\tilde{y}$ for second-order belief (normative expectation), $\bar{x}$ for (average) observed behaviour of peers, π for material pay-offs, u for the utility function. Coefficients A_0, A_1, A_2, A_3 and A_4 will capture the effects of material factors, personal, descriptive, injunctive norms and authorities, respectively, on individual decision-making. We will use this notation throughout the paper irrespective of the ones used in the original papers so all models can be meaningfully compared with each other.

The relationships between our main variables x, y, $\tilde{x}$, $\tilde{y}$ in social dilemmas have been extensively studied through behavioural experiments and heuristic regression models. For example, Fischbacher & Gächter [116] used multiple-round public goods experiments to study the effects of empirical expectations on actions and of observed behaviour on normative expectations (see also [117–119]). Bicchieri & Xiao [120] contrasted the effects of empirical and normative expectations in the dictator game. Other types of social interactions have

also been studied [114,121–124]. Decisions to cooperate strongly depend on whether others are expected to do so [28]. Informing subjects about both peers' actions and beliefs lead to synergistic effects [125]. Empirical and normative expectations can interact with personal norms [126–128]. The effort of authorities can change perceived norms [129]. People can strategically distort their beliefs, including those about norms, to justify self-serving behaviour [130–132]. All these findings highlight further the necessity of explicitly modelling the coevolution of beliefs and decision-making to understand behaviour in social dilemmas.

(a) Early norm-utility models in economics

Early norm-utility models had a significant impact on subsequent research in economics. The pioneering paper by Akerlof [112] modelled complex interactions between labourers and capitalists incorporating consumption, reputation and action-belief alignment into the utility function. Individuals adhered ($x = 1$) or not ($x = 0$) to a norm, believing ($y = 1$) or not ($y = 0$) in it. A reputational loss proportional to norm supporter frequency ($\bar{y}$) occurs upon norm-breaking. These assumptions lead to the utility function:

$$u(x) = \underbrace{A_0\pi(x)}_{\text{material pay-off}} - \underbrace{A_1(1-x)y}_{\text{cognitive dissonance}} - \underbrace{A_2(1-x)\bar{y}}_{\text{reputational loss}}, \tag{3.1}$$

where $\pi(x)$ is the material pay-off resulting from action x and constant parameters A_i measure the relative weights of the corresponding factors. Akerlof focused on norms that decrease individual pay-offs ($\pi(1) < \pi(0)$). The model predicts heterogeneity in both actions and beliefs with disadvantageous norms persisting because breaking them results in reputation loss. The model also predicts that individuals may adhere to norms even if they personally disagree with them (preference falsification [133]).

Follow-up papers applied this approach to several cases. A model on workplace safety beliefs [134] incorporates fear-induced mental costs into economic modelling, providing insights into the spread of innovations, advertisement influences, social security necessity and aspects of crime. A crime model [135] showed that cognitive dissonance can influence individuals to choose criminal activities under harsh penalties but dissuades them when penalties are lenient. Akerlof & Kranton [136,137] modelled situations where individuals optimize utility by selecting effort levels (x) and identities (y), looking at students who exert effort in academic pursuits and classify themselves into 'leading crowd,' 'nerds' and 'burnouts,' each with distinct behavioural norms.

Rabin [138] models the impact of cognitive dissonance on immoral behaviour, such as wearing fur, allowing continuous variation in actions (x) and moral beliefs (y). Individual pay-offs $\pi(x)$ increase with engagement level x, but excessive levels may be morally unacceptable. If $x > y$, cognitive dissonance induced a utility loss $d(x-y)$. Maintaining morally wrong beliefs (y) also led to a psychic cost ($c(y)$). With these assumptions we have:

$$u(x,y) = \underbrace{A_0\pi(x)}_{\text{material pay-off}} - \underbrace{A_1 d(x-y)}_{\text{cognitive dissonance}} - \underbrace{A_2\, c(y)}_{\text{cost of holding belief}}. \tag{3.2}$$

Maximizing the utility function u by considering simultaneously actions (x) and beliefs (y), Rabin showed that amplifying aversion to immorality (raising costs c) can paradoxically increase immoral behaviours owing to cognitive dissonance, where individuals attempt to rationalize immoral behaviours as morally acceptable. When individuals' beliefs influence one another, heightened immorality discomfort can unwittingly encourage collective rationalization of questionable activities, escalating their prevalence. If individuals are primarily influenced by observable behaviours of others rather than by expressed beliefs, increasing the perceived cost of immorality would lead to a decline in immoral activities.

Bernheim [139] modelled individuals who receive material benefits and utility from the prestige granted to them by others, with actions represented by a continuous variable x. Individuals differ in the type θ specifying the action that produces the highest material pay-off. Social interactions are implicit rather than explicit, with the assumption of a universally recognized most prestigious type set at $\theta = 1$. Then, we have:

$$u(x) = \underbrace{-A_0(x-\theta)^2}_{\text{material pay-off}} - \underbrace{A_2(1-x)^2}_{\text{prestige loss}}. \tag{3.3}$$

The model reveals that when societal status outweighs individual preferences (large A_2), many individuals conform to a uniform behavioural standard, disregarding their own preferences. However, groups with significant variation in individual preferences (θ) can resist conformity. The model clarifies why some activities follow behavioural standards while others do not, provides insights into norm evolution owing to preference shifts, and can explain both enduring customs and transient fads. Bénabou & Tirole [140] explored a similar model where the prestige of an action increased with its frequency. They identified conditions for two equilibrium states, each represented by the unanimous selection of one action or the other by all individuals involved. A later paper by the same authors [141] allowed for variation between individuals in intrinsic motivation y to perform a particular action. They used the model to explore the effects of norm-based interventions (such as making descriptive and injunctive norms more salient), aiming to increase the group's welfare.

(b) The Rashevsky model

Next we describe two classical models which initially were formulated without a consideration of utility function but nevertheless can be viewed as examples of the norm-utility approach. By contrast to the models discussed above, these two models are dynamic, directly capture conformity with peer behaviour, and explicitly account for the difference between individuals in characteristics controlling decision-making. The model developed by Rashevsky [106] was the very first attempt to model the effects of social influences on behaviour. (Nicolas Rashevsky is also viewed as the founder of mathematical biology [142] and cliodynamics [143].) Consider a population of N individuals who can take two actions: $x = 0$ and $x = 1$. The probability P of taking action 1 is monotonically increasing with the latent 'position' of the individual with regards to these two actions, written as a sum $y + z$, where y is a constant personal attitude which may depend on expected material or immaterial values associated with the actions. The term z is the net effect of social influence, assumed to be equal for all individuals.

Building on a model of neural discrimination between stimuli [144], Rashevsky [106] described the dynamics of social influence z by a differential equation:

$$\frac{dz}{dt} = \underbrace{\alpha N[2p(z)-1]}_{\text{effect of conformity}} - \underbrace{\beta z}_{\text{decay of social influence}}, \tag{3.4}$$

where $p(z)$ is the frequencies of behaviour 1 in the population. If behaviour 1 is more common (i.e. $p(z) > 0.5$), the first term describes an increase in z, otherwise it describes a decrease in z. The second term describes the decay of social influence to zero. Constant parameters α and β scale the corresponding rates of change in social influence. The model is completed by specifying the density function of the distribution $f(y)$ of personal attitudes y in the population and the function P converting $y+z$ into the probability of choosing action $x=1$. Given these two functions, $p(z) = \int P(y+z)f(y)\,dy$.

Rashevsky [106] demonstrated that z evolves towards an equilibrium, but also that there can be multiple equilibria, hence the final outcome may depend on initial conditions. The population can become 'stuck' in a state where a non-preferred behaviour or norm is maintained. Rashevsky's findings underscore the significance of heterogeneity in attitude y as characterized by function $f(y)$. Small parameter changes can induce tipping point dynamics and sudden shifts in population behaviour. Recent studies have used this model to examine interactions between identity groups and the effects of identity salience and propaganda on group behaviour [145–147].

(c) Granovetter-type models

The model formulated by Granovetter [48] is a generalization of models of spatial segregation developed by Schelling [148,149]. The beauty of Granovetter's formulation is in its simplicity. The model was introduced within the context of riots or social protests which each individual can join ($x=1$) or not ($x=0$) but we can also think about it in terms of other behaviours and norms. Each individual is characterized by a threshold d such that if the frequency $\bar{x}$ of others choosing action 1 is larger than d, the individual does the same. The actual value of d may depend on the perceived costs and benefits of possible actions, on personality, etc. The cumulative distribution of thresholds F in the population is assumed to be constant in time. Then if the current frequency of people choosing action 1 is $\bar{x}_t$, then for the proportion $F(\bar{x}_t)$ of the population $\bar{x}_t$ is larger than their thresholds, so they will choose action 1 as well. This immediately leads to a recurrence equation describing the dynamics of $\bar{x}$:

$$\bar{x}_{t+1} = F(\bar{x}_t).$$

It can be shown that, as time increases, $\bar{x}_t$ converges to an equilibrium. There can be several equilibria $\bar{x}^*$, which are given by solutions of the algebraic equation $\bar{x}^* = F(\bar{x}^*)$. The Granovetter model can also be formulated in continuous time [150].

To analyse the model in more detail we must specify the cumulative distribution function $F(d)$. Figure 2, shows the equilibria when the distribution of thresholds is truncated normal with mean $\bar{d}$ and variance σ^2. When σ is small while $\bar{d}$ is intermediate, there are two stable equilibria (close to $\bar{x}=0$ and $\bar{x}=1$) and an unstable equilibrium with intermediate $\bar{x}$. In this case, there is possibility for a tipping point dynamic when a small change in parameters can cause a dramatic change in the equilibrium frequency of behaviour. For example, in the left most figure increasing $\bar{d}$ beyond approximately 0.77 will cause $\bar{x}$ to drop from about 1 to about 0 while decreasing $\bar{d}$ beyond approximately 0.23 will cause $\bar{x}$ to increase from about 0 to about 1.

Yin [151] contrasted the cases where F is unimodal or bimodal, with equal or unequal peak values, to assess the effectiveness of interventions in promoting or suppressing mass protests. Efferson *et al.* [152] considered the effects of changing the distribution from a unimodal to bimodal (e.g. by educating a certain proportion of the population) to eliminate harmful social norms, such as female genital cutting.

Neither Rashevsky nor Granovetter had much to say about the nature of attitudes/thresholds y which were rather abstract in their models, but they can be linked to psychological factors and forces involved in decision-making. For example, in the model introduced by Kuran [153] individuals suffer moral integrity costs based on the discrepancy between their private attitude y and the action x taken, but receive reputational benefits proportional to the frequency of those exhibiting the same behaviour. Then the utility function becomes:

$$u(x) = -\underbrace{A_1|x-y|}_{\text{moral integrity cost}} - \underbrace{A_3|x-\bar{x}|}_{\text{reputational benefit}}. \tag{3.6}$$

In this model, the threshold value of $\bar{x}$ at which the utility of action ($x=1$) becomes larger than that of non-action ($x=0$) is $d = ((A_1+A_2)/2A_2) - (A_1/A_2)y$, so that a larger attitude y means a smaller threshold d. Now one can use equation (3.5) to describe the dynamics of the frequency $\bar{x}_t$ of people participating in mass protest. All conclusions from the original Granovetter model apply here.

Centola *et al.* [154] studied why people publicly enforce a norm that they secretly wish would go away. In their model, people can privately support ($y=1$) or oppose ($y=-1$) the norm, comply ($x=1$) or not ($x=-1$) with the norm, and punish ($z=1$) norm violators, punish norm-followers ($z=-1$) or do not punish anybody ($z=0$). Mean values $\bar{x}$ and $\bar{z}$ measure the extent of compliance and punishment in the population. At the first step of each round, each individual decides whether to comply or not with the norm by choosing an action x maximizing

$$u(x) = \underbrace{A_1xy}_{\text{cognitive dissonance}} + \underbrace{A_2x\bar{z}}_{\text{social pressure}}. \tag{3.7a}$$

The first term is maximized if action x and belief y match, while the second is maximized when the action complies with the prevailing punishment in the population (i.e. the sign of x matches that of $\bar{z}$).

At the second step, individuals who acted according to their beliefs at the first stage (i.e. those with $x=y$) can punish people with deviating behaviour if the need for enforcement (measured by $w=(1-y\bar{x})/2$) is sufficiently large. Those who acted against their beliefs because of social pressure, can follow with a 'false enforcement', that is, punish people whose behaviour they privately approve if social pressure is strong enough. These assumptions lead to two separate utility functions:

$$u(z|x=y) = -\underbrace{cyz}_{\text{cost of punishing}} + \underbrace{A_1wyz}_{\text{personal norm}}, \tag{3.7b}$$

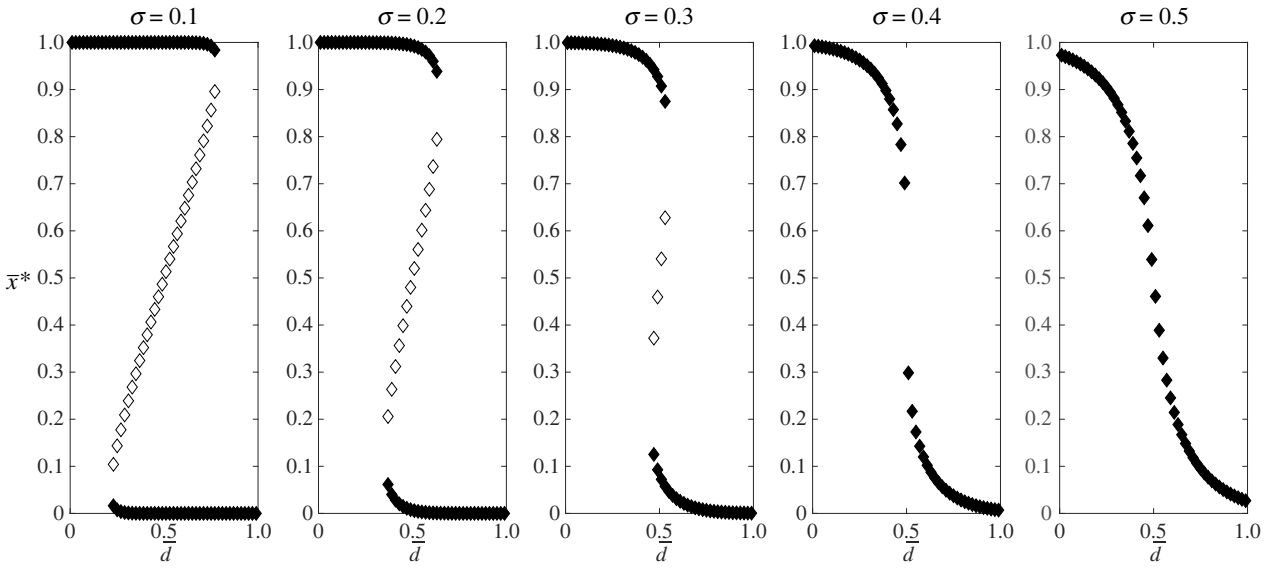

Figure 2. Equilibrium values of frequency $\bar{x}^*$ for the Granovetter model (3.5) assuming a truncated normal distribution of thresholds with mean $\bar{d}$ and variance σ^2. Stable equilibria are shown with solid circles; unstable equilibria are shown with open circles.

and

$$u(z|x=-y) = -\underbrace{(A_1+c)yz}_{\text{cost of punishing}} + \underbrace{\bar{z}z}_{\text{social pressure}}, \tag{3.7c}$$

where c measures the cost of punishing others. Strong conviction (larger A_1) promotes true enforcement (equation (3.7*b*)) and inhibits false enforcement (equation (3.7*c*)). Centola *et al.* [154] numerically studied this model on social networks, showing that when interactions between small neighbourhoods are limited, a small group can ignite cascades leading to almost universal norm adherence and enforcement. Converting false enforcers into true believers does not stabilize high-compliance equilibrium, but instead can trigger its collapse. Certain network features known for promoting the spread of information, innovations, rumours and diseases [155], hinder cascades of false enforcement.

Gavrilets [156] examined a model where individuals can adopt traditional or new behaviours. The traditional behaviour persists owing to its normative status, despite costs. Individuals gain approval or face disapproval based on behavioural alignment with others. Norm-followers have the option to punish norm-violators at a personal cost. The model's dynamics are defined by two Granovetter-type equations for the frequencies of norm-followers and punishers. The model shows that unpopular norms can persist owing to preference falsification, emphasizing the impact of parameters and initial conditions. Changes in the distribution of personal norms can significantly alter norm adherence frequency. Minor parameter adjustments can cause significant societal shifts, and behaviour modifications can be achieved by altering costs, normative values, societal expectations and strategic information dissemination. Gavrilets [156] discusses policy implications in abolishing norms such as footbinding and female genital cutting, reducing college students' drinking and promoting pro-environmental behaviours.

McCullen *et al.* [157] proposed a Granovetter-type model with thresholds being a weighted combination of behaviour frequencies across the entire system and the local neighbourhood. Their findings emphasize two crucial elements influencing the dynamics: the number of connections a node has with its neighbours, and the network's transitivity or clustering, that correlates with the neighbourhoods of interconnected individuals.

(d) Other models of the dynamics of descriptive norms

Norm-utility models in which personal norms/attitudes do not change, predict changes in the average behaviour that can be interpreted as a descriptive norm. In the model of Brock & Durlauf [158], a choice between two competing scientific theories ($x=0$) or ($x=1$) is influenced by existing evidence but also by social factors captured by the mean choice $\bar{x}$ in the population. The utility function is:

$$u(x) = \underbrace{A_0\pi(x)}_{\text{evidence-based utility}} - \underbrace{A_3(x-\bar{x})^2}_{\text{conformity}}. \tag{3.8}$$

The authors showed that social interactions can lead a community consensus away from that theory which is superior by scientific criteria (i.e. the one that has the highest value of π).

With two actions, norm following ($x=1$) and norm-breaking ($x=0$), López-Pérez [102] defined the utility function as:

$$u(x) = \underbrace{A_0\pi(x)}_{\text{material pay-off}} - \underbrace{A_3\bar{x}(1-x).}_{\text{cost of norm-breaking}} \tag{3.9}$$

He used his model to offer a norm-based explanation for why many subjects in experimental games cooperate contrary to their material interest, cooperate in a reciprocal manner, and are willing to punish those who behave unkindly.

Azar [159] modelled tipping. Let x be the tip in percentage of the bill and $\bar{x}$ the average tip in the previous period. The value of $\bar{x}$ is viewed as a descriptive norm. Then the utility function is:

$$u(x) = \underbrace{-cx}_{\text{material pay-off}} + \underbrace{yx}_{\text{moral satisfaction}} - \underbrace{A_3(x-\bar{x})^2,}_{\text{social disapproval}} \tag{3.10}$$

where c is the bill size and y the strength of internalization of the tipping norm. The term yx captures the positive feelings obtained from tipping. Azar shows that if there are consumers with $y > 0$ who get moral satisfaction from tipping,

tipping norms could stabilize (or even grow infinitely under specific extreme parameter conditions).

Azar [160] modelled workplace norms such as the refereeing time in an economics journal. Let y be the reviewer's personally preferred time given their personal characteristic, how busy they are, their interest in the paper, etc. The utility function is

$$u(x) = -\underbrace{A_1(x-y)^2}_{\text{cost of deviating from } y} - \underbrace{A_2(x-\bar{x})^2}_{\text{conformity with existing norm}}, \tag{3.11a}$$

where $\tilde{x}$ is the an existing (descriptive) norm. Azar postulated that the norm is given by a weighted average of the norm in the previous period and the average refereeing delay in the previous period:

$$\tilde{x}_t = \tilde{x}_{t-1} + \underbrace{\alpha(\bar{x}_{t-1} - \tilde{x}_{t-1})}_{\text{learning from observations}}, \tag{3.11b}$$

where α measures the weight of observations in the norm dynamics. Azar showed that the norm that gets established can be larger or smaller than the average preference $\bar{y}$ of individuals depending on the heterogeneity in the population. te Velde [161] modelled the effects of social image motivations on decision-making when the population is divided as to what is right. There are two possible meanings of social image: people may signal their adherence to their personal norm, or they may wish for others to approve their choices. Individuals differ in actions x and personal norms y and the utility function is:

$$u(x) = \underbrace{A_0\pi(x)}_{\text{material pay-off}} - \underbrace{A_1(x-y)^2}_{\text{cognitive dissonance}} + \underbrace{A_2F(x,\bar{y})}_{\text{social image utility}}, \tag{3.12}$$

where the social image term F depends on the action chosen x and the distribution of types in the population. te Velde shows how distinct motives for maintaining social image lead to different outcomes in terms of consensus, hypocrisy, compromise, polarization and destructive posturing. Besides, using social incentives to change behaviour may easily backfire if heterogeneous norms, or approval and respect, are conflated. Earlier Brekke *et al.* [162] studied a similar model but without the last term in equation (3.12).

Houle *et al.* [163] studied cooperation and conflict in a society with multiple factions engaged in economic and political interactions. The model considers two interrelated games: an 'economic game', in which agents of identity-based factions and with different political power can cooperate ($x = 1$) or not ($x = 0$) in the production of a resource, and a 'political game', in which individuals devote a fixed proportion of their resources to a competition the results of which establish the rules of the economic game. The utility function is:

$$u(x) = -\underbrace{A_0\pi(x)}_{\text{material benefit}} + \underbrace{A_3(2\bar{x}-1)x}_{\text{conformity with peers}} + \underbrace{A_4x_sx}_{\text{conformity with state}}, \tag{3.13}$$

where x_s is the action of the most powerful faction (the state). Houle *et al.* showed that high conformity with the state (large A_1) will stabilize cooperation, while high conformity with peers (large A_2), can counter-intuitively, destabilize cooperation, because once a majority of low-power factions are defecting, the other factions are 'pulled' to defect as well. Houle *et al.* tested various modelling predictions using social unrest as a proxy for the breakdown of cooperation in society and data covering 75 countries worldwide between 1991 and 2016.

Yang *et al.* [164] used a game-theoretic model to explore the socio-cultural factors influencing mask-wearing during the COVID-19 pandemic. The utility of mask-wearing depended on perceived infection risk, strength of the descriptive social norm, institutional signals promoting mask-wearing, and individual sensitivity to these signals. The mask-wearing benefit correlated with the susceptible-exposed-infectious-recovered-susceptible infection model's frequency of infected people. They found that increased pathogen spread or stricter policies could trigger a behavioural cascade, leading to full mask adoption. While cultural tightness can slow initial adoption (because people are more reluctant to modify their behaviour), it accelerates adoption once a tipping point is reached, helping establish mask-wearing as a norm. The tighter the culture, the more likely it is that collective mask-wearing will continue, even when the risk of infection decreases and policies are relaxed.

(e) Dynamics of descriptive and personal norms

An important limitation of most models considered above is that they assume attitudes y remain constant. Next we discuss models explicitly accounting for the dynamics of attitudes.

Kuran & Sandholm [165] introduced a model of 'cultural integration', in which individuals have personal norms y, potentially related to their social identity, regarding behavioural acts x. However, they also benefit from coordinating their actions with others. We can capture these assumptions by an utility function:

$$u(x) = -\underbrace{A_0(x-\bar{x})^2}_{\text{material pay-off}} - \underbrace{A_1(x-y)^2}_{\text{cognitive dissonance}}. \tag{3.14a}$$

With constant personal norms, Kuran & Sandholm [165] show that the equilibrium behaviours of individuals reflect compromises between their own preferences and the need to coordinate with others. Kuran & Sandholm [165] also studied the case when personal norms change, by adapting the DeGroot model [107] of opinion change:

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \underbrace{\alpha(x_i^* - y)}_{\text{cognitive dissonance}}. \tag{3.14b}$$

That is, each agent's personal preference changes over time towards his current 'action' to reduce cognitive dissonance. In this case, preferences (y) and actions (x) converge to the initial mean preference $\bar{y}_0$. This convergence can be interpreted as the emergence of a single 'melting pot' scenario. Kuran & Sandholm [165] have extended the model to two partially segregated communities, where members have limited interactions with members from the other community. Their analysis focused on the extent of cultural segregation and the efficiency of policies aimed at preserving cultural distinctness or promoting cultural integration. Della Lena & Dindo [166] study different generalizations of the Kuran and Sandholm model.

Martins [167] considered a model in which individuals have a discrete set of alternative actions. Individual attitude/preferemce is specified by a probability distribution defined over this set. Each individual chooses the action with highest value, which means the utility function coincides with the personal norm. After choosing an action and

observing groupmates' behaviour, individuals update their personal norm using the Bayes rule. Numerical simulations on a network demonstrated the emergence of extreme personal norms, where individuals believe that one alternative is significantly superior to all others. Clusters consisting of individuals with similar attitudes arose, with central nodes in these clusters representing individuals with extreme personal norms.

Acharya *et al.* [168] considered strategic interactions between two agents. Utility function accounted for cognitive dissonance, conformity and a loss of utility owing to the deviation of the current personal norm from its initial value. They showed that at the equilibrium, personal norms match actions and that stronger conformity leads to large deviations from initial personal norms. Their results highlight that interactions between individuals expressing diverse perspectives can facilitate empathetic changes in actions.

Calabuig *et al.* [169–171] studied the coevolution of actions and personal norms in a linear public goods game with quadratic costs in heterogeneous groups. Both actions (x) and personal norms (y) are continuous variables. Individuals differ in the efficiencies of their efforts s and the shares v of the reward they secure from the good produced. These differences lead to differences in the efforts $\theta = vs$ maximizing individual material pay-off. Individuals are motivated by material pay-offs but also prefer to follow their personal norms y. The utility function is:

$$u(x) = \underbrace{A_0 \pi}_{\text{material pay-off}} - \underbrace{A_1 (x-y)^2}_{\text{cognitive dissonance}}. \tag{3.15a}$$

After choosing an action maximizing utility and observing groupmates' choices, individuals update their personal norms, driven by cognitive dissonance and conformity with groupmates. It is described by a DeGroot-type recurrence equation analogous to equation (3.14*b*) above:

$$y' = y + \underbrace{\alpha(x-y)}_{\text{cognitive dissonance}} + \underbrace{\beta(\bar{x}-y)}_{\text{conformity}}, \tag{3.15b}$$

where $\bar{x}$ is the average action and α and β measure the weight of the corresponding factors. The model allows for variation in all parameters.

Calabuig *et al.* [169] show that the population evolves to an equilibrium with the average action $\bar{x}$ and the average personal norm $\bar{y}$ matching $\bar{\theta}$. At the same time, individuals deviate from the values θ maximizing their pay-offs. At equilibrium, the observed variances satisfy the inequalities: $\mathrm{var}(y) \leq \mathrm{var}(x) \leq \mathrm{var}(\theta)$. Their results predict that (cultural) variation in personal norms and behaviour increases with the variance of skills ($\mathrm{var}(s)$), the average group skill level $\bar{s}$ and the variance of the income sharing rule ($\mathrm{var}(v)$). Increasing conformity (i.e. larger β) decreases this variation while increasing cognitive dissonance (i.e. larger α) or the weight of material factors (larger A_0) have opposite effects. Figure 3 illustrates that ignoring the fact that personal norms can change can lead to very different predictions about the equilibrium distributions of actions and beliefs.

Calabuig *et al.* [171] used the above model to study the effects of culture on group productivity. They demonstrated that individualism increases the equilibrium efforts of individuals with above-average revenue and decreases them for those with lower revenues. Conversely, collectivism raises the equilibrium effort of individuals with below-average revenue and reduces it for high earners. In teams with diverse skills, individualism can affect both team revenue and costs depending on specific team parameters. In homogeneous teams, individualism only increases costs, but with unequal revenue sharing, full collectivism maximizes team production. The optimal balance between individualism and collectivism depends on the team's income distribution and skill diversity. Lastly, the team's culture can either amplify or mitigate the changes in skill or income distribution within the team.

Building on earlier work [172], Zino *et al.* [173] considered two possible actions ($x = \pm 1$), with the attitude y which can take any value within the range [–1, 1]. Individuals update their actions and opinions after interacting on a two-layer network. The utility function includes the terms for cognitive dissonance and for material pay-offs from coordination with neighbours in the so-called 'influence layer'. Attitudes are updated according to a DeGroot-type model weighting communications and observations on the 'communication layer'. The model exhibits a range of dynamics: rapid shifts to new sets of beliefs, where the majority adopts an innovation, or development and maintenance of an unpopular norm, where despite overwhelming support for an innovation, individuals fail to embrace it. Under some conditions the community favours the status quo over any innovation.

Mo & Sun [174] extend the above model by introducing an 'opinion regulator', an agent who can communicate with some nodes/individuals affecting their opinions and 'impulsive stimulation', a periodic reward or punishment for a specific behaviour administered to some individuals to promote or inhibit choosing this behaviour. Mo & Sun discuss optimal strategies of opinion regulating and impulsive stimulation for shifting behaviours in the population.

Aghbolagh *et al.* [175] used a similar model but with additional utility function components describing individual prejudices (unchangeable personal norms) and an external influence source. They identify the conditions necessary for the emergence and stability of polarized equilibria, in which the population divides into two factions endorsing and pursuing different courses of action. They also study conditions for pluralistic ignorance, when a social group mistakenly infers the opinions of others based on observed actions.

(f) Dynamics of actions and descriptive, personal and injunctive norms

We are aware of only one paper jointly modelling the dynamics of normative ($\tilde{y}$) and empirical ($\tilde{x}$) expectations in addition to actions (x) and personal (y) norms [176]. Inspired by recent behavioural experiments [119,123,126–128,132], Gavrilets [176] described quantitatively the dynamics of these variables in social dilemmas. Besides social influences by peers, Gavrilets' model also accounted for the influence by an external authority promoting a particular action G. Each individual chooses an action x to maximize the subjective utility function

$$\left.\begin{aligned} u(x) = {} & \underbrace{A_0\, \pi(x,\tilde{x})}_{\text{material pay-off}} - \underbrace{\frac{1}{2} A_1 (x-y)^2}_{\text{cognitive dissonance}} - \underbrace{\frac{1}{2} A_2 (x-\tilde{y})^2}_{\text{disapproval by peers}} \\ & - \underbrace{\frac{1}{2} A_3 (x-\tilde{x})^2}_{\text{conformity w/peers}} - \underbrace{\frac{1}{2} A_4 (x-G)^2}_{\text{compliance w/authority}}. \end{aligned}\right\} \tag{3.16a}$$

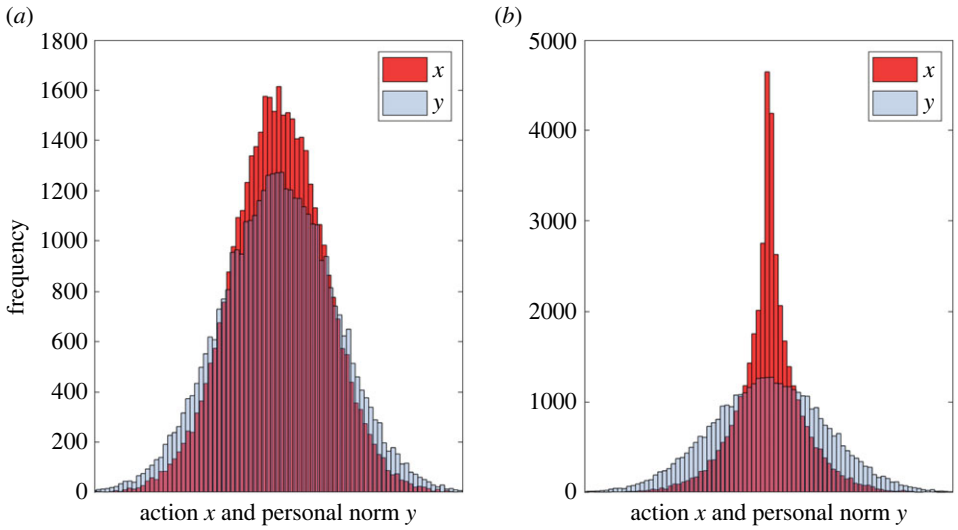

Figure 3. An example of equilibrium distributions of actions and personal norms in the model of Calabuig *et al.* [171] without (*a*) and with (*b*) evolution in personal norms. All averages are close to $\bar{\theta}$ but the variance in y is larger than that in x on the left but smaller on the right. The distribution of θ and the initial distribution of y are normal with same mean and standard deviations. The distributions of A_i, α_i, β_i are uniform on [0, 1]. The population size is 40 000. (Online version in colour.)

After taking actions and observing groupmates' behaviour, the attitude and beliefs of the individual change as described by the linear deGroot-type recurrence equations:

$$y' = y + \underbrace{\alpha_1(x-y)}_{\text{cognitive dissonance}} + \underbrace{\beta_1(X-y)}_{\text{conformity w/peers}} + \underbrace{\gamma_1(G-y)}_{\text{compliance w/authority}}, \tag{3.16b}$$

$$\tilde{y}' = \tilde{y} + \underbrace{\alpha_2(y-\tilde{y})}_{\text{social projection}} + \underbrace{\beta_2(X-\tilde{y})}_{\text{learning about others}} + \underbrace{\gamma_2(G-\tilde{y})}_{\text{compliance w/authority}}, \tag{3.16c}$$

$$\text{and} \quad \tilde{x}' = \tilde{x} + \underbrace{\alpha_3(\tilde{y}-\tilde{x})}_{\text{logic constraints}} + \underbrace{\beta_3(X-\tilde{x})}_{\text{learning about others}} + \underbrace{\gamma_3(G-\tilde{x})}_{\text{compliance w/authority}}, \tag{3.16d}$$

where prime indicates the next time step, X is the average action of groupmates observed by the focal individual (so different individuals can have different X), and α_i, β_i, γ_i are non-negative constant coefficients measuring the strength of the corresponding forces.

Gavrilets [176] examined social interactions characterized by quadratic pay-off functions. With no messaging and in the absence of 'stubborn' individuals who refuse to change, the population progresses towards a state where the average behaviour aligns with behaviour maximizing individual material pay-offs, in agreement with standard game-theoretic models. On average, individuals develop attitudes and beliefs justifying (or matching) their behaviours. In equilibrium, substantial inter-individual variability exists in all variables, reflecting individual psychological traits. With messaging by an external authority, long-term equilibrium encapsulates a balance of diverse forces, often deviating from game-theoretic predictions. Attempts by an external authority to direct group behaviour can trigger an opposing behaviour (backfiring effect). Gavrilets [176] also studied how various factors can affect differences in tightness/looseness of social norms between groups and societies, highlighting societal heterogeneity, societal threats, authority effects, cultural variations in collectivism versus individualism, population size and subsistence style as significant factors. Tverskoi *et al.* [177] tested this model using data from a long-term common pool resources experiment without and with messaging promoting group beneficial actions. Figure 4 shows that the match between model-based predictions and observed data is good.

Tverskoi *et al.* [178] adopted the model developed in [176] to investigate new technology diffusion using a model encompassing individual proficiency in the technology, shifts in attitudes (y) and adoption decisions (x). The model predicts that early adopters exhibit low dissonance and peer conformity, but are swayed by authority. Also, individualistic societies fare better in early technology adoption, societies with strong normative factors and conformity with authorities promoting new technology achieve high adoption rates, those with high cognitive dissonance resist new technologies, and future-oriented cultures embrace innovations. The dynamic nature of personal norms is crucial for these predictions.

Gavrilets & Richerson [179] simplified the model in [176] to analyse evolution of food sharing in small-scale societies, involvement in political protests, and the impact of priming social identity in behavioural experiments. For each application, their approach provides different (or simpler) explanations of human behaviour compared to other methods. Moreover, they precisely determined and characterized the extent of discrepancy between individual actions and attitudes.

(g) Evolutionary emergence of norm-utility

The various norm-utility models discussed above presume specific non-material utility function components. The emergence of norm-utility has been explored in several studies. Alger & Weibull [180] demonstrated that assortative interaction based on chosen strategies can lead to the evolution of individual utility functions, turning socially optimal behaviour into personal norms. Gavrilets & Richerson [181]

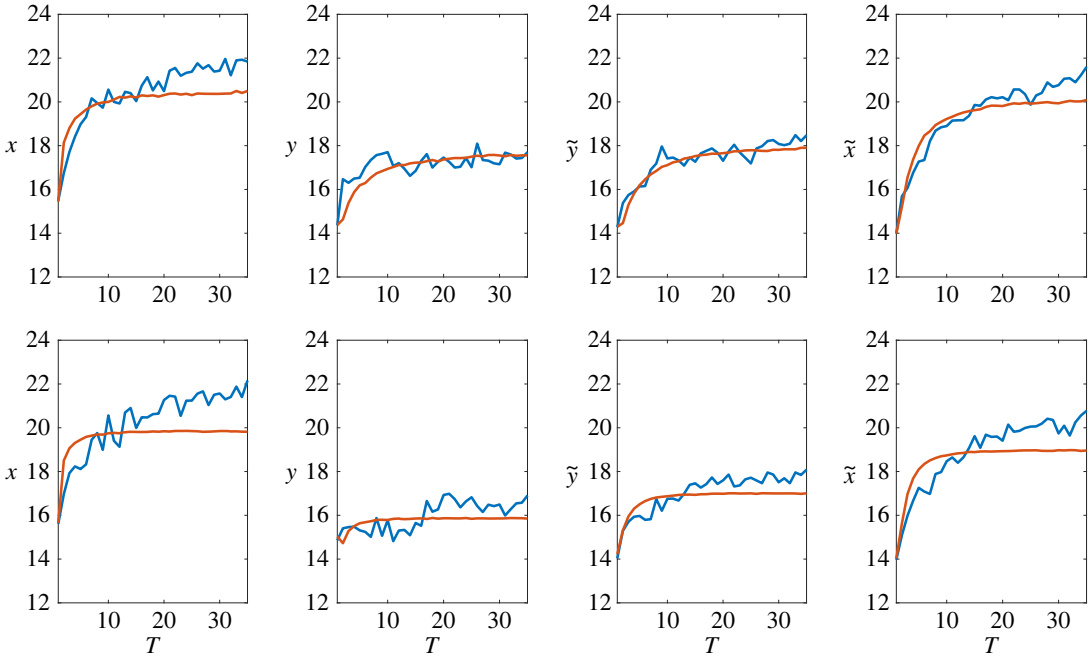

Figure 4. Observed (blue) and simulated (red) mean trajectories. The 'simulated trajectories' were obtained by repeatedly iterating dynamic equations describing the model using the obtained estimates of parameters and the actual individual data in the first round. Shown are at top: the results with no messaging; shown are at bottom: the results with messaging (after fig. S6 in [177].) (Online version in colour.)

explored the genetic evolution of the capacity to internalize social norms within populations undertaking collective actions. This model was expanded by Lozano *et al*. [182] for competitive within-group dynamics. Akcay & van Cleve [183] showed that populations engaging in social interactions could evolve to internalize the necessity to conform to majority behaviour. Kimbrough and co-workers [184,185] studied the origin of personal norms and normative expectations, accounting for individual differences in consumption utility. They found that injunctive norms could arise from minimizing overall consumption-related dissatisfaction as agents interact. If consumption utilities are unknown, personal norms could emerge from minimizing perceived dissatisfaction based on beliefs about others' consumption utilities. Normative expectations emerge as individuals' perceptions of others' personal norms based on current information.

4. Discussion

By integrating norm-utility approaches with belief dynamics, recognizing cognitive forces, accounting for individual differences and considering the role of authority influences, we can effectively and flexibly model the emergence, persistence and evolution of social norms. Such models allow for a rich, multifaceted exploration of the complex coevolution of norms and beliefs over time and across different spheres of human life.

Several general patterns arise from the models discussed above. First, certain behavioural patterns can persist within populations for a long time. These could be some advantageous behaviours, like cooperation [102,103,163,176,179], but also behaviours detrimental for individuals' material well-being or privately disapproved of [112,139,154–156]. Mechanisms contributing to the stability of such norms include preference falsification (publicly expressing preferences disagreeing with their true private ones [133]), pluralistic ignorance (mistakenly believing that one's private beliefs are in the minority even if they are widely shared [9–11]), false enforcement (enforcing a norm privately disapproved of, [154]) or the 'spirals of silence' (hesitating to voice dissenting opinions or divergent behaviour [186]).

Psychological and cognitive processes play crucial roles in maintaining and transforming social norms. Among these processes, cognitive dissonance (and its consequences for behaviour and beliefs), having being widely modelled [112,135,138,154,161,165,168–171,173,175,176,178], stands out as a significant factor that can give rise to backfiring effects. Models show that imposing stricter penalties may surprisingly lead to an increase rather than a decrease in criminal behaviour [135]. Similarly, a heightened public shaming and disapproval of amoral conduct can unexpectedly contribute to an upsurge of such behaviour [138]. People's reactions to messaging and nudging may steer them in the opposite direction of the intended one, and variations in social projection and cognitive constraints on beliefs can result in diverse dynamics of actions and preferences [176].

Models focusing on descriptive norms assume that people correctly identify them from observations, i.e. beliefs are correct [106,158,160,163,169–171,178]. When interactions happen on social networks, people have information only about the average behaviour among their social partners [175]. Only a few papers considered that people's empirical expectations can differ from observed behaviours, and even less models look at the dynamics of injunctive norms [176]. Nevertheless, the models show that incorrect perception of norms will strongly affect group behaviour and belief dynamics. For example, one consequence are self-fulfilling prophecies—predictions that, by being made, directly or indirectly make themselves true [4]: if it is collectively believed that some behaviour is the norm, individuals are likely to conform to that behaviour, thereby making the prediction true [156,176].

Mathematical models depict dynamics leading to tipping points, where infrequent behaviour suddenly becomes

widespread. This can happen after a significant shift in external circumstances (e.g. environmental or political) making a different behaviour more advantageous compared to previous practises. Alternatively, there may be a mass realization that long-held beliefs about personal circumstances, identity, or perceptions of others are flawed. More intriguingly, situations leading to tipping points can arise from minute changes. In mathematical models, this requires the existence of multiple equilibria such as those shown in figure 2. Alternative behaviours yielding higher pay-offs, strong conformity or mismatches between high pay-off strategies and authority-promoted norms promote multiple stable states. The exact conditions for tipping point dynamics largely depend on model specifics, parameters and belief distributions within the population.

Models predict that polarization in behaviour and beliefs can be sustained by differing behaviour pay-offs, allowing disparate belief systems to remain stable in the population despite varied societal advantages [112,161,162]. These models highlight how norms and beliefs are susceptible to manipulation by those with specific agendas. Models also suggest that norms and beliefs are highly susceptible to manipulation. Individuals or groups with particular agendas may exploit this vulnerability, significantly altering shared norms and collective beliefs. Individual and cultural differences can greatly impact social change dynamics and outcomes [164–166,169,171,176,178]. Additionally, the structure of social networks, including individual connections and information flow, significantly influences the spread of new behaviours and beliefs. Models also stress the importance of initial conditions, particularly the location of behaviour emergence, with some suggesting that innovations arising on a network's periphery have a higher success rate.

Mathematical models of social norms dynamics provide an invaluable foundation for understanding how norms develop and evolve. Extending these approaches is crucial to more accurately reflect key factors shaping our societies. While cooperation and coordination have been successfully modelled using norm-utility approaches, other types of norms may require different methods. For example, the signalling norm [187] is described by a sequential game for which the norm-utility approach would not be practical to apply. Punishment of norm violators is pivotal for both the establishment and preservation of social norms [2,19,21,26–29]. Despite its importance, there has been relatively scant effort to integrate punishment mechanisms into norm-utility models [154,156,181]. Much of the modelling work on punishment has been conducted within the framework of evolutionary game theory, where individuals are generally pre-programmed to either penalize defectors or emulate those with the highest pay-offs [32,93,94]. Enhancing norm-utility models to more comprehensively include punishment mechanisms would substantially elevate both their realism and applicability. We also need to better incorporate network structure, the intricate web of relationships that influence the propagation of norms [62,188–190]. We need to address the emergence of social norms online, and the rise of new cultural authorities in digital spaces [191]. Also, we should account for the evolutionary emergence of differences in the parameters of utility functions and belief dynamics, considering how different cultural contexts shape individual and collective values, preferences and beliefs, and how these differences play out in social norm dynamics. Beyond network structure, intra- and inter-individual forces and culture, attention should be paid to how groups, their identities and between-group relationships are formed and change. In particular, the feedback loops between identity-linked social norms and forces changing the group boundaries may also be very important. Finally, we need more detailed and realistic models linked to tangible real-world processes, such as ecological and environmental shifts, economic fluctuations, or epidemiological trends. This would enrich our understanding of how rewards and penalties associated with different behaviours can shape the formation, persistence and change of social norms. Such enhanced models of social norm dynamics, if properly validated, parameterized and tested (e.g. [177]), could more accurately reflect the nuanced and complex reality of human social behaviour and be applied for mitigating various challenges faced by our society [192].

Ethics. This work did not require ethical approval from a human subject or animal welfare committee.

Data accessibility. This article does not contain any additional data.

Declaration of AI use. Yes, we have used AI-assisted technologies in creating this article. AI was used to shorten certain pieces of text.

Authors' contributions. S.G.: conceptualization, investigation, project administration, writing—original draft, writing—review and editing; D.T.: investigation, writing—review and editing; A.S.: investigation, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. A.S. was supported by grant nos. PGC2018-098186-B-I00 (BASIC) funded by MCIN/AEI/10.13039/501100011033 and by 'ERDF A way of making Europe'. S.G. was supported by the Office of Naval Research (grant no. W911NF-17-1-0150), the Air Force Office of Scientific Research (grant nos. FA9550-21-1-0217 and FA9550-22-1-025) and the John Templeton Foundation.

Acknowledgements. We thank D. Garber, M. Lipatov, W. Przepiorka and a reviewer for comments and suggestions.

References

1. Cialdini RL, Reno RR, Kallgren CA. 1990 A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Pers. Soc. Psychol.* **58**, 1015–1026. (doi:10.1037/0022-3514.58.6.1015)
2. Bicchieri C. 2006 *The grammar of society. The nature and dynamics of social norms*. Cambridge, UK: Cambridge University Press.
3. Horne C, Mollborn S. 2020 Norms: an integrated framework. *Ann. Rev. Sociol.* **46**, 467–487. (doi:10.1146/annurev-soc-121919-054658)
4. Merton RK. 1948 The self fulfilling prophecy. *Antioch Rev.* **8**, 173–190. (doi:10.2307/4609267)
5. Pelto PJ. 1968 The differences between 'tight' and 'loose' societies. *Transaction* **5**, 37–40. (doi:10.1007/BF03180447)
6. Gelfand MJ *et al*. 2011 Differences between tight and loose cultures: a 33-nation study. *Science* **332**, 1100–1104. (doi:10.1126/science.1197754)
7. Harrington JR, Gelfand MJ. 2014 Tightness-looseness across the 50 united states.. *Proc. Natl Acad. Sci. USA* **111**, 7990–7995. (doi:10.1073/pnas.1317937111)

8. Jackson JC, Gelfand M, Ember CR. 2020 A global analysis of cultural tightness in non-industrial societies. *Proc. R. Soc. B* **287**, 20201036. (doi:10.1098/rspb.2020.1036)
9. Miller DT, McFarland C. 1987 Pluralistic ignorance: when similarity is interpreted as dissimilarity. *J. Pers. Soc. Psychol.* **53**, 298–305. (doi:10.1037/0022-3514.53.2.298)
10. Miller DT, McFarland C. 1991 When social comparison goes away: the case of pluralistic ignorance. In *Social comparison: contemporary theory and research* (eds J Suls. TA Wills), pp. 287–313. Mahwah, NI: Laurence Erlbaum Associates, Inc.
11. Shamir J, Shamir M. 1007 Pluralistic ignorance across issues and over time: information cues and biases. *Public Opin. Q.* **61**, 227–260. (doi:10.1086/297794)
12. Elster J. 1985 Rationality, morality, and collective action. *Ethics* **96**, 136–155. (doi:10.1086/292725)
13. Bicchieri C, Muldoon R, Sontuoso A. 2011 Social norms. In *The Stanford encyclopedia of philosophy* (winter 2018 edition) (ed. EN Zalta). Stanford, CA: Stanford University Press.
14. Zelizer VAR. 1979 *Morals and markets: the development of life insurance in the United States*. New York, NY: Columbia University Press.
15. Hitlin S, Vaisey S. 2013 The new sociology of morality. *Ann. Rev. Sociol.* **39**, 51–68. (doi:10.1146/annurev-soc-071312-145628)
16. Ensminger J, Henrich J. 2014 *Theoretical foundations: the coevolution of social norms, intrinsic motivation, markets, and the institutions of complex societies*. New York, NY: Russel Sage Foundation.
17. Richerson PJ, Gavrilets S, de Waal FBM. 2021 Modern theories of human evolution foreshadowed by Darwin's Descent of Man. *Science* **372**, eaba3776. (doi:10.1126/science.aba3776)
18. Kroneberg C, Kalter F. 2012 Rational choice theory and empirical research: methodological and theoretical contributions in Europe. *Ann. Rev. Sociol.* **38**, 73–92. (doi:10.1146/annurev-soc-071811-145441)
19. Hawkins RX, Goodman ND, Goldstone RL. 2019 The emergence of social norms and conventions. *Trends Cogn. Sci.* **23**, 158–169. (doi:10.1016/j.tics.2018.11.003)
20. Legros S, Cislaghi B. 2019 Mapping the social-norms literature: an overview of reviews. *Perspect. Psychol. Sci.* **15**, 62–80. (doi:10.1177/1745691619866455)
21. Constantino SM, Sparkman G, Kraft-Todd GT, Bicchieri C, Centola D, Shell-Duncan B, Vogt S, Weber EU. 2022 Scaling up change: a critical review and practical guide to harnessing social norms for climate action. *Psychol. Sci. Public Interest* **23**, 50–97. (doi:10.1177/15291006221105279)
22. Gelfand M, Gavrilets S, Nunn N. 2024 Norm dynamics: interdisciplinary perspectives on social norm emergence, persistence, and change. *Annu. Rev. Psychol.* **75**, 1. (doi:10.1146/annurev-psych-033020-013319)
23. Tajfel H. 1981 *Human groups and social categories: studies in social psychology*. Cambridge, UK: Cambridge University Press.
24. Bavel JJV, Pereira A. 2018 The partisan brain: an identity-based model of political belief. *Trends Cogn. Sci.* **22**, 213–224. (doi:10.1016/j.tics.2018.01.004)
25. Bavel JJV, Pärnamets P. 2022 How neurons, norms, and institutions shape group cooperation. *Adv. Exp. Soc. Psychol.* **66**, 59–105. (doi:10.1016/bs.aesp.2022.04.004)
26. Boyd R, Richerson PJ. 1992 Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* **13**, 171–195. (doi:10.1016/0162-3095(92)90032-Y)
27. Fehr E, Fischbacher U. 2004 Third-party punishment and social norms. *Evol. Hum. Behav.* **25**, 63–87. (doi:10.1016/S1090-5138(04)00005-4)
28. Fehr E, Schurtenberger I. 2018 Normative foundations of human cooperation. *Nat. Hum. Behav.* **2**, 458–468. (doi:10.1038/s41562-018-0385-5)
29. Willard AK, 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. *Evol. Hum. Behav.* **41**, 385–396. (doi:10.1016/j.evolhumbehav.2020.07.001)
30. Wrong D. 1961 The oversocialized concept of man in modern sociology. *Am. Sociol. Rev.* **26**, 183–193. (doi:10.2307/2089854)
31. Granovetter M. 1985 Economic action and social structure: the problem of embeddedness. *Am. J. Sociol.* **91**, 481–510. (doi:10.1086/228311)
32. Axelrod R. 1986 An evolutionary approach to norms. *Am. Polit. Sci. Rev.* **80**, 1095–1111. (doi:10.2307/1960858)
33. Gintis H. 2003 The hitchhiker's guide to altruism: gene-culture coevolution, and the internalization of norms. *J. Theor. Biol.* **220**, 407–418. (doi:10.1006/jtbi.2003.3104)
34. Gintis H. 2003 Solving the puzzle of prosociality. *Ration. Soc.* **15**, 155–187. (doi:10.1177/1043463103015002001)
35. Henrich J, Ensminger J. 2014 Theoretical foundations: the coevolution of social norms, intrinsic motivation, markets, and the institutions of complex societies. In *Experimenting with social norms: fairness and punishment in cross-cultural perspective* (eds J Ensminger, J Henrich), pp. 19–44. New York, NY: Russel Sage Foundation.
36. Cooter R. 2000 Do good laws make good citizens? An economic analysis of internalized norms. *Va. Law Rev.* **86**, 1577–1601. (doi:10.2307/1073825)
37. Chudek M, Henrich J. 2011 Culture–gene coevolution, norm-psychology and the emergence of human prosociality. *Trends Cogn. Sci.* **15**, 218–226. (doi:10.1016/j.tics.2011.03.003)
38. Kelly D, Setman S. 2020 The psychology of normative cognition. In *The Stanford encyclopedia of philosophy* (ed. E Zalta). Stanford, CA: Stanford University Press. See https://plato.stanford.edu/entries/psychology-normative-cognition/.
39. Young HP. 2015 The evolution of social norms. *Economics* **7**, 359–387. (doi:10.1146/annurev-economics-080614-115322
40. Dannals JE, Miller DT. 2017 Social norm perception in groups with outliers. *J. Exp. Psychol.: Gen.* **146**, 1342. (doi:10.1037/xge0000336)
41. Mackie G, Moneti F, Shakya H, Denny E. 2015 What are social norms? How are they measured. University of California at San Diego-UNICEF Working Paper, San Diego, CA, USA.
42. Siu AMH, Shek DTL, Law B. 2012 Prosocial norms as a positive youth development construct: a conceptual review. *Sci. World J.* **2012**, 832026. (doi:10.1100/2012/832026)
43. Etzioni A. 2000 Social norms: internalization, persuasion, and history. *Law Soc. Rev.* **34**, 157–178. (doi:10.2307/3115119)
44. Lapinski MK, Rimal RN. 2005 An explication of social norms. *Commun. Theory* **15**, 127–147. (doi:10.1111/j.1468-2885.2005.tb00329.x)
45. Burke MA, Young HP. 2011 Social norms. In *Handbook of social economics*, (eds J Benhabib, A Basic, MO Jackson) pp. 311–338. New York, NY: Elsevier.
46. Bicchieri C, Funcke A. 2018 Norm change: trendsetters and social structure. *Soc. Res.: Int. Q.* **85**, 1–21. (doi:10.1353/sor.2018.0002)
47. Centola D, Becker J, Brackbill D, Baronchelli A. 2018 Experimental evidence for tipping points in social convention. *Science* **360**, 1116–1119. (doi:10.1126/science.aas8827)
48. Granovetter M. 1978 Threshold models of collective behavior. *Am. J. Sociol.* **83**, 1420–1443. (doi:10.1086/226707)
49. Nunn N. 2014 On the dynamics of human behavior: the past, present, and future of culture, conflict, and cooperation. *AEA Pap. Proc.* **112**, 15–37. (doi:10.1257/pandp.20221126)
50. Loewenstein G, Molnar A. 2018 The renaissance of belief-based utility in economics. *Nat. Hum. Behav.* **2**, 166–167. (doi:10.1038/s41562-018-0301-z)
51. Molnar A, Loewenstein G. 2022 Thoughts and players: an introduction to old and new economic perspectives on beliefs. In *The cognitive science of belief: a multidisciplinary approach* (eds J Musolino, J Sommer, P Hemmer), pp. 321–350. Cambridge, UK: Cambridge University Press.
52. Thomas WI. 1928 *The child in America: behavior problems and programs*. New York, NY: Alfred A. Knopf.
53. Olson JM, Zanna MP. 1993 Attitudes and attitude change. *Annu. Rev. Psychol.* **44**, 117–154. (doi:10.1146/annurev.ps.44.020193.001001)
54. Petty RE, Wegener DT, Fabrigar LR. 1997 Attitude and attitude change. *Annu. Rev. Psychol.* **48**, 609–647. (doi:10.1146/annurev.psych.48.1.609)
55. Bohner G, Dickel N. 2011 Attitudes and attitude change. *Annu. Rev. Psychol.* **62**, 391–417. (doi:10.1146/annurev.psych.121208.131609)
56. Albarracin D, Shavitt S. 2017 Attitudes and attitude change. *Annu. Rev. Psychol.* **69**, 299–327. (doi:10.1146/annurev-psych-122216-011911)
57. Tormos R. 2020 *The rhythm of modernization. How values change over time*. Leiden, The Netherlands: Brill.

58. Bursztyn L, Egorov G, Fiorin S. 2020 From extreme to mainstream: the erosion of social norms. *Am. Econ. Rev.* **110**, 3522–3548. (doi:10.1257/aer.20171175)
59. Levy N. 2021 Not so hypocritical after all: belief revision is adaptive and often unnoticed. In *Empirically engaged evolutionary ethics* (eds J De Smedt, H De Cruz), pp. 41–61. Cham, Switzerland: Springer Nature.
60. Inglehart R, Norris P. 2003 *Rising tide: gender equality and cultural change around the world*. Cambridge, UK: Cambridge University Press.
61. Flores AR, Barclay S. 2016 Backlash, consensus, legitimacy, or polarization: the effect of same-sex marriage policy on mass attitudes. *Polit. Res. Q.* **69**, 43–58. (doi:10.1177/1065912915621175)
62. Centola D, Macy M. 2007 Complex contagions and the weakness of long ties. *Am. J. Sociol.* **113**, 702–734. (doi:10.1086/521848)
63. Williams D. 2023 The marketplace of rationalizations. *Econ. Philos.* **39**, 99–123. (doi:10.1017/S0266267121000389)
64. Galesic M, Olsson H, Dalege J, van der Does T, Stein DL. 2021 Integrating social and cognitive aspects of belief dynamics: towards a unifying framework. *J. R. Soc. Interface* **18**, 20200857. (doi:10.1098/rsif.2020.0857)
65. Festinger L. 1957 *A theory of cognitive dissonance*. Evanstone, IL: Row, Peterson and Company.
66. Harmon-Jones E, Harmon-Jones C. 2007 Cognitive dissonance theory after 50 years of development. *Z. Sozialpsychol.* **38**, 7–16. (doi:10.1024/0044-3514.38.1.7)
67. McGrath A. 2017 Dealing with dissonance: a review of cognitive dissonance reduction. *Soc. Pers. Psychol. Compass* **11**, e12362. (doi:10.1111/spc3.12362)
68. Harmon-Jones E, Mills J. 2019 An introduction to cognitive dissonance theory and an overview of current perspectives on the theory. In *Cognitive dissonance: re-examining a pivotal theory in psychology* (ed. E Harmon-Jones). Washington, DC: American Psychological Association.
69. Fudenberg D, Lanzani G, Strack P. 2022 Selective memory equilibrium. Available at SSRN 4015313.
70. Krueger J. 2000 The projective perception of the social world: a building block of social comparison processes. In *Handbook of social comparison: theory and research* (eds J Suls, L Wheeler), pp. 323–351. Springer, New York, NY.
71. Robbins JM, Krueger JI. 2005 Social projection to ingroups and outgroups: a review and meta-analysis. *Pers. Soc. Psychol. Rev.* **9**, 32–47. (doi:10.1207/s15327957pspr0901_3)
72. Krueger JI. 2007 From social projection to social behaviour. *Eur. Rev. Soc. Psychol.* **18**, 1–35. (doi:10.1080/10463280701284645)
73. Premack D, Woodruff G. 1979 Does the chimpanzee have a theory of mind. *Behav. Brain Sci.* **1**, 515–526. (doi:10.1017/S0140525X00076512)
74. Leslie AM, Friedman O, German TP. 2004 Core mechanisms in 'theory of mind'. *Trends Cogn. Sci.* **8**, 528–533. (doi:10.1016/j.tics.2004.10.001)
75. Schaafsma SM, Pfaff DW, Spunt RP, Adolphs R. 2015 Deconstructing and reconstructing theory of mind. *Trends Cogn. Sci.* **19**, 65–72. (doi:10.1016/j.tics.2014.11.007)
76. Wellman HM. 2018 Theory of mind: the state of the art. *Eur. J. Dev. Psychol.* **15**, 728–755. (doi:10.1080/17405629.2018.1435413)
77. Brehm JW. 1966 *The theory of psychological reactance*. New York, NY: Academic Press.
78. Miron AM, Brehm JW. 2006 Reactance theory—40 years later. *Z. Sozialpsychol.* **37**, 9–18. (doi:10.1024/0044-3514.37.1.9)
79. Steindl C, Jonas E, Sittenthaler S, Traut-Mattausch E, Greenberg J. 2015 Understanding psychological reactance: new developments and findings. *Z. Psychol.* **223**, 205–214. (doi:10.1027/2151-2604/a000222)
80. Lerner JS, Li Y, Valdesolo P, Kassam KS. 2015 Emotion and decision making. *Annu. Rev. Psychol.* **66**, 799–823. (doi:10.1146/annurev-psych-010213-115043)
81. Livet P. 2016 Emotions, beliefs, and revisions. *Emot. Rev.* **8**, 240–249. (doi:10.1177/1754073915619019)
82. Pittig A, Brand M, Pawlikowski M, Alpers GW. 2014 The cost of fear: avoidant decision making in a spider gambling task. *J. Anxiety Disord.* **28**, 326–334. (doi:10.1016/j.janxdis.2014.03.001)
83. Bettiga D, Lamberti L. 2020 Future-oriented happiness: its nature and role in consumer decision-making for new products. *Front. Psychol.* **11**, 929. (doi:10.3389/fpsyg.2020.00929)
84. Kumalasari AD, Karremans JC, Dijksterhuis A. 2020 Do people choose happiness? Anticipated happiness affects both intuitive and deliberative decision-making. *Curr. Psychol.* **41**, 6500–6510. (doi:10.1007/s12144-020-01144-x)
85. Tajfel H, Turner JC. 1979 An integrative theory of intergroup conflict. In *The social psychology of intergroup relation* (eds WG Austin, S Worche), pp. 33–47. Monterey, CA: Brooks/Cole.
86. Nail PR, MacDonald G, Levy DA. 2000 Proposal of a four-dimensional model of social response. *Psychol. Bull.* **126**, 454–470. (doi:10.1037/0033-2909.126.3.454)
87. Henrich J. 2020 *The WEIRDest people in the world. How the West became psychologically peculiar and particularly prospective*. New York, NY: Farrar, Straus, and Giroux.
88. Davis R, Campbell R, Hildon Z, Hobbs L, Michie S. 2015 Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. *Health Psychol. Rev.* **9**, 323–344. (doi:10.1080/17437199.2014.941722)
89. Fudenberg D, Tirole J. 1992 *Game theory*. Cambridge, MA: MIT Press.
90. Sandholm WH. 2010 *Population games and evolutionary dynamics*. Cambridge, MA: MIT Press.
91. Young HP. 1998 Social norms and economic welfare. *Eur. Econ. Rev.* **42**, 821–830. (doi:10.1016/S0014-2921(97)00138-4)
92. Bendor J, Swistak P. 2001 The evolution of norms. *Am. J. Sociol.* **106**, 1493–1545. (doi:10.1086/321298)
93. Boyd R, Gintis H, Bowles S, Richerson PJ. 2003 The evolution of altruistic punishment. *Proc. Natl Acad. Sci. USA* **100**, 3531–3535. (doi:10.1073/pnas.0630443100)
94. Hauert C, Traulsen A, Brandt H, Nowak MA, Sigmund K. 2007 Via freedom to coercion: the emergence of costly punishment. *Science* **316**, 1905–1907. (doi:10.1126/science.1141588)
95. Geanakoplos J, Pearce D, Stacchetti E. 1989 Psychological games and sequential rationality. *Games Econ. Behav.* **1**, 60–79. (doi:10.1016/0899-8256(89)90005-5)
96. Rabin M. 1993 Incorporating fairness into game theory and economics. *Am. Econ. Rev.* **83**, 1281–1302.
97. Battigalli P, Dufwenberg M. 2009 Dynamic psychological games. *J. Econ. Theory* **144**, 1–35. (doi:10.1016/j.jet.2008.01.004)
98. Tamarit I, Sánchez A. 2016 Emotions and strategic behaviour: the case of the ultimatum game. *PLoS ONE* **11**, e0158733. (doi:10.1371/journal.pone.0158733)
99. Battigalli P, Corrao R, Dufwenberg M. 2019 Incorporating belief-dependent motivation in games. *J. Econ. Behav. Organ.* **167**, 185–218. (doi:10.1016/j.jebo.2019.04.009)
100. Dufwenberg M, Patel A. 2019 Introduction to special issue on psychological game theory. *J. Econ. Behav. Organ.* **167**, 181–184. (doi:10.1016/j.jebo.2019.05.034)
101. Battigalli P, Dufwenberg M. 2022 Belief-dependent motivations and psychological game theory. *J. Econ. Lit.* **60**, 833–882. (doi:10.1257/jel.20201378)
102. López-Pérez R. 2008 Aversion to norm-breaking: a model. *Games Econ. Behav.* **64**, 237–267. (doi:10.1016/j.geb.2007.10.009)
103. Ridinger G, McBride M. 2020 Reciprocity in games with unknown types. In *Handbook of experimental game theory* (eds C Capra, R Croson, M Rigdon, T Rosenblatt), pp. 271–287. Northampton, MA: Edward Elgar Publishing Limited.
104. Cavalli-Sforza LL, Feldman MW. 1981 *Cultural transmission and evolution: a quantitative approach*. Princeton, NJ: Princeton University Press.
105. Boyd R, Richerson PJ. 1985 *Culture and the evolutionary process*. Chicago, IL: University of Chicago Press.
106. Rashevsky N. 1949 Mathematical biology of social behavior. III. *Bull. Math. Biol.* **11**, 255–271. (doi:10.1007/BF02477979)
107. DeGroot M. 1974 Reaching a consensus. *J. Am. Stat. Assoc.* **69**, 118–121. (doi:10.1080/01621459.1974.10480137)
108. Flache A, Mäs M, Feliciani T, Chattoe-Brown E, Lorenz J. 2017 Models of social influence: towards the next frontiers. *J. Artif. Soc. Soc. Simul.* **20**, 2. (doi:10.18564/jasss.3521)
109. Watts DJ. 2002 A simple model of global cascades on random networks. *Proc. Natl Acad. Sci. USA* **99**, 5766–5771. (doi:10.1073/pnas.082090499)
110. Jackson M. 2010 *Social and economic networks*. Princeton, NJ: Princeton University Press.

111. Easley D, Kleinberg J. 2010 *Networks, crowds and markets*. Cambridge, UK: Cambridge University Press.
112. Akerlof G. 1980 A theory of social custom, of which unemployment may be one consequence. *Q. J. Econ.* **94**, 749–775. (doi:10.2307/1885667)
113. Elster J. 1989 Social norms and economic theory. *J. Econ. Perspect.* **3**, 99–117. (doi:10.1257/jep.3.4.99)
114. Krupka EL, Weber RA. 2013 Identifying social norms using coordination games: why does dictator game sharing vary? *J. Eur. Econ. Assoc.* **11**, 495–524. (doi:10.1111/jeea.12006)
115. Friedkin NE, Johnsen EC. 1990 Social influence and opinion. *J. Math. Sociol.* **15**, 193–205. (doi:10.1080/0022250X.1990.9990069)
116. Fischbacher U, Gächter S. 2010 Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *Am. Econ. Rev.* **100**, 541–556. (doi:10.1257/aer.100.1.541)
117. Smith JE *et al.* 2016 Leadership in mammalian societies: emergence, distribution, power, and payoff. *Trends Ecol. Evol.* **31**, 54–66. (doi:10.1016/j.tree.2015.09.013)
118. Théroude V, Zylbersztejn A. 2019 Cooperation in a risky world. *J. Public Econ. Theory* **22**, 388–407.
119. Andreozzi L, Ploner M, Saral AS. 2020 The stability of conditional cooperation: beliefs alone cannot explain the decline of cooperation in social dilemmas. *Sci. Rep.* **10**, 13610. (doi:10.1038/s41598-020-70681-z)
120. Bicchieri C, Xiao E. 2009 Do the right thing: but only if others do so. *J. Behav. Decis. Making* **22**, 191–208. (doi:10.1002/bdm.621)
121. de Oliveira ACM, Spraggon JM, Denny MJ. 2016 Instrumenting beliefs in threshold public goods. *PLoS ONE* **11**, e0147043. (doi:10.1371/journal.pone.0147043)
122. Kim J, Putterman L, Zhang X. 2022 Trust, beliefs and cooperation: excavating a foundation of strong economies. *Eur. Econ. Rev.* **147**, 104166. (doi:10.1016/j.euroecorev.2022.104166)
123. Gächter S, Molleman L, Nosenzo D. 2021 Why people follow rules. *Open Science Framework*. See https://osf.io/7wz4f/.
124. Jachimowicz JM, Hauser OP, O'Brien JD, Sherman E, Galinsky AD. 2018 The critical role of second-order normative beliefs in predicting energy conservation. *Nat. Hum. Behav.* **2**, 757–764. (doi:10.1038/s41562-018-0434-0)
125. Bonan J, Cattaneo C, Tavoni M. 2020 The interaction of descriptive and injunctive social norms in promoting energy conservation. *Nat. Energy* **5**, 900–909. (doi:10.1038/s41560-020-00719-z)
126. d'Adda G, Dufwenberg M, Passarelli F, Tabellin G. 2020 Social norms with private values: theory and experiments. *Games Econ. Behav.* **124**, 288–304.
127. Górges L, Nosenzo D. 2020 Measuring social norms in economics: why it is important and how it is done. *Anal. Krit.* **42**, 285–311.
128. 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. *Nat. Commun.* **12**, 5452. (doi:10.1038/s41467-021-25734-w)
129. Casoria F, Galeotti F, Villeval MC. 2021 Perceived social norm and behavior quickly adjusted to legal changes during the COVID-19 pandemic. *J. Econ. Behav. Organ.* **190**, 54–65. (doi:10.1016/j.jebo.2021.07.030)
130. Bénabou R, Tirole J. 2016 Mindful economics: the production, consumption, and value of beliefs. *J. Econ. Perspect.* **30**, 141–164.
131. Amelio A, Zimmermann F. 2023 Motivated memory in economics—a review. *Games* **14**, 15. (doi:10.3390/g14010015)
132. Bicchieri C, Dimant E, Sonderegger S. 2023 It's not a lie if you believe the norm does not apply: conditional norm-following and belief distortion. *Games Econ. Behav.* **138**, 321–354. (doi:10.1016/j.geb.2023.01.005)
133. Kuran T. 1995 *Private truths, public lies*. Cambridge, MA: Harvard University Press.
134. Akerlof GA, Dickens WT. 1982 The economic consequences of cognitive dissonance. *Am. Econ. Rev.* **72**, 307–319.
135. Dickens WT. 1986 Crime and punishment again. The economic approach with a psychological twist. *J. Public Econ.* **30**, 97–107. (doi:10.1016/0047-2727(86)90079-4)
136. Akerlof GA, Kranton RE. 2000 Economics and identity. *Q. J. Econ.* **115**, 715–753. (doi:10.1162/003355300554881)
137. Akerlof GA, Kranton RE. 2002 Identity and schooling: some lessons for the economics of education. *J. Econ. Lit.* **40**, 1167–1201. (doi:10.1257/.40.4.1167)
138. Rabin M. 1994 Cognitive dissonance and social change. *J. Econ. Behav. Organ.* **24**, 177–194. (doi:10.1016/0167-2681(94)90066-3)
139. Bernheim D. 1994 A theory of conformity. *J. Polit. Econ.* **102**, 841–877. (doi:10.1086/261957)
140. Bénabou R, Tirole J. 2006 Incentives and prosocial behavior. *Am. Econ. Rev.* **96**, 1652–1678. (doi:10.1257/aer.96.5.1652)
141. Bénabou R, Tirole J. 2011 Laws and norms. *NBER Working Paper Series*, working paper no. 17579. National Bureau of Economic Research. See https://www.nber.org/system/files/working_papers/w17579/w17579.pdf.
142. Shmailov MM. 2016 *Intellectual pursuits of Nicolas Rashevsky*. Basel, Switzerland: Birkhäuser.
143. Rashevsky N. 1968 *Looking at history through mathematics*. Cambridge, MA: MIT Press.
144. Landahl HD. 1938 A contributions to the mathematical biophysics of error elimination. *Psychometrika* **3**, 107–125. (doi:10.1007/BF02289306)
145. Petrov AP, Proncheva OG. 2018 Modeling propaganda battle: decision-making, homophily, and echo chambers. In *Proceedings of the conference on artificial intelligence and natural language AINL 2018* (eds D Ustalov, A Filchenkov, L Pivovarova, J Žižka), pp. 197–209. Berlin, Germany: Springer.
146. Petrov AP, Proncheva OG. 2020a Stationary states in a model of position selection by individuals. *Comput. Math. Math. Phys.* **60**, 1737–1746. (doi:10.1134/S0965542520100115)
147. Petrov AP, Proncheva OG. 2020 Modeling position selection by individuals during informational warfare with a two-component Agenda. *Math. Models Comput. Simul.* **12**, 154–163. (doi:10.1134/S207004822002009X)
148. Schelling T. 1971 Dynamic models of segregation. *J. Math. Sociol.* **1**, 143–186. (doi:10.1080/0022250X.1971.9989794)
149. Schelling T. 1971 On the ecology of micromotives. *Public Int.* **25**, 61–98.
150. Young H. 2009 Innovation diffusion in heterogeneous populations: contagion, social influence, and social learning. *Am. Econ. Rev.* **99**, 1899–1924. (doi:10.1257/aer.99.5.1899)
151. Yin CC. 1998 Equilibria of collective action in different distribution of protest thresholds. *Public Choice* **97**, 536–567. (doi:10.1023/A:1004921725560)
152. Efferson C, Vogt S, Fehr E. 2019 The promise and the peril of using social influence to reverse harmful traditions. *Nat. Hum. Behav.* **4**, 55–68. (doi:10.1038/s41562–019–0768–2)
153. Kuran T. 1989 Sparks and prairie fires: a theory of unanticipated political revolution. *Public Choice* **61**, 41–74. (doi:10.1007/BF00116762)
154. Centola D, Willer R, Macy M. 2005 The emperor's dilemma: a computational model of self-enforcing norms. *Am. J. Sociol.* **110**, 1009–1040. (doi:10.1086/427321)
155. Granovetter MS. 1973 The strength of weak ties. *Am. J. Sociol.* **78**, 1360–1380. (doi:10.1086/225469)
156. Gavrilets S. 2020 The dynamics of injunctive social norms. *Evol. Hum. Sci.* **2**, e60. (doi:10.1017/ehs.2020.58)
157. McCullen NJ, Rucklidge AM, Foxon CSEBTJ, Gale WF. 2013 Multiparameter models of innovation diffusion on complex networks. *SIAM J. Appl. Dyn. Syst.* **12**, 515–532. (doi:10.1137/120885371)
158. Brock WA, Durlauf SN. 2001 Discrete choice with social interactions. *Rev. Econ. Stud.* **68**, 235–260. (doi:10.1111/1467-937X.00168)
159. Azar O. 2004 What sustains social norms and how they evolve? The case of tipping. *J. Econ. Behav. Organ.* **54**, 49–64. (doi:10.1016/j.jebo.2003.06.001)
160. Azar OH. 2008 Evolution of social norms with heterogeneous preferences: a general model and an application to the academic review process. *J. Econ. Behav. Organ.* **65**, 420–435. (doi:10.1016/j.jebo.2006.03.006)
161. te Velde VL. 2022 Heterogeneous norms: social image and social pressure when people disagree. *J. Econ. Behav. Organ.* **194**, 319–340. (doi:10.1016/j.jebo.2021.12.013)
162. Brekke KA, Kverndokk S, Nyborg K. 2003 An economic model of moral motivation. *J. Public Econ.* **87**, 1967–1983. (doi:10.1016/S0047-2727(01)00222-5)
163. Houle C, Ruck DJ, Bentley RA, Gavrilets S. 2022 Inequality between identity groups and social

unrest. *R. Soc. Interface* **19**, 20210725. (doi:10.1098/rsif.2021.0725)

164. Yang L, Constantino SM, Grenfell BT, Weber EU, Levin SA, Vasconcelos VV. 2022 Sociocultural determinants of global mask-wearing behavior. *Proc. Natl Acad. Sci. USA* **119**, e2213525119. (doi:10.1073/pnas.2213525119)
165. Kuran T, Sandholm WH. 2008 Cultural integration and its discontents. *Rev. Econ. Stud.* **75**, 201–228. (doi:10.1111/j.1467-937X.2007.00469.x)
166. Della Lena S, Dindo P. 2019 On the evolution of norms in strategic environments. Working paper no. 16/WP/2019, ISSN 1827–3580. Social Science Research Network. See https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3407246.
167. Martins AC. 2020 Discrete opinion dynamics with *M* choices. *Eur. Phys. J. B* **93**, 1–10. (doi:10.1140/epjb/e2019-100298-3)
168. Acharya A, Blackwell M, Sen M. 2018 Explaining preferences from behavior: a cognitive dissonance approach. *J. Polit.* **80**, 400–411. (doi:10.1086/694541)
169. Calabuig V, Olcina G, Panebianco F. 2016 The erosion of personal norms and cognitive dissonance. *Appl. Econ. Lett.* **23**, 1265–1268. (doi:10.1080/13504851.2016.1150940)
170. Calabuig V, Olcina G, Panebianco F. 2017 The dynamics of personal norms and the determinants of cultural homogeneity. *Ration. Soc.* **29**, 322–354. (doi:10.1177/1043463117717233)
171. Calabuig V, Olcina G, Panebianco F. 2018 Culture and team production. *J. Econ. Behav. Organ.* **149**, 32–45. (doi:10.1016/j.jebo.2018.03.004)
172. Martins ACR. 2009 Continuous opinions and discrete actions in opinion dynamics problems. *Int. J. Mod. Phys. C* **19**, 617–624. (doi:10.1142/S0129183108012339)
173. 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. (doi:10.1063/5.0004787)
174. Mo Y, Sun J. 2022 Coevolution of collective opinions and actions under two different control inputs. *Inf. Sci.* **608**, 1632–1650. (doi:10.1016/j.ins.2022.06.094)
175. Aghbolagh HD, Ye M, Zino L, Cao M, Chen Z. 2022 Coevolutionary dynamics of actions and opinions in social networks. *IEEE Trans. Automat. Cont.* **68**, 7708–7723.
176. Gavrilets S. 2021 Coevolution of actions, personal norms, and beliefs about others in social dilemmas. *Evol. Hum. Sci.* **3**, e44. (doi:10.1017/ehs.2021.40)
177. Tverskoi D, Guido A, Andrighetto G, Sánchez A, Gavrilets S. 2023 Disentangling material, social, and cognitive determinants of human behavior and beliefs. *Humanit. Soc. Sci. Commun.* **10**, 1–13. (doi:10.1057/s41599-023-01745-4)
178. Tverskoi D, Babu S, Gavrilets S. 2022 The spread of technological innovations: effects of psychology, culture and policy interventions. *R. Soc. Open Sci.* **9**, 211833. (doi:10.1098/rsos.211833)
179. Gavrilets S, Richerson PJ. 2022 Authority matters: propaganda and the coevolution of behaviour and attitudes. *Evol. Hum. Sci.* **4**, e51. (doi:10.1017/ehs.2022.48)
180. Alger I, Weibull JW. 2013 Homo morales—preference evolution under incomplete information and assortative matching. *Econometrica* **81**, 2269–2302. (doi:10.3982/ECTA10637)
181. Gavrilets S, Richerson PJ. 2017 Collective action and the evolution of social norm internalization. *Proc. Natl Acad. Sci. USA* **114**, 6068–6073. (doi:10.1073/pnas.1703857114)
182. Lozano P, Gavrilets S, Sánchez A. 2020 cooperation, social norm internalization, and hierarchical societies. *Sci. Rep.* **21**, 15359. (doi:10.1038/s41598-020-71664-w)
183. Akcay E, van Cleve J. 2020 Internalizing cooperative norms in group-structured populations. In *Social cooperation and conflict: biological mechanisms at the interface* (eds W Wilczynski, S Brosnan), pp. 26-43. Cambridge, UK: Cambridge University Press.
184. Kimbrough EO, Vostroknutov A. 2019 A theory of injunctive norms. See http://www.vostroknutov.com/pdfs/axinorms12_02.pdf.
185. Tremewan J, Vostroknutov A. 2021 An informational framework for studying social norms. In *A research agenda for experimental economics* (ed. A Chaudhuri), pp. 19–42. Northampton, MA: Edward Elgar Publishing.
186. Scheufle DA, Moy P. 2000 Twenty-five years of the spiral of silence: a conceptual review and empirical outlook. *Int. J. Public Opin. Res.* **12**, 3–28. (doi:10.1093/ijpor/12.1.3)
187. Przepiorka W, Diekmann A. 2021 Parochial cooperation and the emergence of signalling norms. *Phil. Trans. R. Soc. B* **376**, 20200294. (doi:10.1098/rstb.2020.0294)
188. Nakamaru M, Levin SA. 2004 Spread of two linked social norms on complex interaction networks. *J. Theor. Biol.* **230**, 57–64. (doi:10.1016/j.jtbi.2004.04.028)
189. Zhang Y, Leezer J. 2009 Emergence of social norms in complex networks. In *2009 Int. Conf. on Computational Science and Engineering*, vol. 4, pp. 549–555. New York, NY: IEEE.
190. Ushchev P, Zenou Y. 2020 Social norms in networks. *J. Econ. Theory* **185**, 104969. (doi:10.1016/j.jet.2019.104969)
191. Yee N, Bailenson JN, Urbanek M, Chang F, Merget D. 2007 The unbearable likeness of being digital: the persistence of nonverbal social norms in online virtual environments. *CyberPsychol. Behav.* **10**, 115–121. (doi:10.1089/cpb.2006.9984)
192. Nyborg K *et al*. 2016 Social norms as solutions. *Science* **354**, 42–43. (doi:10.1126/science.aaf8317)