REVIEW ARTICLE
https://doi.org/10.1038/s41562-018-0385-5

nature human behaviour

Normative foundations of human cooperation

Ernst Fehr* and Ivo Schurtenberger

A large literature shares the view that social norms shape human cooperation, but without a clean empirical identification of the relevant norms almost every behaviour can be rationalized as norm driven, thus rendering norms useless as an explanatory construct. This raises the question of whether social norms are indeed causal drivers of behaviour and can convincingly explain major cooperation-related regularities. Here, we show that the norm of conditional cooperation provides such an explanation, that powerful methods for its empirical identification exist and that social norms have causal effects. Norm compliance rests on fundamental human motives ('social preferences') that also imply a willingness to punish free-riders, but normative constraints on peer punishment are important for its effectiveness and welfare properties. If given the chance, a large majority of people favour the imposition of such constraints through the migration to institutional environments that enable the normative guidance of cooperation and norm enforcement behaviours.

Normative constraints and prescriptions are ubiquitous and pervade almost every aspect of human social life, from the mundane to the most profound. They appear to play a role in all social groups and have been documented for a large number of ancient societies[1,2], but also play a role in contemporary societies. Norms are part of the weave of social life and, if obeyed, they make it predictable, constitute social order and become the cement of society[3], but if compliance with fundamental norms breaks down — as it sometimes happens in the aftermath of lost wars or natural disasters — disorder, revolt or revolutionary chaos prevails, and life becomes "solitary, poor, nasty, brutish and short"[4].

Human cooperation is an equally ubiquitous phenomenon that is present in some form in almost every social relationship and is key for the success of social units from the family to the nation state to global organizations[5]. Sometimes, cooperation is in the material self-interest of people, but here we are interested in those aspects of cooperation where economic incentives alone are not sufficient to induce individuals to cooperate because free-riding would maximize their private gains. Throughout human history, myriad scenarios are characterized by such social dilemmas. Every successful sequential exchange, in which one party provides the quid pro quo first, constitutes an act of cooperation. Our ancestors also faced social dilemmas when they hunted large game, during tribal warfare or during reciprocal food sharing in times of need. Contemporary humans encounter them in team production settings and whenever there is a tension between one's own interest and the reputation of the company, when paying taxes despite low probabilities of being caught in tax evasion or in the context of problems of a truly global scale such as climate change.

To what extent and how do social norms shape human cooperation? There are social norms, such as the norm to keep a promise or the honesty norm, that affect behaviour in cooperative contexts, but are not directly related to cooperation. For example, the honesty norm proscribes lying and that implies that one should also not lie to evade taxes and the norm to keep one's promises implies that one should also keep promises made to an exchange partner, but these norms have implications that go far beyond cooperative contexts. In this Review, we focus instead on social norms that directly prescribe, and limit their prescription to, cooperation and punishment behaviours in social dilemma and collective action contexts. An example of such a norm is the 'conditional cooperation norm', which we define in more detail below. We ask whether these norms can, in principle, explain major behavioural regularities observed in collective action contexts, what the properties of these norms are and which motivational forces ensure compliance with them, and whether they indeed guide or are the causal drivers of behaviour in collective action.

To answer these questions requires a clear definition of social norms. We define them as commonly known standards of behaviour that are based on widely shared views of how individual group members ought to behave in a given situation[3,6,7]. This definition entails three crucial features of social norms. First, a social norm establishes a normative standard of behaviour that applies to a particular group and to a particular situation. Second, the norm is not defined in terms of group members' actual behaviour nor in terms of their motives, their compliance or the conditions under which compliance occurs; it is exclusively defined in terms of a normative behavioural standard, that is, how group members ought to behave. Third, this normative standard and its widely shared approval is commonly known by group members.

Because a norm requires that the normative standard is widely shared, non-compliance with the norm automatically triggers some disapproval. Therefore, if individuals dislike the thought that others disapprove of them, they automatically have some incentive to comply, although, as we will see, this incentive may not necessarily be sufficient to induce compliance. We will therefore also ask which kind of other motives and mechanisms support compliance with social cooperation norms and whether they act as a constraint on potential non-compliers or are part of the 'intrinsic' motivation of individuals. In this context, we will also ask whether the (peer) punishment of norm violators is itself a social norm or whether it is driven by other motivational sources.

Regularities in cooperation-related behaviours?

To assess the role of social norms for human cooperation, we first describe major behavioural regularities observed in experimental social dilemma games. With the exception of experiments that allow for face-to-face communication, the subjects in these games are anonymous to each other. They play for real money under conditions where complete free-riding is the dominant strategy for selfish individuals in one-shot games and backward induction implies that complete free-riding is also predicted in the finitely repeated

University of Zurich, Department of Economics, Zurich, Switzerland. *e-mail: ernst.fehr@econ.uzh.ch

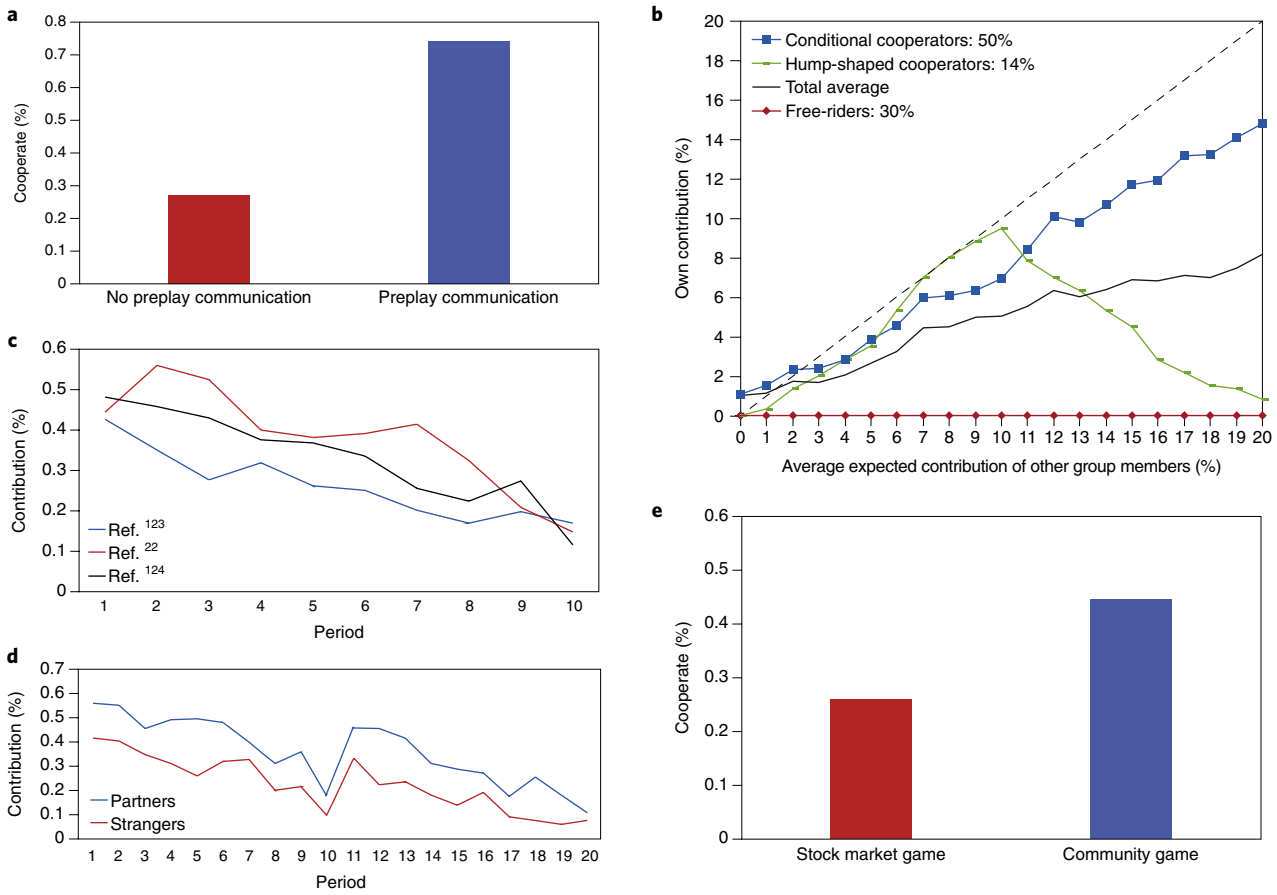

Fig. 1 | Illustrations of behavioural regularities 1–5 in cooperation experiments. a, Cooperation rates in a one-shot social dilemma game with and without pre-play communication among the subjects[8] (regularity 1). **b**, Higher expectations of other group members' cooperation causes on average an increase in individual's own cooperation (regularity 2), but individuals are heterogeneous with, typically, a majority of conditional cooperators, a significant minority of full free-riders and some share of hump-shaped conditional cooperators[12]. The dashed line is the 45° line. **c**, Decline in cooperation rates over time in finitely repeated public goods experiments in which free-riding is the payoff maximizing strategy for selfish subjects[22,123,124] (regularity 3). **d**, Cooperation rates in partner treatments are typically higher than those in stranger treatments. In this study, subjects initially believed they had to interact for ten periods after which the experimenter implemented a surprise restart of the same ten-period experiment (regularities 3 and 4)[19]. **e**, Merely calling the prisoners' dilemma a community game — as opposed to a stock market game — increases cooperation (regularity 5)[20]; but if the game is played sequentially, this framing effect vanishes. Panels adapted from: **a**, ref. [8], APA; **b**, ref. [12], Elsevier; **c**, ref. [22], AEA; ref. [123], Springer Nature; ref. [124], Elsevier; **d**, ref. [19], Elsevier; **e**, ref. [20], Elsevier.

game. We deliberately restrict ourselves to these experimental settings because to precisely identify the role of social norms, their predictions must differ from the self-interest model. Field evidence, in contrast, typically does not allow self-interest to be ruled out with perfect certainty, but below we point out that many lab observations resemble regularities that are observed in naturally occurring environments. Second, we discuss the ability of social norms to provide a parsimonious explanation for the regularities.

The following patterns are among the key findings in the literature:

(1) Although complete free-riding is a dominant strategy, a substantial share of the subjects cooperate in one-shot social dilemmas but free-riding frequently also prevails[8,9]. However, if subjects can communicate about the game before they play it cooperation strongly increases[8,10,11] (Fig. 1a).

(2) A large proportion of subjects are conditional cooperators, that is, the belief that other group members cooperate at high levels induces them to also cooperate at high levels but if others are believed to decrease their cooperation, these individuals also decrease their cooperation[12–14] (Fig. 1b).

(3) In finitely repeated public goods games, cooperation is initially relatively high but often declines to very low levels towards the final periods[15,16]. This holds regardless of whether the game is framed as a public goods game or as a common-pool resource game[17]. If subjects play the finitely repeated game several times — but each time with a new composition of group members — cooperation always starts high and becomes very low towards the end of the game[18] (Fig. 1c).

(4) In finitely repeated public goods games, cooperation is generally higher in groups with a stable group composition ('partner matching') compared with random reassignment of individuals to groups in every period ('stranger matching')[14,18,19] (Fig. 1d).

(5) Merely framing a simultaneously played prisoners' dilemma game differently by calling it community game instead of stock market game typically causes substantial increases in cooperation rates. However, if the game is played sequentially this framing effect vanishes[20,21] (Fig. 1e).

(6) There is a widespread willingness to punish free-riders even in one-shot interactions although it is costly for the punisher[22–24] (Fig. 2a). Furthermore, peer punishment opportunities in

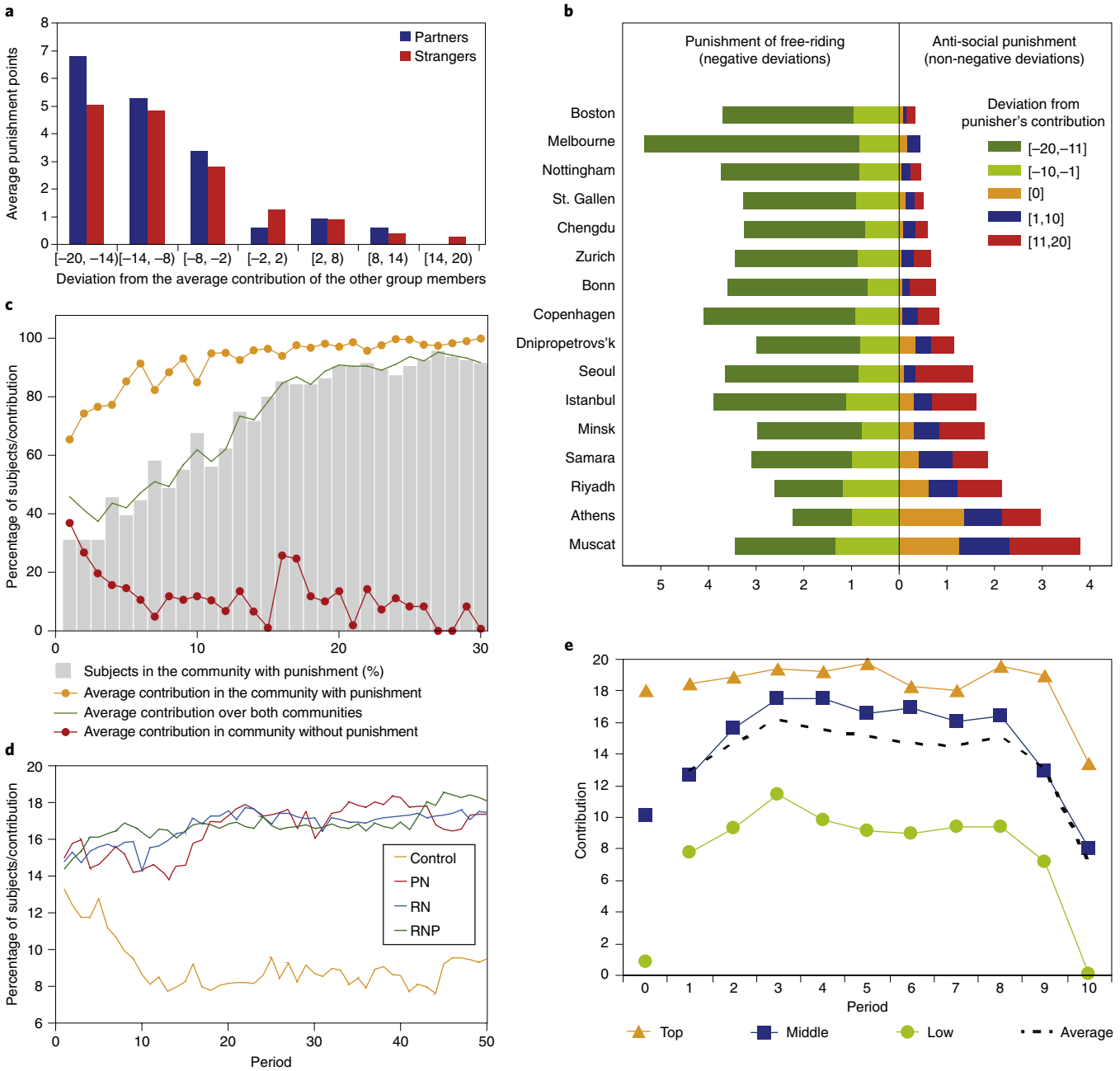

Fig. 2 | Illustrations of behavioural regularities 6–10 in cooperation experiments. a, Punishment of group members — measured in terms of the experienced percentage reduction in income — as a function of the deviation of their cooperation level from the average cooperation of other group members (regularities 6 and 7)[22]. Punishment of free-riders is very high but above average cooperators also face some 'perverse' punishment. **b**, Evidence for strong cultural differences in antisocial punishment of cooperators (regularity 7)[31]. **c**, Subjects could choose in every period whether they want to be in the community with a peer punishment opportunity or the community without this opportunity. The vast majority of subjects eventually preferred the community with peer punishment (regularity 8)[34]. **d**, The opportunity to punish peers after they observed others' cooperation levels (treatment PN) leads to large increases in cooperation relative to a control treatment without peer punishment (control). The opportunity to mutually reward each other (RN) leads to similarly high cooperation levels compared with PN and treatments with both reward and punishment (RNP)[38] (regularity 6 and 9). **e**, High cooperators in a one-shot prisoner's dilemma are grouped together in a subsequent ten-period public goods game. Likewise, the middle and the low cooperators are grouped together. High cooperators achieve very high cooperation rates during the first nine periods (regularity 10)[40]. Panels adapted from: **a**, ref. [22], AEA; **c**, ref. [34], Elsevier. Panels reproduced from: **b**, ref. [31], AAAS; **d**, ref. [38], AAAS; **e**, ref. [40], Oxford Univ. Press.

repeated interactions cause large cooperation increases and often lead to near complete and stable cooperation under partner matching[22,25] (Fig. 2d). These opportunities are, however, also associated with high initial costs such that group welfare does not increase (or even decreases) for roughly ten periods[22,25].

(7) The effectiveness of peer punishment in enhancing cooperation is undermined if punishment threats signals selfish intentions[26–29] and by 'perverse'[30] or 'antisocial' punishment[31,32] of cooperators in public goods games by those who free-ride — a tendency that varies strongly across different cultures (Fig. 2b).

(8) Despite the high initial cost caused by peer punishment, subjects eventually prefer environments with a peer punishment opportunity almost unanimously over an environment that rules out peer punishment[33,34] (Fig. 2c).
(9) The opportunity to reward cooperators — either through the preferred choices of cooperative partners[35] or through the direct rewarding of those with a high reputation for cooperation[36–39] causes large cooperation increases (Fig. 2d).
(10) Stable cooperation at very high levels can be achieved either when cooperative individuals are exogenously matched together[40,41] (Fig. 2e) or in intergenerational public goods games when individuals can give advice that is common knowledge to the next generation[42].

An important question is how insights gained in lab experiments relate to behaviour in naturally occurring environments. Several studies[43–52] demonstrate that individuals' behaviour in the lab is predictive of their behaviour in relevant field settings. For instance, people who tend to contribute more in public goods games are more likely to participate in local and national accountability institutions[43]. Fishermen who exhibit more cooperation in a laboratory public goods game also show more cooperative behaviour in a real world common-pool resource problem by employing more sustainable fishing techniques; they use buckets with larger holes such that younger shrimps are not yet caught[44]. Another study[45] shows that Ethiopian communities that face serious common-pool resource problems are better able to maintain the commons if they have a higher share of people that display conditional cooperation in a public goods experiment. This study also provides evidence suggesting that causality runs from conditional cooperation to better maintenance of the commons resource. Behaviours consistent with conditional cooperation are also observed in field experiments[46].

Can social norms explain cooperation-related behaviours?

All the abovementioned regularities are largely incompatible with the pure self-interest model, that is, they cannot be explained if it is common knowledge that all actors are rational and selfish. If free-riding is the dominant strategy at each contribution stage, there is also no incentive to enact costly punishment/rewards to induce cooperation, and reassortment or communication will not be effective either.

However, many of these regularities can, at least in principle, be explained if one directly assumes that a significant share of individuals has a desire to comply with a social cooperation norm[53]. We call this the direct social norms approach[7,41,54,55] because it directly assumes (1) the existence of a norm c^* that is defined in terms of a specific behaviour and (2) that individuals have an intrinsic desire to comply with c^* without providing a deeper micro-foundation of c^* and motives for norm compliance. In the context of cooperation, c^* describes the smallest cooperation level that is consistent with the normative prescription. Formally, this can be modelled by a utility function u_i in which individual i's utility depends positively on i's own material payoff x_i (which depends on all players' choices) while negative deviations of i's behaviour c_i from the social norm c^* ($c_i < c^*$) generate some disutility:

$$u_i = \begin{cases} x_i - \gamma_i(c_i - c^*)^2 & \text{if } c_i < c^* \\ x_i & \text{if } c_i \geq c^* \end{cases} \quad (1)$$

The term $\gamma_i(c_i - c^*)^2$ denotes the psychic cost of deviating from the social norm (for simplicity these costs increase quadratically with negative deviations from the norm ($c_i - c^*$) and $\gamma_i \geq 0$ captures an individual's strength of the desire to conform to the norm. This approach represents a simple theory of conformism based on the assumption that negative deviations from the norm are, for some reason, psychologically costly for individuals with a strictly positive γ_i. In the context of cooperation, higher individual cooperation levels c_i are costly and thus reduce the individual's material payoff x_i but if c_i is below the norm c^* an increase in c_i reduces the costs of non-conformity $\gamma_i(c_i - c^*)^2$. For a sufficiently large level of γ_i, the individual has therefore an incentive to obey the social norm c^*. Note that we assume for simplicity that positive deviations from the norm c^* have no psychological costs or benefits.

It is almost surely the case that the psychological cost of negative deviations from c^* (that is, the γ_i's) vary across people but the assumption that there are some psychological costs of negative deviations makes sense in the light of the definition of a social norm because that definition implies that group members widely approve of the norm and that this is known by the subjects. Thus, subjects know that if they violate a social norm they are likely to face the disapproval of other people and for some people even the mere thought that others might disapprove of their action could constitute a psychological cost. In principle, γ_i could also represent the cost of deviating from a behavioural habit acquired in social life. Or the psychological cost of noncompliance could positively depend on how widely the norm is shared among the group members. However, in the following we assume for simplicity that γ_i is fixed and varies across individuals.

Unconditional normative prescriptions like 'be selfless', 'do the right thing' or 'be moral' cannot explain the behavioural regularities described above. For example, they cannot explain communication effects (regularity 1), the decline in cooperation over time (regularity 2) or the higher levels of cooperation in a partner compared with stranger matching (regularity 3). In contrast, a social norm of conditional cooperation can help explain all regularities but those described in regularities 6–8. This norm prescribes full cooperation as long as other group members also cooperate fully, but if others' average cooperation becomes smaller it is normatively justified to match this reduction, that is, the conditional cooperation norm prescribes to contribute at least as much as others' average contribution. Note that this implies that subjects' empirical beliefs about others' average cooperation become an important determinant of their cooperation levels — the more others cooperate, the higher is the incentive to cooperate for an individual with a positive γ_i, which explains regularity 2.

But this norm can also explain regularity 1: subjects with a very small γ_i ($\gamma_i \approx 0$) will defect, while those with a sufficiently large γ_i and a high expectation about others' cooperation will cooperate in one-shot social dilemmas. Moreover, under face-to-face communication, subjects often promise to each other to cooperate[56], which is very likely to increase beliefs about others' cooperation. This increase in others' expected cooperation will then induce individuals with a sufficiently positive γ_i to increase their cooperation levels.

It has been shown[5,57] that the existence of imperfect conditional cooperators is the key ingredient for explaining regularity 3 — the decay of cooperation over time in finitely repeated games. Conditional cooperation is imperfect if an individual does not match other group member's average cooperation perfectly but cooperates somewhat less than others are expected to cooperate on average. The above utility function assumes that people care positively for their own payoff and, therefore, individuals with a positive yet sufficiently low γ_i will not obey the norm c^* perfectly but reduce c_i somewhat below c^*, which implies imperfect conditional cooperation. However, if many individuals cooperate less than what each of them expect others to cooperate, jointly their expectations are too optimistic, which results in a downwards revision of their expectations and this then leads — via conditional cooperation — to a further decline in their cooperation rates, and so on.

The existence of a conditional cooperation norm can also explain regularity 4 — the higher cooperation rates under a stable group composition — and regularity 5, the existence of a framing

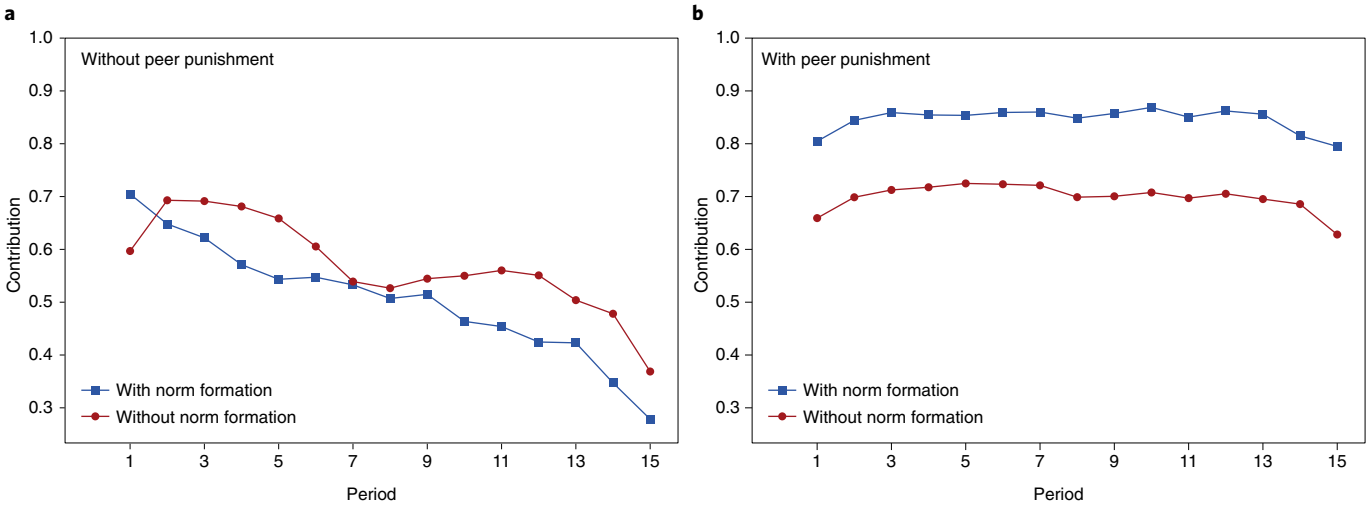

Fig. 3 | The effect of social norms with and without punishment[88]**.** Average normalized contributions over time (1 = full contributions; 0 = complete free-riding) in fixed groups of four subjects that play a public goods game for 15 periods. **a**, Treatments without punishment. **b**, Treatments with punishment. Treatments with a punishment opportunity allow for the counter-punishment of those who punish free-riders to examine whether norms have a causal impact in an environment that has been shown to be hostile for human cooperation[32].

effect on cooperation in the simultaneously played prisoners' dilemma but not in the sequentially played prisoners' dilemma[20]. When there is a stable group composition, even selfish individuals (that is, those with $\gamma_i \approx 0$) have temporarily a strong incentive to cooperate because this generates benefits in future periods by inducing conditional cooperators to keep contributing (regularity 4). To explain regularity 5, recall that if there is a norm of conditional cooperation, subjects who derive disutility from norm violations adjust their cooperation level to what they believe the other player will do in the simultaneously played prisoners' dilemma. For optimistic beliefs they cooperate; for pessimistic beliefs they defect. Under the plausible assumption that the label 'community game' renders beliefs about the partner's cooperation more optimistic, conditionally cooperative subjects will cooperate with higher frequency. However, for the second mover in the sequential prisoners' dilemma, beliefs are irrelevant because this player already knows exactly what the first mover did. Thus, the frame can no longer change beliefs and therefore becomes irrelevant; and if a rational first mover anticipates the absence of a framing effect (s)he has no reason to condition behaviour on the frame either. Note that this explanation does not assume that the conditional cooperation norm changes across frames or between simultaneous and sequential play.

The conditional cooperation norm can also explain why the addition of mutual reward opportunities to a public goods game increases cooperation (regularity 9). In the presence of mutual reward opportunities, subjects can observe the cooperation level of other group members in the public goods game, after which they can spend money on rewarding other group members that costs them less than it benefits the rewarded subjects. This basically boils down to the opportunity of playing another bilateral prisoners' dilemma with each of the other group members after they observed others' cooperation levels. Obviously, the norm of conditional cooperation also applies to these prisoners' dilemma games and because cooperation in the public goods game can serve as a signal of cooperative intent, cooperation in the public goods game fosters the belief that an individual will also cooperate in the prisoners' dilemma. Therefore, mutual reward opportunities increase the incentive to cooperate in the public goods game.

Finally, the conditional cooperation norm can also help explain regularity 10, that is, why the assignment of cooperative individuals to the same group may cause high and stable cooperation. In terms of the direct social norms approach, cooperative individuals may be viewed as those with a sufficiently high γ_i such that for them perfect obedience with the norm ($c_i = c^*$) becomes optimal. If, in addition, these subjects are told that they are grouped together with other cooperators[40] they start with high expectations that trigger high cooperation, which confirms the initial high expectation. The publicly known sorting of cooperative individuals into a group thus renders cooperation an equilibrium outcome.

For a similar reason, the existence of a conditional cooperation norm may also explain why cooperative advice by a previous generation of players that is made common knowledge among all current group members (regularity 10) causes large increases in cooperation rates. Cooperative advice that is common knowledge induces a general increase in the expected cooperation of other group members[42]. Together with the norm of conditional cooperation, the increased expectations then give rise to a general increase in cooperation rates.

However, there are of course other motives — such as equity or reciprocity motives — that make similar predictions to those described above. Moreover, a norm of conditional cooperation cannot explain why subjects punish free-riders (regularity 6) nor subjects' preferences for playing the public goods game in an environment that allows for peer punishment (regularity 8). This follows simply from the fact that the conditional cooperation norm is defined in the space of cooperation behaviour and not in the space of punishment behaviour. One may, of course, stipulate the existence of another norm that renders punishment of free-riders a socially desirable act but there is little evidence for this and it shows one of the drawbacks of an unconstrained direct social norms approach. By stipulating that a particular behaviour constitutes a social norm, it is possible to explain any behaviour, which renders such an approach irrefutable and thus empty — a problem that we take up later.

In real life, peer punishment ranges in severity and costliness from a simple raised eye brow or a hurtful smile to outright ridicule or ostracism and the expulsion from social groups. Nevertheless, punishments in the lab capture key features of real-life sanctions and also teach us that a significant share of participants will enact punishment systematically, even when it is costly and there is no personal material benefit for them.

The psychology of norm compliance

To make progress in understanding the potential impact of social norms on human cooperation, it is important to examine more closely the psychological reasons that induce individuals to comply with social norms. The direct social norms approach stipulates a normative behavioural standard and a psychological cost of non-compliance but does not provide a microfoundation for the behavioural standard and is typically not very explicit about the psychological cost of non-compliance. In principle, these costs could arise because individuals may be averse to actual, anticipated or merely imagined disapproval when deviating from the norm. In this case, compliance rests on an internalized desire for conformism, which has been challenged long ago as a general and sufficient basis for norm compliance[58].

Another reason for psychological costs of norm compliance arises if individuals have an intrinsic desire for equity or fairness and social norms play a role in defining what is perceived as equitable or fair[59–61]. This case is also methodologically interesting because it implies that a collective phenomenon — the social norm — substantively affects the content of individuals' motivation by influencing what is perceived as fair, while the intrinsic desire for fairness then ensures compliance with the norm. A third reason for costs of deviating from the social norm could be that individuals have a desire to reciprocate the behaviour of relevant others[62–64]. In this case, the reciprocity motive applies, that is, the tendency to reward kind intentions with kindness ('positive reciprocity') and to punish hostile or unkind intentions ('negative reciprocity'). Note, however, that this motive requires a definition of what constitutes kind and unkind behaviour, which is typically also based on some normative notion of fairness/equity. For a reciprocally motivated individual, psychic costs of non-compliance arise, if the individual fails to reciprocate to a kind act with kindness or does not retaliate to a hostile act with a hostile response. Therefore, as in the case of fairness/equity motives, the reciprocity motive becomes operative on the basis of what is perceived as fair/kind and unfair/unkind.

A fourth reason for psychic costs of non-compliance arises if individuals have a propensity towards guilt aversion[65–67]. This theory rests on the idea that individuals experience the aversive, utility-decreasing emotion of guilt if they disappoint others. A social norm only exists if group members widely approve of the norm, and if there is widespread compliance then an individual act of non-compliance is almost surely disappointing other individuals. For example, if a subject believes that her partner in the prisoners' dilemma expects her to cooperate, then she disappoints him/her if she defects, and if the subject feels guilt and anticipates this emotion, she has an incentive to cooperate. Therefore, to the extent to which social norms generate the belief that others expect the individual to comply — a very likely belief in the presence of widespread compliance — a guilt-averse individual has some incentive to cooperate. However, if a social norm is systematically violated, such that the individual does not face a general expectation of compliance, a guilt-averse individual has no reason to comply with the norm. Guilt aversion is thus likely to generate conditional norm compliance behaviour that is mediated by individuals' beliefs about what others expect from them.

Finally, self-image theory assumes that individuals assign an intrinsic value to their self-image as a prosocial individual[68]. In this case, non-compliance with socially beneficial norms is detrimental for their self-image and provides a psychological deterrent for non-compliance. Similar to the case of fairness and reciprocity theories, this approach rests on some pre-existing notion — the notion of 'prosociality' — which is likely to be shaped by social norms.

It is interesting that all the abovementioned approaches rest on assumptions about individuals' intrinsic motivational properties. These motives — for example, the desire for fairness — are assumed to be stable across contexts. Stability in the desire for fairness does not mean, however, that the content of what is defined as fair is stable across contexts. It only means that individuals' preferences for implementing what is defined as fair, that is, their willingness to pay to implement the fair action, is stable while what is defined in a given society or group as fair or prosocial can be malleable. Thus, a main difference between social preference theories of equity, reciprocity, guilt aversion, and self-image and the direct social norms approach is that these theories are concrete about the motivational basis of norm compliance and the motives are assumed to be stable across contexts whereas the direct social norms approach remains vague with respect to the motives underlying norm compliance.

For example, both conditionally cooperative behaviour and the willingness to punish free-riders in a public goods game can arise from a desire for fairness or reciprocity. In other words, inequity-averse subjects and reciprocity-motivated subjects are often conditional cooperators as well as punishers[59,64] and, therefore, these motives contribute to the explanation of all the major qualitative regularities mentioned above (except the existence of antisocial or perverse punishment, which we discuss below). Likewise, the communication effects (regularity 1) as well as the framing effects (regularity 5) can be explained by stable preferences for equity or reciprocity because these preferences imply conditionally cooperative behaviour such that if frames and pre-play communication renders expectations about others' cooperation more optimistic, subjects will cooperate more.

Or take, for example, regularity 4 that 'partners' generally cooperate more than 'strangers'. The theory of inequity aversion or reciprocity can explain this finding by the regularity that the existence of inequity-averse or reciprocal subjects generates incentives for selfish individuals in a partner treatment to invest into cooperation during the early periods of a finitely repeated game[18]. This investment is profitable because it maintains the cooperation of the inequity-averse or reciprocal subjects in future periods. However, this incentive is absent in a stranger treatment where all interactions are one-shot so that there are no future gains. Note that this theory also explains that in a partner treatment, cooperation declines over time but restarts again if subjects play another finitely repeated game[18]. And because the theories explain why people punish free-riders, they can account for the punishment-related regularities 6–8.

In summary, social preferences for fairness/equity, reciprocity or a prosocial self-image and the desire to avoid guilt are likely to play an important role in norm compliance. They provide an intrinsic motive to obey the normative standard to some extent and/or to sanction those who violate it. All of these theories are consistent with the notion that emotions are a key driver of the social preference although — with the exception of guilt-aversion theory, which models the emotion of guilt — they do not explicitly incorporate emotions in the model.

Although social preferences help in achieving norm compliance, it is important to distinguish them conceptually from social norms, which are defined as widely shared and approved normative standards. These standards are the essence of a social norm and they affect social preferences by defining what is considered as fair/equitable, kind or prosocial but they are conceptually nevertheless distinct. The direct norm approach is silent about the underlying motives that induce individuals to comply with a prevailing social norm and theoretical papers that apply this approach[54] often make ad hoc assumptions about the social norm while empirical studies do not define ex ante the content of the normative standard but instead measure the norm empirically[55,69]. This renders the direct norm approach more flexible and more difficult to refute unless it is possible to reliably identify the normative standard empirically over the relevant range of situations.

How can we identify social norms?

There are several methods for the identification of social norms[24,55,70–72]. One method builds on the premise that humans are willing to incur personal costs to sanction the violation of a norm even if they are not directly hurt by the violation. One reason for this willingness may be that norm violations have been shown to cause indignation or even outrage[23,73,74] and these emotions may provide the raw material for the willingness to punish. Another reason may be that norm violators are typically perceived to deserve punishment[75] and, therefore, sanctioning them provides satisfaction — a hypothesis that is consistent with the finding that reward-related brain areas are activated during the punishment of norm violators[76] and that already preschool children and chimpanzees are willing to pay for watching the punishment of antisocial actors[77].

Whatever the precise reason may be, if norm violations trigger the desire to punish the perpetrators, we have a potential tool for identifying the norm as part of those behaviours that are not punished by uninvolved third parties. Various studies have therefore employed a third-party punishment paradigm for the study of social norms[24,78–81]. In these experiments, third parties more readily punish those who free-ride against a cooperative partner compared with bilateral defectors or cooperators, providing evidence for a norm of conditional cooperation[24,82,83]. Survey studies confirm that participants judge defection against a cooperative partner more harshly than mutual defection[70].

An important method for the identification of social norms is based on the idea that social norms provide a focal point such that subjects' normative judgements are coordinated on this focal point[55]. This approach provides an incentivized measure of social norms by asking subjects to rate the extent to which an action is 'socially appropriate and consistent with moral or proper social behaviour'. Subjects are not asked to provide their own personal evaluation, but to indicate what they believe is the most common answer, and they earn a monetary reward if their rating coincides with the modal answer of others. This method has already been employed in several studies to elicit the social norm in social dilemmas[41,84] and in one of these studies[41] identifies a conditional cooperation norm in the public goods game.

One study[85] applied this method to measure whether the punishment of unfair proposers in the ultimatum game — by rejecting their offer — is a social norm. Interestingly, the study shows that this is clearly not the case. We conjecture that this is also likely to hold in social dilemma situations, suggesting that the desire to punish free-riders derives from other motives such as to avoid inequity[59,86] or to reciprocate to unfair actions[64,83].

Another method for the identification of social norms in social dilemma games has recently been presented in two papers[87,88]. Here, each subject of the group is asked to indicate what other group members should contribute to the public goods. The average of subjects' normative requests is afterwards conveyed to all group members and is likely to constitute a general normative standard of cooperation because it is commonly known and reflects the group members' views. Moreover, the higher subjects' agreement in their normative requests, the more the average request will constitute a legitimate normative standard[88]. One advantage of this method is that it can be easily implemented in every period of a public goods game such that the level and the strength of the norm can be identified continuously. Also, this method supports the existence of a conditional cooperation, that is, the average requested contribution in a period is declining in subjects' average actual contributions in the previous period. In addition, the data show that when direct targeted punishment of free-riders is possible, subjects strongly obey the average normative request in their actual cooperation choices[88].

Thus, taken together, there is ample and diverse evidence for the existence of a conditional cooperation norm in social dilemma situations while there is little or no evidence that punishment of free-riders constitutes a social norm. These results show that one can provide discipline to the direct social norms approach and they strengthen the conjecture that a conditional cooperation norm shapes human cooperation. However, these norm elicitation approaches do not yet prove that cooperation behaviour is causally affected by social norms because they — so far — only establish a correlation between the social norm and actual cooperative behaviour[41].

Do social norms causally affect cooperation behaviour?

The potential causal effect of social norms on behaviour has been studied in various ways. A prominent approach[89,90] assumes that social norms need to be activated, that is, become the focus of subjects' attention to affect behaviour. Based on this view, a causal effect of social norms can be identified by varying the salience of the norm with various priming techniques. This literature shows that when subjects' attention is shifted towards social norms they begin to act in a more norm-congruent way[89–93]. For example, in one study[90], car drivers, who did not know that they were part of an experiment, saw the following handbill on their windshield: "April is Keep Arizona Beautiful Month. Please Do Not Litter". In a second condition, the text on the handbill was "April is Conserve Arizona's Energy Month. Please Turn Off Unnecessary Lights" and in a third (control) condition they could read "April is Arizona's Fine Arts Month. Please Visit Your Local Art Museum". In line with the hypothesis that a stronger activation of the anti-littering norm leads to less littering, car drivers threw the handbill on the ground in only in 10% of the cases in the first treatment, and in 18% and 25% of the cases in the second and third conditions, respectively. Findings like these raise the question of which aspect of the social norm is the causal driver of the behaviour change. Does the increase in the salience of the norm change the social appropriateness rating of norm-compliant behaviour? Or does it merely change subjects' views about how widely the norm is shared? Or does it change subjects' feelings of guilt if they litter? Unfortunately, we do not know the answer to these questions.

The abovementioned method for norm identification[87,88] through individual normative requests can also be used to study the causal impact of social norms on behaviour. In treatments with normative requests, the average request constitutes a commonly known standard of behaviour that is absent in treatments without normative requests. In one study[88], the authors introduce the norm formation opportunity in finitely repeated public goods games where the possibility to punish other group members is either absent or present. Interestingly, when the possibility of punishment is absent, the opportunity to form a normative standard has no impact on behaviour while in the presence of the possibility to punish, the normative standard causes a significant and stable increase in cooperation rates (Fig. 3).

This radically different impact of social norms on cooperation when there are punishment opportunities exists despite the fact that the normative standard in the punishment and no-punishment treatment is very high and statistically indistinguishable during the first three periods. Nevertheless, substantial norm deviations occur in the absence of punishment from the very beginning while in the presence of punishment the norm is largely obeyed throughout the whole experiment. Thus, the existence of a normative standard that renders high cooperation the socially most appropriate action, and focuses attention on the normative standard, is per se not sufficient to induce a change in cooperation behaviour, suggesting that intrinsic motives for norm compliance are not sufficiently strong and that the punishment threat is needed to establish a stable norm-driven behaviour change in a population of heterogeneously motivated actors.

Normative constraints and peer punishment (in)efficiency

The existence of punishment opportunities in public goods games causes strong cooperation increases in many, but not in all,

cultures[31,94,95]. In particular, in those countries that have weak norms of civic cooperation — defined as the willingness to evade taxes, make fraudulent claims to receive welfare state benefits or dodging fares on public transport — the antisocial punishment of cooperators is particularly strong and is associated with detrimental effects on overall cooperation rates. This finding is consistent with the view that norms of civic cooperation have a causal, constraining effect on antisocial punishment. However, the finding does not prove causality because there could be other reasons that may account for the correlation between antisocial punishment and norms of civic cooperation. For example, countries with low norms of civic cooperation often also have bad schools (for example, because of teacher absenteeism or low teacher quality[96,97]) and school or teacher quality might shape both norms of civic cooperation and restraints on antisocial punishment.

Although the antisocial punishment of above-average cooperators by those who cooperate less tends to be rare in Western cultures, it has been observed from the beginning and several potential reasons for its existence have been mentioned[22]. First, in rare cases, it may simply reflect a random choice error. Second, there is evidence that a small, yet significant proportion of subjects regularly displays envious or spiteful motives[98,99], implying that they prefer to spend money to hurt others regardless of their level of prosociality. Third, antisocial punishment may be the result of a coordination failure among reciprocally motivated subjects that are in principle willing to cooperate. Consider a reciprocal subject with pessimistic beliefs about others' cooperation. These subjects may cautiously start with an intermediate or low level of cooperation while other subjects have optimistic expectations, start with high cooperation and punish those who cooperate less. The pessimistic, yet willing, low contributor may view this as an unfair punishment and may thus retaliate in the next period against the high contributors. These events may spoil the whole group and lead to a process of punishment and counter-punishment with detrimental effects on cooperation. In fact, if subjects are given explicit counter-punishment opportunities[30,32], some subjects use them to the detriment of the group's cooperation and welfare by punishing those who punished them for free-riding. More generally, public goods experiments that allow for peer punishment often fail to increase the overall welfare of the group members for an extended period of time despite the large increase in cooperation rates[22,25,38]. The reason for this is the high collateral cost associated with peer punishment.

However, the very fact that peer punishment can get out of control suggests that societies have developed mechanisms to constrain and control it. After all, peer punishment is physically always possible when two or more individuals directly interact with each other. It appears impossible for society to ever control or constrain all the different forms of peer punishment — that range from a raised eye brow or verbal insult to mobbing, ostracism, public shaming and corporal punishment — except through the normative control of people's behaviour. The literature on simple societies[100,101] provides ample evidence of the ways in which societies impose constraints on punishment. One study[101], for example, reports how the Ju/'hoansi bushmen, a group of hunter-gatherers living in Botswana, exert peer punishment according to strong habitual and normative constraints. For instance, if a man is publicly criticized for norm violations, this is often done by a women to avoid the escalation of arguments among men.

Rather than rely on peer-to-peer sanctioning, individuals will often prefer some type of institutional arrangement to regulate punishment by either ruling out peer punishment completely[102] or replacing it with a centralized state that automatically imposes taxes to finance public goods[103] or by an enforcement mechanism that rules out antisocial peer punishment[104–107]. But how is it possible to achieve this without also ruling out peer punishment altogether and more fundamentally, how is it ever possible to rule out peer punishment altogether in a world in which people socially interact with each other and in which the centralized legal enforcement of rules is always imperfect?

This question can be answered by comparing the punishment patterns in settings with and without the opportunity for normative requests[88]. It turns out that when subjects can form a normative cooperation standard, the punishment of free-riders becomes less severe. Thus, the normative standard increases cooperation while simultaneously decreasing the punishment of free-riders, suggesting that the punishment of free-riders becomes more effective. In fact, punished free-riders indeed increase their cooperation subsequently more strongly when the normative standard is present[88]. Antisocial punishment also decreases in the presence of a normative cooperation standard, thus lending support to the hypothesis that norms of civic cooperation may causally reduce antisocial punishment.

Despite the high potential collateral cost of normatively unconstrained peer punishment, it has been observed that participants will prefer this over a setting with no opportunities for targeted punishment (regularity 8). However, if subjects additionally can migrate to normatively coordinated peer punishment and normative coordination and punishment by a central authority, participants never enter the uncoordinated peer punishment setting. The institutions with normative coordination minimize or fully eradicate antisocial punishment and generate high levels of cooperation without the collateral damages associated with uncoordinated peer punishment[87]. This demonstrates that the traditional uncoordinated peer punishment institution fails to capture a very important dimension: the strong demand for normative coordination and regulation — a demand that societies who inevitably have to rely on some forms of peer sanctioning typically satisfy through the formation of social norms that put constraints on individuals' sanctioning behaviour. Of course, groups will not automatically solve inefficient peer sanctioning through informal constraints, but it seems likely that those groups who do solve this problem in a more efficient way will be more successful because they are better able to solve their collective action problems[1,87,108,109]. Therefore, they are better able to compete with other groups. Thus, conclusions regarding the effectiveness and the welfare properties of peer punishment may provide a misleading picture if they are based on institutional settings that rule out suitable normative consensus building opportunities that can put constraints on peer sanctioning.

Summary and open questions

The pervasiveness of social norms and the ubiquity of cooperation among non-kin are two salient features of human societies. Many social norms are beneficial for overall society and compliance with them can be viewed as acts of cooperation. Although humans are by no means the only species displaying cooperation among individuals, it has often been pointed out that the breadth and depth of human large-scale cooperation among non-kin in a globalized world, as well as the observed cooperation in one-shot encounters, appear unique in the animal kingdom[5,110–112]. Several potential factors — such as limited memory or excessive time discounting[111,113] — may constitute evolutionary obstacles to cooperation in animal species but perhaps the cognitive prerequisites for social norms are also relevant. For example, the very notion of a normative standard — what ought to be done — is rather complex and perhaps even impossible to identify reliably in species that lack sophisticated language. The same applies to the notion of normative approval and disapproval. Therefore, it is perhaps not surprising that our closest living relatives do not seem to share some of our most fundamental norms of fairness and cooperation[114–116] (although see refs[117,118]) and that there seems to be no evidence for third-party punishment of norm violations harming non-kin in non-human species[119]. In contrast, third-party punishment of non-kin and even strangers is

Box 1 | Important unsolved research problems

(1) What are micro-sociological and psychological processes that facilitate and hinder the development of a social norm?
(2) What is — at the conceptual level — the precise relationship between social preferences and social norms and how can we distinguish them empirically? How do social norms influence the motivational content of social preferences and, for given social preferences, how do they affect compliance with normative standards?
(3) What determines individuals' agreement with the 'ought component' of norms[88]? How do they come to internalize or reject a normative standard?
(4) What explains the formation and the decay of social norms and how can we explain changes in the normative content, that is, the 'ought component' of social norms[88]?
(5) What are the long-run environmental and economic determinants of social norms[125–128]? And how do normative standards evolve in the context of conflicting economic interests[71]?
(6) How do economic incentives, the human desire for social approval and normative standards interact? When are they complements and when do economic incentives undermine normative standards and approval incentives[129]?
(7) How does actual compliance and non-compliance shape the development of normative standards[88]?
(8) Through which interventions and public policies is it possible to shape social norms[72] and which aspect of the norm and norm-related behaviours — the content of the normative standard, social agreement with the normative standard, behavioural compliance with the standard — is changed by the intervention?
(9) How do legal institutions — apart from their sanctioning capacity — affect social norms and how do social norms affect the effectiveness of legal institutions[129,130]? To what extent do legal institutions shape normative standards by setting precedent, fall back rules or through expressing what is normatively approved and expected[131]?
(10) To what extent and in which ways do social norms influence important economic and social patterns[87,88,132–134]?

frequent in humans[24,78,120] and young children already have a working knowledge of social norms[116,121,122]. The widespread prevalence of social norms may therefore well be one of the defining characteristics of our species and a crucial determinant of human cooperation.

The evidence suggests that human cooperation is strongly affected by normative considerations. Various methods indicate the existence of a strong conditional cooperation norm. The behavioural strength of the conditional cooperation norm probably also derives from its relation to principles of equity and reciprocity. Compliance with social norms relies on the existence of social preferences that incorporate abstract normative principles such as equity or reciprocity — which also provide foundations for the willingness to punish norm violators — or are based on the desire for avoiding disapproval, a prosocial self-image or the avoidance of disappointing others. Social norms also appear to guide and constrain punishment behaviour and subjects have a strong desire for environments that enable normative coordination.

There are, however, still many important unanswered questions. Reliable empirical knowledge about the precise channels through which norms have a causal impact is, for example, still scarce. Does the normative standard shape behaviour directly via an intrinsic utility component or does it have an impact by affecting and coordinating beliefs about others' cooperation. Or does it guide the punishment of free-riders and affect beliefs about punishment in case of non-compliance? In addition, there are many other intriguing and exciting questions that are awaiting an answer (see Box 1 on important unresolved research problems), implying that there is still much to discover in this area of research.

Published online: 9 July 2018

References

1. Boyd, R. & Richerson, P. J. The evolution of norms — an anthropological view. *J. Inst. Theor. Econ.* **150**, 72–87 (1994).
2. Sober, E. & Wilson, D. S. *Unto Others: The Evolution and Psychology of Unselfish Behavior* (Harvard Univ. Press, Cambridge, MA, 1999).
3. Elster, J. *The Cement of Society: A Survey of Social Order* (Cambridge Univ. Press, Cambridge, 1989).
4. Hobbes, T. *Leviathan* (Continuum, New York, NY, 2005).
5. Fehr, E. & Fischbacher, U. The nature of human altruism. *Nature* **425**, 785–791 (2003).
6. Fehr, E. & Fischbacher, U. Social norms and human cooperation. *Trends Cogn. Sci.* **8**, 185–190 (2004).
7. Bicchieri, C. *The Grammar of Society: The Nature and Dynamics of Social Norms* (Cambridge Univ. Press, 2006).
8. Dawes, R. M., McTavish, J. & Shaklee, H. Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation. *J. Pers. Soc. Psychol.* **35**, 1–11 (1977).
9. Dawes, R. M. Social dilemmas. *Ann. Rev. Psychol.* **31**, 169–193 (1980).
10. Isaac, R. M. & Walker, J. M. Communication and free-riding behavior: the voluntary contribution mechanism. *Econ. Inq.* **26**, 585–608 (1988).
11. Sally, D. Conversation and cooperation in social dilemmas: a meta-analysis of experiments from 1958 to 1992. *Ration. Soc.* **7**, 58–92 (1995).
12. Fischbacher, U., Gächter, S. & Fehr, E. Are people conditionally cooperative? Evidence from a public goods experiment. *Econ. Lett.* **71**, 397–404 (2001).
13. Kocher, M. G., Cherry, T., Kroll, S., Netzer, R. J. & Sutter, M. Conditional cooperation on three continents. *Econ. Lett.* **101**, 175–178 (2008).
14. Chaudhuri, A. Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Exp. Econ.* **14**, 47–83 (2011).
15. Isaac, M. R., McCue, K. & Plott, C. R. Public goods provision in an experimental environment. *J. Public Econ.* **26**, 51–74 (1985).
16. Kim, O. & Walker, J. M. The free rider problem: experimental evidence. *Public Choice* **43**, 3–24 (1984).
17. Andreoni, J. Warm-glow versus cold-prickle: the effects of positive and negative framing on cooperation in experiments. *Quart. J. Econ.* **110**, 1–21 (1995).
18. Ambrus, A. & Pathak, P. A. Cooperation over finite horizons: a theory and experiments. *J. Public Econ.* **95**, 500–512 (2011).
19. Croson, R. Partners and strangers revisited. *Econ. Lett.* **53**, 25–32 (1996).
20. Ellingsen, T., Johannesson, M., Mollerstrom, J. & Munkhammar, S. Social framing effects: preferences or beliefs?. *Games Econ. Behav.* **76**, 117–130 (2012).
21. Liberman, V., Samuels, S. M. & Ross, L. The name of the game: predictive power of reputations versus situational labels in determining prisoner's dilemma game moves. *Pers. Soc. Psychol. B* **30**, 1175–1185 (2004).
22. Fehr, E. & Gächter, S. Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **90**, 980–994 (2000).
23. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137–140 (2002).
24. Fehr, E. & Fischbacher, U. Third-party punishment and social norms. *Evol. Hum. Behav.* **25**, 63–87 (2004).
25. Gächter, S., Renner, E. & Sefton, M. The long-run benefits of punishment. *Science* **322**, 1510–1510 (2008).
26. Fehr, E. & Rockenbach, B. Detrimental effects of sanctions on human altruism. *Nature* **422**, 137–140 (2003).
27. Houser, D., Xiao, E., McCabe, K. & Smith, V. When punishment fails: research on sanctions, intentions and non-cooperation. *Games Econ. Behav.* **62**, 509–532 (2008).
28. Xiao, E. T. Profit-seeking punishment corrupts norm obedience. *Games Econ. Behav.* **77**, 321–344 (2013).
29. Fehr, E. & List, J. A. The hidden costs and returns of incentives-trust and trustworthiness among CEOs. *J. Eur. Econ. Assoc.* **2**, 743–771 (2004).
30. Cinyabuguma, M., Page, T. & Putterman, L. Can second-order punishment deter perverse punishment?. *Exp. Econ.* **9**, 265–279 (2006).
31. Herrmann, B., Thöni, C. & Gächter, S. Antisocial punishment across societies. *Science* **319**, 1362–1367 (2008).
32. Nikiforakis, N. Punishment and counter-punishment in public good games: Can we really govern ourselves?. *J. Public Econ.* **92**, 91–112 (2008).
33. Gürerk, Ö., Irlenbusch, B. & Rockenbach, B. The competitive advantage of sanctioning institutions. *Science* **312**, 108–111 (2006).

34. Gürerk, Ö., Irlenbusch, B. & Rockenbach, B. On cooperation in open communities. *J. Public Econ.* **120**, 220–230 (2014).
35. Brown, M., Falk, A. & Fehr, E. Relational contracts and the nature of market interactions. *Econometrica* **72**, 747–780 (2004).
36. Rockenbach, B. & Milinski, M. The efficient interaction of indirect reciprocity and costly punishment. *Nature* **444**, 718–723 (2006).
37. Sefton, M., Shupp, R. & Walker, J. M. The effect of rewards and sanctions in provision of public goods. *Econ. Inq.* **45**, 671–690 (2007).
38. Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D. & Nowak, M. A. Positive interactions promote public cooperation. *Science* **325**, 1272–1275 (2009).
39. Balliet, D., Mulder, L. B. & Van Lange, P. A. M. Reward, punishment, and cooperation: a meta-analysis. *Psychol. Bull.* **137**, 594–615 2011).
40. Gächter, S. & Thöni, C. Social learning and voluntary cooperation among like-minded people. *J. Eur. Econ. Assoc.* **3**, 303–314 (2005).
41. Kimbrough, E. O. & Vostroknutov, A. Norms make preferences social. *J. Eur. Econ. Assoc.* **14**, 608–638 (2016).
42. Chaudhuri, A., Graziano, S. & Maitra, P. Social learning and norms in a public goods experiment with inter-generational advice. *Rev. Econ. Stud.* **73**, 357–380 (2006).
43. Barr, A., Packard, T. Serra, D. Participatory accountability and collective action: experimental evidence from Albania. *Eur. Econ. Rev.* **68**, 250–269 (2014).
44. Fehr, E. & Leibbrandt, A. A. A field study on cooperativeness and impatience in the tragedy of the commons. *J. Public Econ.* **95**, 1144–1155 (2011).
45. Rustagi, D., Engel, S. & Kosfeld, M. Conditional cooperation and costly monitoring explain success in forest commons management. *Science* **330**, 961–965 (2010).
46. Keizer, K., Lindenberg, S. & Steg, L. The spreading of disorder. *Science* **322**, 1681–1685 (2008).
47. Kosfeld, M. & Rustagi, D. Leader punishment and cooperation in groups: experimental field evidence from commons management in Ethiopia. *Am. Econ. Rev.* **105**, 747–783 (2015).
48. Breza, E., Kaur, S. & Krishnaswamy, N. *Scabs: Norm-driven Suppression of Labor Supply* Working Paper (2018).
49. Kaur, S. Nominal wage rigidity in village labor markets. *Am. Econ. Rev.* (in the press).
50. Gelcich, S., Guzman, R., Rodríguez-Sickert, C., Castilla, J. C. & Cárdenas, J. C. Exploring external validity of common pool resource experiments: insights from artisanal benthic fisheries in Chile. *Ecol. Soc.* **18**, 2 (2013).
51. Burks, S. et al. *Lab Measures of Other-regarding Preferences can Predict some Related On-the-job Behavior: Evidence from a Large Scale Field Experiment* IZA Discussion Paper No. 9767 (SSRN, 2016).
52. Carlsson, F., Johansson-Stenman, O. & Nam, P. K. Social preferences are stable over long periods of time. *J. Public Econ.* **117**, 104–114 (2014).
53. Ostrom, E. Collective action and the evolution of social norms. *J. Econ. Perspect.* **14**, 137–158 (2000).
54. Lindbeck, A., Nyberg, S. & Weibull, J. W. Social norms and economic incentives in the welfare state. *Quart. J. Econ.* **114**, 1–35 (1999).
55. Krupka, E. L. & Weber, R. A. Identifying social norms using coordination games: why does dictator game sharing vary?. *J. Eur. Econ. Assoc.* **11**, 495–524 (2013).
56. Bicchieri, C. Covenants without swords: group identity, norms, and communication in social dilemmas. *Ration. Soc.* **14**, 192–228 (2002).
57. Fischbacher, U. & Gächter, S. Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *Am. Econ. Rev.* **100**, 541–556 (2010).
58. Wrong, D. H. The oversocialized conception of man in modern sociology. *Am. Sociol. Rev.* **26**, 183–193 (1961).
59. Fehr, E. & Schmidt, K. M. A theory of fairness, competition, and cooperation. *Quart. J. Econ.* **114**, 817–868 (1999).
60. Bolton, G. E. & Ockenfels, A. ERC: a theory of equity, reciprocity, and competition. *Am. Econ. Rev.* **90**, 166–193 (2000).
61. Lopez-Perez, R. Aversion to norm-breaking: a model. *Games Econ. Behav.* **64**, 237–267 (2008).
62. Rabin, M. Incorporating fairness into game theory and economics. *Am. Econ. Rev.* **83**, 1281–1302 (1993).
63. Dufwenberg, M. & Kirchsteiger, G. A theory of sequential reciprocity. *Games Econ. Behav.* **47**, 268–298 (2004).
64. Falk, A. & Fischbacher, U. A theory of reciprocity. *Games Econ. Behav.* **54**, 293–315 (2006).
65. Battigalli, P. & Dufwenberg, M. Guilt in games. *Am. Econ. Rev.* **97**, 170–176 (2007).
66. Dufwenberg, M., Gachter, S. & Hennig-Schmidt, H. The framing of games and the psychology of play. *Games Econ. Behav.* **73**, 459–478 (2011).
67. Dhami, S., Wei, M. & Al-Nowaihi, A. Public goods games and psychological utility: theory and evidence. *J. Econ. Behav. Organ.* (in the press).
68. Benabou, R. & Tirole, J. Identity, morals, and taboos: beliefs as assets. *Quart. J. Econ.* **126**, 805–855 (2011).
69. Krupka, E. L., Leider, S. & Jiang, M. A. A meeting of the minds: informal agreements and social norms. *Manag. Sci.* **63**, 1708–1729 (2016).
70. Cubitt, R. P., Drouvelis, M., Gächter, S. & Kabalin, R. Moral judgments in social dilemmas: How bad is free riding?. *J. Public Econ.* **95**, 253–264 (2011).
71. Reuben, E. & Riedl, A. Enforcement of contribution norms in public good games with heterogeneous populations. *Games Econ. Behav.* **77**, 122–137 (2013).
72. Bicchieri, C. *Norms in the Wild* (Oxford Univ. Press, Oxford, 2017).
73. Xiao, E. & Houser, D. Emotion expression in human punishment behavior. *Proc. Natl. Acad. Sci. USA* **102**, 7398–7401 (2005).
74. Bosman, R., Sutter, M. & van Winden, F. The impact of real effort and emotions in the power-to-take game. *J. Econ. Psych.* **26**, 407–429 (2005).
75. Carlsmith, K. M., Darley, J. M. & Robinson, P. H. Why do we punish? Deterrence and just deserts as motives for punishment. *J. Pers. Soc. Psychol.* **83**, 284–299 (2002).
76. DeQuervain, D. et al. The neural basis of altruistic punishment. *Science* **305**, 1254–1258 (2004).
77. Mendes, N., Steinbeis, N., Bueno-Guerra, N., Call, J. & Singer, T. Preschool children and chimpanzees incur costs to watch punishment of antisocial others. *Nat. Hum. Behav.* **2**, 45–51 (2018).
78. Henrich, J. et al. Costly punishment across human societies. *Science* **312**, 1767–1770 (2006).
79. Marlowe, F. W. et al. More 'altruistic' punishment in larger societies. *Proc. R. Soc. B* **275**, 587–592 (2008).
80. Lewisch, P. G., Ottone, S. & Ponzano, F. Free-riding on altruistic punishment? An experimental comparison of third-party punishment in a stand-alone and in an in-group environment. *Rev. Law. Econ.* **7**, 161–190 (2011).
81. Lergetporer, P., Angerer, S., Glätzle-Rützler, D. & Sutter, M. Third-party punishment increases cooperation in children through (misaligned) expectations and conditional cooperation. *Proc. Natl. Acad. Sci. USA* **111**, 6916–6921 (2014).
82. Kamei, K. *Altruistic Norm Enforcement and Decision-making Format in a Dilemma: Experimental Evidence* Working Paper (SSRN, 2017).
83. Carpenter, J. P. & Matthews, P. H. Norm enforcement: anger, indignation, or reciprocity?. *J. Eur. Econ. Assoc.* **10**, 555–572 (2012).
84. Gächter, S., Nosenzo, D. & Sefton, M. Peer effects in pro-social behavior: social norms or social preferences?. *J. Eur. Econ. Assoc.* **11**, 548–573 (2013).
85. Bartling, B. & Özdemir, Y. *The Limits to Moral Erosion in Markets: Social Norms and the Replacement Excuse* Working Paper Series ISSN ISSN 1664-705X, No. 263 (Department of Economics, University of Zurich, 2017).
86. Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R. & Smirnov, O. Egalitarian motives in humans. *Nature* **446**, 794–796 (2007).
87. Fehr, E. & Williams, T. *Social Norms, Eendogenous Sorting and the Culture of Cooperation* Working Paper (Department of Economics, University of Zurich, 2018); www.econ.uzh.ch/static/wp/econwp267.pdf
88. Fehr, E. & Schurtenberger, I. *The Dynamics of Norm Formation and Norm Decay* Working Paper (Department of Economics, University of Zurich, 2018).
89. Cialdini, R. B., Kallgren, C. A. & Reno, R. R. A focus theory of normative conduct: a theoretical refinement and reevaluation of the role of norms in human behavior. *Adv. Exp. Soc. Psychol.* **24**, 201–234 (1991).
90. Kallgren, C. A., Reno, R. R. & Cialdini, R. B. A focus theory of normative conduct: when norms do and do not affect behavior. *Pers. Soc. Psychol. B.* **26**, 1002–1012 (2000).
91. Berkowitz, L. & Daniels, L. R. Affecting the salience of the social responsibility norm: effects of past help on the response to dependency relationships. *J. Abnorm. Soc. Psychol.* **68**, 275–281 (1964).
92. Berkowitz, L. Social norms, feelings, and other factors affecting helping and altruism. *Adv. Exp. Soc. Psychol.* **6**, 63–108 (1972).
93. Hallsworth, M., List, J. A., Metcalfe, R. D. & Vlaev, I. The behavioralist as tax collector: using natural field experiments to enhance tax compliance. *J. Public Econ.* **148**, 14–31 (2017).
94. Gächter, S. & Herrmann, B. Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philos. Trans. R. Soc. Lond. Ser. B* **364**, 791–806 (2009).
95. Gächter, S. & Herrmann, B. The limits of self-governance when cooperators get punished: experimental evidence from urban and rural Russia. *Eur. Econ. Rev.* **55**, 193–210 (2011).
96. Hanushek, E. A. & Woessmann, L. Knowledge capital, growth, and the East Asian miracle access to schools achieves only so much if quality is poor. *Science* **351**, 344–345 (2016).
97. Hanushek, E. A., & Rivkin, S. G. The distribution of teacher quality and implications for policy. *Ann. Rev. Econ.* **4**, 131–158 (2012).
98. Fehr, E., Hoff, K. & Kshetramade, M. Spite and development. *Am. Econ. Rev.* **98**, 494–499 (2008).
99. Bruhin, A., Fehr, E. & Schunk, D. The many faces of human prosociality. *J. Eur. Econ. Assoc.* (in the press).

100. Mathew, S. & Boyd, R. Punishment sustains large-scale cooperation in prestate warfare. *Proc. Natl. Acad. Sci. USA* **108**, 11375–11380 (2011).
101. Wiessner, P. Norm enforcement among the Ju/'hoansi Bushmen — a case of strong reciprocity?. *Hum. Nat.* **16**, 115–145 (2005).
102. Sutter, M., Haigner, S. & Kocher, M. G. Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations. *Rev. Econ. Stud.* **77**, 1540–1566 (2010).
103. Markussen, T., Putterman, L. & Tyran, J. R. Self-organization for collective action: an experimental study of voting on sanction regimes. *Rev. Econ. Stud.* **81**, 301–324 (2014).
104. Yamagishi, T. The provision of a sanctioning system as a public good. *J. Pers. Soc. Psychol.* **51**, 110–116 (1986).
105. Ertan, A., Page, T. & Putterman, L. Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *Eur. Econ. Rev.* **53**, 495–511 (2009).
106. Traulsen, A., Röhl, T. & Milinski, M. An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proc. R. Soc. B* https://doi.org/10.1098/rspb.2012.0937 (2012).
107. Andreoni, J. & Gee, L. K. Gun for hire: delegated enforcement and peer punishment in public goods provision. *J. Public Econ.* **96**, 1036–1046 (2012).
108. Boyd, R. & Richerson, P. J. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* **13**, 171–195 (1992).
109. Henrich, J. Cultural group selection, coevolutionary processes and large-scale cooperation. *J. Econ. Behav. Organ.* **53**, 3–35 (2004).
110. Hammerstein, P. *Genetic and Cultural Evolution of Cooperation* (MIT Press, Cambridge, MA, 2003).
111. Stevens, J. R. & Hauser, M. D. Why be nice? psychological constraints on the evolution of cooperation. *Trends Cogn. Sci.* **8**, 60–65 (2004).
112. Boyd, R. & Richerson, P. in *Evolution and Culture* (eds Levinson, S. & Jaisson, P.) 105–132 (MIT Press, Cambridge, MA, 2006).
113. Stephens, D. W., McLinn, C. M. & Stevens, J. R. Discounting and reciprocity in an iterated prisoner's dilemma. *Science* **298**, 2216–2218 (2002).
114. Jensen, K., Call, J. & Tomasello, M. Chimpanzees are rational maximizers in an ultimatum game. *Science* **318**, 107–109 (2007).
115. Jensen, K., Call, J. & Tomasello, M. Chimpanzees are vengeful but not spiteful. *Proc. Natl Acad. Sci. USA* **104**, 13046–13050 (2007).
116. Ulber, J., Hamann, K. & Tomasello, M. Young children, but not chimpanzees, are averse to disadvantageous and advantageous inequities. *J. Exp. Child Psychol.* **155**, 48–66 (2017).
117. Proctor, D., Williamson, R. A., Waal, F. B. M. & Brosnan, S. F. Chimpanzees play the ultimatum game. *Proc. Natl Acad. Sci. USA* **110**, 2070–2075 (2013).
118. Brosnan, S. F., Schiff, H. C. & De Waal, F. B. Tolerance for inequity may increase with social closeness in chimpanzees. *Proc. R. Soc. B* **272**, 253–258 (2005).
119. Riedl, K., Jensen, K., Call, J. & Tomasello, M. No third-party punishment in chimpanzees. *Proc. Natl. Acad. Sci. USA* **109**, 14824–14829 (2012).
120. Jordan, J. J., Hoffman, M., Bloom, P. & Rand, D. G. Third-party punishment as a costly signal of trustworthiness. *Nature* **530**, 473–476 (2016).
121. McAuliffe, K., Jordan, J. J. & Warneken, F. Costly third-party punishment in young children. *Cognition* **134**, 1–10 (2015).
122. Cummins, D. D. Evidence of deontic reasoning in 3-and 4-year-old children. *Mem. Cogn.* **24**, 823–829 (1996).
123. Isaac, M. R., Walker, J. M. & Thomas, S. H. Divergent evidence on free riding: an experimental examination of some possible explanations. *Public Choice* **43**, 113–149 (1984).
124. Andreoni, J. Why free ride? strategies and learning in public goods experiments. *J. Public Econ.* **37**, 291–304 (1988).
125. Henrich, J. et al. Markets, religion, community size, and the evolution of fairness and punishment. *Science* **327**, 1480–1484 (2010).
126. Alesina, A., Giuliano, P. & Nunn, N. On the origins of gender roles: women and the plough. *Quart. J. Econ.* **128**, 469–530 (2013).
127. Ellickson, R. C. in *Social Norms* (eds Hechter, M. & Opp, K. D.) 35–75 (Russell Sage Foundation, New York, NY, 2001).
128. Lowes, S., Nunn, N., Robinson, J. A. & Weigel, J. L. The evolution of culture and institutions: evidence from the Kuba Kingdom. *Econometrica* **85**, 1065–1091 (2017).
129. Benabou, R. & Tirole, J. *Laws and Norms* Working Paper No. 17579 (NBER, 2011).
130. Posner, E. A. *Law and Social Norms* (Harvard Univ. Press, Cambridge, MA, 2000).
131. Sunstein, C. R. On the expressive function of law. *Univ. PA Law Rev.* **144**, 2021–2053 (1996).
132. Akerlof, G. A. The missing motivation in macroeconomics. *Am. Econ. Rev.* **97**, 5–36 (2007).
133. Allcott, H. Social norms and energy conservation. *J. Public Econ.* **95**, 1082–1095 (2011).
134. Nolan, J. M., Schultz, P. W., Cialdini, R. B., Goldstein, N. J. & Griskevicius, V. Normative social influence is underdetected. *Pers. Soc. Psychol. B* **34**, 913–923 (2008).

Acknowledgements

E.F. acknowledges support from the European Researcher Council (advanced ERC grant on the Foundations of Economic Preferences).

Author contributions

E.F. and I.S. contributed equally to all parts of the research and writing.

Competing interests

The authors declare no competing interests.

Additional information

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence should be addressed to E.F.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.