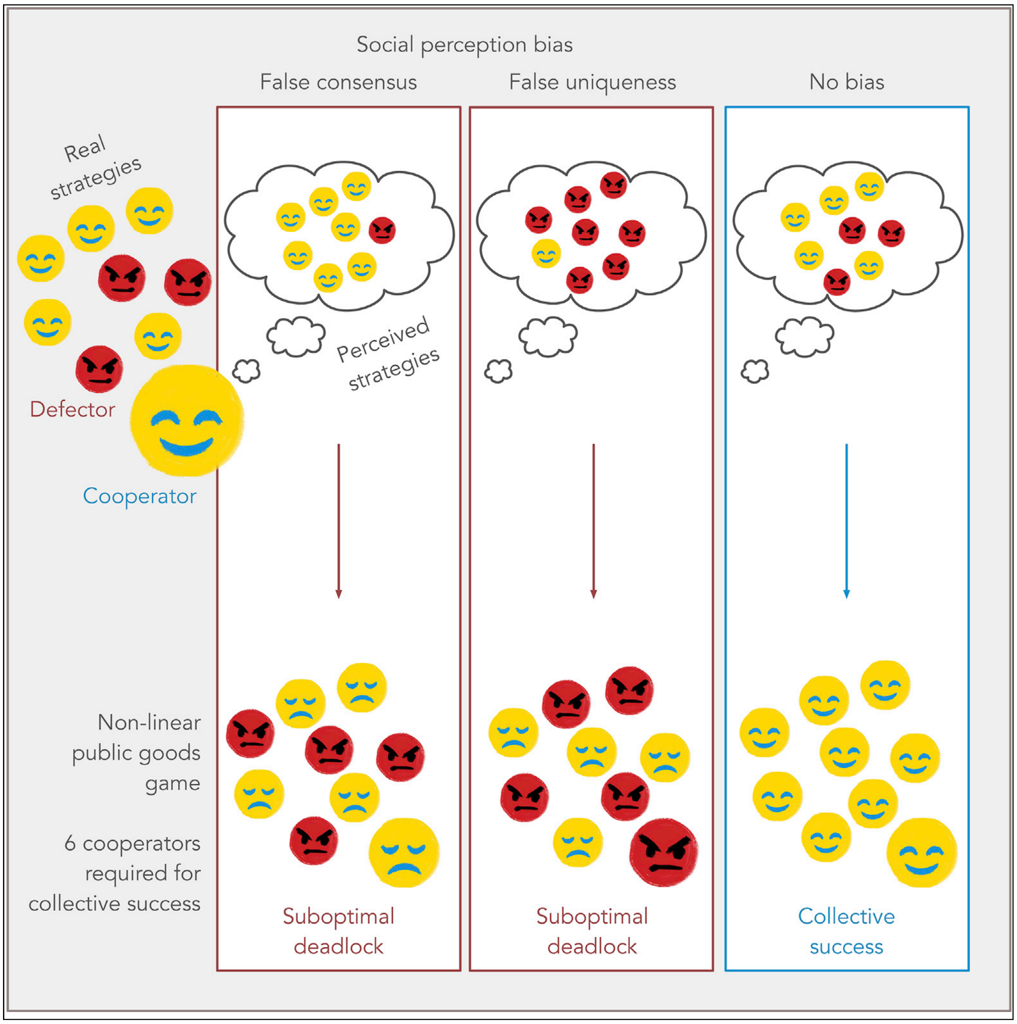
iScience

CellPress OPEN ACCESS

Article

Biased perceptions explain collective action deadlocks and suggest new mechanisms to prompt cooperation



Fernando P. Santos, Simon A. Levin, Vítor V. Vasconcelos

fppdsantos@gmail.com (F.P.S.)
v.v.vasconcelos@uva.nl (V.V.V.)

Highlights

Individuals often misperceive the real cooperation levels in a population

We model the impact of such biases in non-linear public goods games dynamics

False uniqueness and false consensus can lock groups in suboptimal states

Addressing perception biases can be more effective than typical monetary incentives

Santos et al., iScience *24*, 102375
April 23, 2021 © 2021 The Authors.
https://doi.org/10.1016/j.isci.2021.102375

CellPress
OPEN ACCESS

Article

Biased perceptions explain collective action deadlocks and suggest new mechanisms to prompt cooperation

Fernando P. Santos,[1,2,3,10,*] Simon A. Levin,[1,2,4,5,6] and Vítor V. Vasconcelos[2,3,7,6,8,9,*]

SUMMARY

When individuals face collective action problems, their expectations about others' willingness to contribute affect their motivation to cooperate. Individuals, however, often misperceive the cooperation levels in a population. In the context of climate action, people underestimate the pro-climate positions of others. Designing incentives to enable cooperation and a sustainable future must thereby consider how social perception biases affect collective action. We propose a theoretical model and investigate the effect of social perception bias in non-linear public goods games. We show that different types of bias play a distinct role in cooperation dynamics. False uniqueness (underestimating own views) and false consensus (overestimating own views) both explain why communities get locked in suboptimal states. Such dynamics also impact the effectiveness of typical monetary incentives, such as fees. Our work contributes to understanding how targeting biases, e.g., by changing the information available to individuals, can comprise a fundamental mechanism to prompt collective action.

INTRODUCTION

Many of the most pressing problems humanity faces today share the perils of public goods dilemmas (Dietz et al., 2003; Olson, 1965). These are dilemmas in which reaching a minimum level of cooperation is necessary to achieve the best social outcome, but in which refusing to do so (free-riding) is the immediate rational action to follow. Greenhouse gas emissions, overexploitation of natural resources, low vaccination coverage, antibiotics abuse, or fertilizer overuse are challenges in which incentivizing cooperation is arduous yet necessary to obtain results that benefit all (Dietz et al., 2003; Keohane and Victor, 2016; Levin, 1999; Smith et al., 2005). Failing to do so leads to the infamous *tragedy of the commons* (Hardin, 1968), engendering ecological breakdown and increased inequality, resource depletion, failure to achieve herd immunity, antimicrobial resistance, or groundwater contamination. Averting those scenarios requires judiciously designing incentives, interventions, and institutions.

Cooperation in public goods games is constrained not only by the costs and benefits involved but also by the social environment wherein the interactions take place. Experiments in the laboratory (Fischbacher et al., 2001) and the field (Frey and Meier, 2004) reveal that "those who believe others will cooperate in social dilemmas are more likely to cooperate themselves (Ostrom, 2000)." Elinor Ostrom identifies this as one of the seven stylized facts about public goods games—results replicated so frequently that they can be considered core facts. In fact, this finding has accompanied public goods games since the very first experiments with this interaction paradigm, which already indicate that assumptions about others' behavior impact the decision to cooperate (Dawes et al., 1977). Recent research reinforces this idea, revealing that second-order beliefs (i.e., beliefs about others' beliefs) are good predictors of one's own behavior (Jachimowicz et al., 2018). This observation underscores the potential effectiveness of norm-based interventions whereby informing individuals about the cooperative actions of others constitutes a trigger for cooperation (Bicchieri, 2016; Carattini et al., 2019; Miller and Prentice, 2016; Nyborg et al., 2016).

Although there is a link between cooperation and beliefs about others cooperating, humans reveal social perception biases, e.g., systematic errors in estimating the distribution of cooperative behaviors in a population. In a paradigmatic example, Monin and Norton report that, in a field study during a water shortage crisis in which students were asked to reduce the number of showers to save water, individuals

[1] Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA

[2] Center for BioComplexity, High Meadows Environmental Institute, Princeton University, Princeton, NJ 08544, USA

[3] Informatics Institute, University of Amsterdam, Science Park 904, 1098XH Amsterdam, The Netherlands

[4] Resources for the Future, Washington, DC, USA

[5] Beijer Institute of Ecological Economics, Stockholm, Sweden

[6] Andlinger Center for Energy and the Environment, Princeton University, Princeton, NJ 08544, USA

[7] Institute for Advanced Study, University of Amsterdam, 1012 GC Amsterdam, The Netherlands

[8] Centre for Urban Mental Health, University of Amsterdam, Amsterdam, The Netherlands

[9] Princeton Institute for International and Regional Studies, Princeton University, Princeton, NJ 08544, USA

[10] Lead contact

*Correspondence:
fppdsantos@gmail.com
(F.P.S.),
v.v.vasconcelos@uva.nl
(V.V.V.)

https://doi.org/10.1016/j.isci.2021.102375

systematically failed to estimate the prosocial behavior of others (Monin and Norton, 2003). Limiting water usage (reducing the number of showers) has all the ingredients of cooperation, whereas refusing to do so implies defecting on the public good. Survey results show that students concurred in false consensus, uniqueness bias, pluralistic ignorance, and other typical social perception biases. Beyond local public goods, the existence of perception bias extends to climate change beliefs. Research has shown that both the mass public and political elites—in China, the United States, and Germany—tend to underestimate the pro-climate positions of others (Mildenberger and Tingley, 2019; Taddicken et al., 2019). Likewise, Leviston et al. investigate the existence of pluralistic ignorance and false consensus effects regarding climate change beliefs in Australia, finding that opinions are subject to strong false consensus; in general, people underestimate the number of others who agree with the existence of climate change (Leviston et al., 2013). Although those opinions do not directly translate into cooperation or defection behaviors, they can be thought of as a proxy for engaging (or not) in climate action. The existence of such social perception biases was recently pointed out as an impediment to discussions about climate change (Geiger and Swim, 2016)—leading to the so-called spiral of silence (Noelle-Neumann, 1974)—being one possible reason for inhibition to take part in collective climate action (Kjeldahl and Hendricks, 2018). All the biases mentioned have for long been known in social psychology: Pluralistic ignorance is known as a situation in which people erroneously believe that their private opinions or behaviors are different from everybody else's (Miller and McFarland, 1987; Prentice and Miller, 1993)—which corresponds to false uniqueness or uniqueness bias when actions map with personal injunctive norms (Goethals et al., 1991; Suls and Wan, 1987); False consensus is known as the tendency to overestimate the representativeness of one's opinion or behavior in a population (Ross et al., 1977). Given the above-mentioned connection between cooperation in public goods dilemmas and beliefs about others' cooperative behavior, it is likely that such biases play an influential role in collective action itself.

The effect of perception biases is likely to be exacerbated in non-linear public goods games, in which collective action cannot be decomposed into pairwise interactions. A prototypical example is that of threshold public goods games, where the benefits of cooperation are not realizable until a certain fraction of cooperators exists (e.g., the advantages of reducing carbon emissions only ensue once a certain fraction of countries or industries do so) (Milinski et al., 2008; Pacheco et al., 2009; Santos and Pacheco, 2011; Tavoni et al., 2011). Threshold formulations for interactions typically lead to tipping points, characteristic of social behavior influenced by social norm change and expected to play a critical role in transitions to sustainability, e.g., mass adoption of sustainable technologies, implementation of collective insurance and risk-mitigation strategies (Santos et al., 2021), or changes in diets (Nyborg et al., 2016). Cooperation might, in this case, be hampered by failing to estimate accurately the number of individuals willing to cooperate, either by overestimating their real number (''*there are so many cooperators, I do not need to cooperate''*) or underestimating it (''*there are too few cooperators, it is not worth it for me to cooperate''*). Likewise, biases may create the illusion that the required number of cooperators is closer to the goal than it is, thus motivating cooperation. Importantly, these (incorrect) expectations about others can persist even after repeated interactions (Ackermann and Murphy, 2019). Therefore, it is fundamental to (1) understand the role of social perception bias in the dynamics of (non-linear) public goods games and (2) understand how to design cooperation incentives and interventions in situations where perception bias is prevalent.

We provide a theoretical model to analyze the effect of perception bias in public goods cooperation dynamics. We consider a population of (boundedly) rational individuals who adapt their behavior through a (smooth) best response (Fudenberg et al., 1998) while possibly incurring perception bias—either under- or overestimating the overall levels of cooperation. As detailed below (see transparent methods, supplemental information), we assume a population in which each individual can either adopt strategy C (cooperate) or D (defect). Interacting groups are formed randomly. Each cooperator pays a cost $c > 0$, and, when there are more than a threshold number, M, of cooperators, everyone gets a benefit, $b > c$, plus an enhanced share of the contributions of cooperators. We focus the analysis in situations where the enhancement, f, is such that there is an individual incentive to cooperate above the threshold ($f > 1$), and both full cooperation and full defection are Nash equilibria—with full cooperation being the social optimum. Above the threshold, cooperation is self-enforceable (Keohane and Victor, 2016), yet it is potentially hard to trigger in the first place, when below the threshold. This regime allows us to focus on the simpler situation in which collective action dynamics, in the absence of bias, are characterized by a single coordination barrier (see the supplemental information for further exploration of the parameters, where we show that the effects of biases discussed in the main text extend to other types of collective action dilemmas). For a given

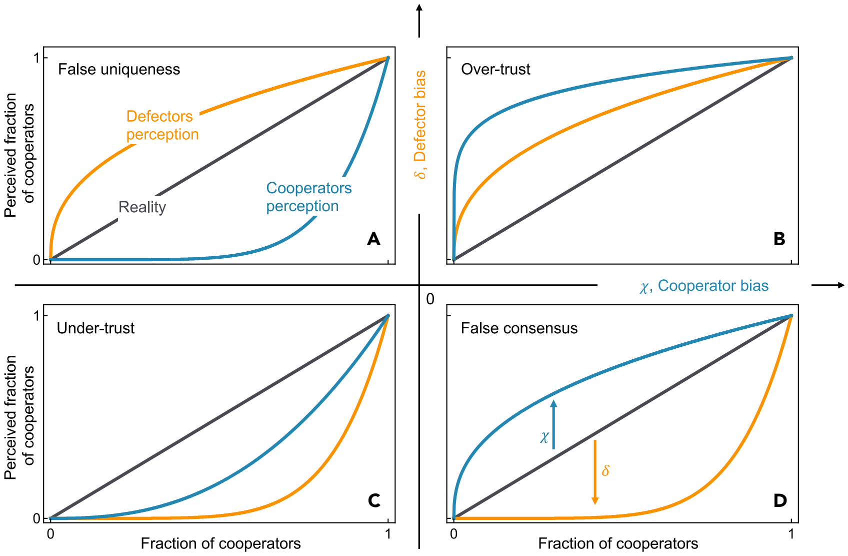


Figure 1. Individual perception biases toward cooperation

Individuals can be affected by different biases, depending on their behavior. Cooperators can perceive a higher or lower fraction of cooperators than in reality, and so do defectors. This creates the four different scenarios represented.

(A) ''False uniqueness'' corresponds to a case in which both cooperators and defectors believe their representation in the population is a smaller fraction than it is.

(B) ''Over-trust'' reflects biases where all individuals believe there are more cooperators than there are.

(C) ''Under-trust'' reflects a belief that there are fewer cooperators than there are.

(D) In the ''false consensus'' scenario, cooperators and defectors believe their representation is broader than it is.

configuration of the population, individuals will adapt by selecting the strategy maximizing their payoffs, given an estimate of the current distribution of strategies. Such estimates can be biased. As Figure 1 conveys, all the perception biases we consider here can be situated in a two-dimensional space (χ,δ), defined by a bias in the level of cooperators by cooperators (χ, where $\chi < 0$ implies an underestimation and $\chi > 0$ an overestimation of the number of other cooperators in the population) and a bias in the level of cooperators by defectors (δ, where $\delta < 0$ implies an underestimation and $\delta > 0$ an overestimation in the number of cooperators by defectors). Within this space, we can identify four distinct types of social perception biases: (1) ''False uniqueness'' ($\delta > 0$, $\chi < 0$), in which both cooperators and defectors believe their representation in the population is a smaller fraction than it is (we include a note on this definition of false uniqueness in supplemental information); (2) ''Over-trust'' ($\delta > 0$, $\chi > 0$), which reflects biases where all individuals believe there are more cooperators than there is; (3) ''Under-trust'' ($\delta < 0$, $\chi < 0$), which reflects a belief that there is less cooperation than there is; and (4) ''False consensus''' ($\delta < 0$, $\chi > 0$), whereby both cooperators and defectors believe their representation is broader than it is.

RESULTS

The aforementioned biases have substantial impacts on the dynamics of cooperation. We first focus on the role of homodirectional biases, affecting cooperators and defectors alike (over-trust and under-trust, Figures 1B and 1C), and then move to heterodirectional biases, which affect cooperators and defectors in opposite ways (false uniqueness and false consensus, Figures 1A and 1D).

Under-trust and over-trust impact the likelihood to reach optimal coordination

In the game considered here, and detailed above, collective benefits are distributed—and cooperation becomes desirable both for the group and the individuals—when a minimum fraction of cooperators exist in a population. In Figure 2, we control δ and χ such that we navigate from a scenario of under-trust (Figure 1C) into a scenario of over-trust (Figure 1B). We can observe that increasing cooperation bias (i.e., increasing both δ and χ) eases the coordination toward full cooperation entailed by the non-linear public goods with $f > 1$. If individuals mistakenly perceive that there are more cooperators in a population than there truly are, they may recognize that the collective benefits of cooperation can be attained, even in a configuration

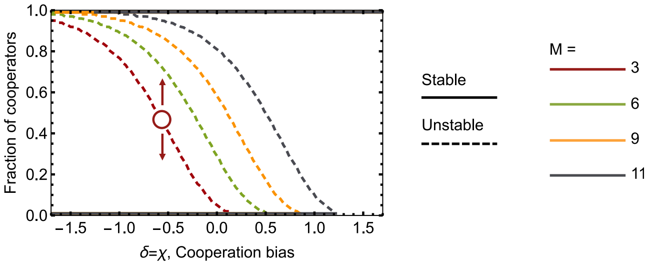


Figure 2. Under-trust and over-trust (homodirectional biases) impact the likelihood of reaching optimal coordination

In a coordination dilemma ($f > 1$), when there is no bias ($\delta = \chi = 0$), the dynamics of the population are characterized by a coordination threshold that corresponds to the fraction of cooperators above which the population will evolve toward full cooperation and below which it will evolve toward defection. That coordination threshold depends on the threshold within the interacting group, M, necessary for getting the reward. The dashed lines represent unstable equilibria: below them, there are insufficient cooperators, and the population evolves to a state of full defection; above, the population evolves to a state of full cooperation. Full lines at 0 and 1 represent stable equilibria in the fraction of cooperators. The left side of the figure, with negative biases toward cooperation ($\delta = \chi < 0$), is part of the under-trust region. The right side, with positive biases toward cooperation ($\delta = \chi > 0$), is part of the over-trust region. Over-trust promotes the coordination of a population into a cooperative state, whereas under-trust does the opposite. Effectively, biases toward the existence of cooperators reduce the coordination threshold, facilitating cooperation. Parameters: $N = 11$, $c = 1$, $b = 10$, $f = 1.5$, and $\chi = \delta$.

where the number of cooperators is still insufficient. Conversely, reducing cooperation bias (i.e., decreasing both δ and χ) induces individuals to understand that the collective benefits of cooperation are harder to be reached, even in situations where, actually, there are a sufficient number of cooperators to realize collective success. As such, under-trust hinders coordination toward full cooperation, requiring a higher number of cooperators to have a population self-organize toward the socially desirable outcomes. The effect of over- and under-trust on coordination toward cooperation can be grasped by the position of the coordination point in Figure 2: for different values of M, increasing δ and χ reduces the position of the coordination point (represented with dashed lines), implying that a smaller fraction of cooperators is needed to evolve toward full cooperation.

False uniqueness and false consensus lead to suboptimal deadlocks

The effects observed in Figure 2 result from homodirectional bias, that is, situations in which both cooperators and defectors over- or underestimate the real number of cooperators in a population. Social perception bias can, however, affect cooperators or defectors in different directions. In the case of false consensus (Figure 1D), individuals overestimate the adoption of their own strategy in a population, meaning that cooperators will overestimate the fraction of cooperators and defectors will overestimate the fraction of defectors. If one considers heterodirectional biases of this kind, the effects on cooperation dynamics become more intricate. Figure 3 summarizes the effects of heterodirectional bias on cooperation dynamics, considering false uniqueness ($\delta = -\chi$, $\chi < 0$, left half of the figure) and false consensus ($\delta = -\chi$, $\chi > 0$, right half). We can observe that false uniqueness induces a stable coexistence of cooperators and defectors, which may not be sufficient to support high levels of collective success (see transparent methods, supplemental information, for more details on group achievement). On the other hand, false consensus introduces a ''neutral region'' in which both cooperators and defectors stick to their current strategy.

The different impacts of false consensus and false uniqueness on cooperation dynamics can be further understood if we examine the gradients of selection and the decisions characterizing each type of bias. Figure 4B shows the original selection gradient in the absence of any bias. As already discussed, in this case, the dynamics are simply characterized by a coordination threshold that corresponds to the fraction of cooperators above which the population will evolve toward full cooperation and below which it will evolve toward defection. As Figure 4E reveals, below that threshold, cooperators turn into defectors with high probability and defectors remain defectors, making the gradient of selection of cooperators negative. Above that threshold, defectors are likely to turn into cooperators, and cooperators stick to their strategy, making the gradient of selection of cooperators positive. If individuals undergo false uniqueness biases (Figure 4A), we observe that, at the macroscopic level, the population is likely to remain in a state where

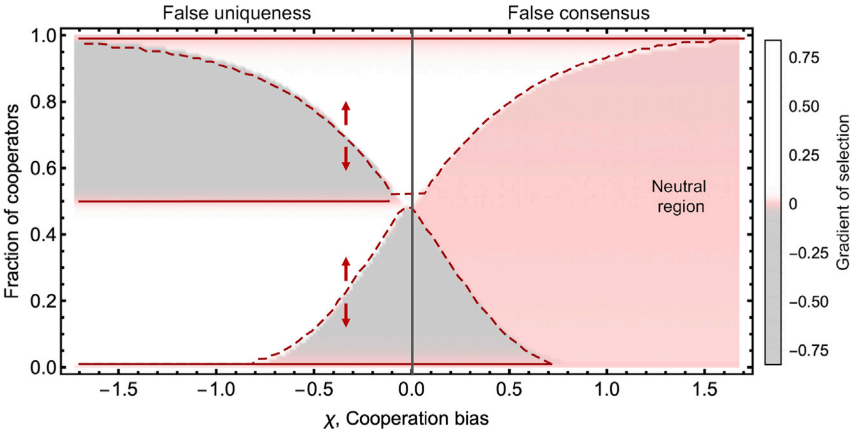


Figure 3. False uniqueness and false consensus (heterodirectional biases) lead to deadlocks resulting in individual and collective suboptimal configurations

We show the position of the equilibrium points associated with different biases. Dashed lines represent unstable equilibria, and full lines represent stable equilibria. Positive (negative) values of the gradient of selection, in white (gray), indicate a tendency for the number of cooperators to increase (decrease). False uniqueness ($\chi < 0$, left) is characterized by the existence of a stable configuration in which cooperators and defectors coexist, and the population is unable to solve the coordination dilemma. From the social-optimum point of view, this is the worst-case scenario because individuals contribute but not enough to surpass the threshold. A second—higher—coordination needs to be achieved for the population to reach a fully cooperative state. False consensus introduces a region where individuals believe there are no incentives to changing strategy even though the population is in a suboptimal configuration from the individual and collective point of view. In such a region, individuals do not change strategies, and the gradient of selection is 0 (neutral region, pink). Again, a second, higher, coordination needs to be achieved for the population to reach a fully cooperative state. Same parameters as Figure 1, with $M = 8$ and $\delta = -\chi$.

cooperators and defectors coexist. In Figure 4D, we can observe that this coexistence is motivated by a set of configurations in which both cooperators and defectors are likely to change their strategies: cooperators believe themselves to be surrounded by defectors, which motivates them to alter their strategy to defection; conversely, defectors expect that more cooperation exists than what actually occurs, which encourages themselves to become cooperators. A different dynamic is sustained by false consensus (Figure 4C). In this case, we observe an area in which any change in behaviors only occurs through exogenous factors (see supplemental Information). By further inspecting the likelihood that individuals change their strategy (Figure 4F), we realize that a neutral region appears when neither cooperators nor defectors are incentivized to alter their strategies: as everyone overestimates the representativeness of their own strategy in the population, cooperators believe that the cooperation threshold will be achieved, thus expecting to receive high benefits for cooperating, and defectors are convinced that such threshold is hardly attained, assuming no benefits for starting cooperating.

The previous results are confirmed in Figure 5 by a time-series analysis, where we assume that a large population of individuals ($Z = 1{,}000$) evolve following the best-response process detailed above (and in the transparent methods section, supplemental information). We confirm that false uniqueness originates a prevalent cooperator-defector coexistence, and false consensus introduces a neutral region where, over time, individuals maintain their strategies; both scenarios are sub-optimal, leading to many groups failing to achieve collective success.

Perception biases affect the effectiveness of monetary incentives

The previous effects of perception bias on cooperation dynamics imply that (1) different biases may have an impact on achieving high levels of collective success and (2) interventions are likely to have a different impact depending on whether individuals in a given population reveal a specific bias. Reasoning about bias and incentives simultaneously also suggests comparing the effect of interventions based on (possibly monetary) incentives such as rewards and punishment (Andreoni et al., 2003; Balliet et al., 2011; Couto et al., 2020; Dreber et al., 2008; Góis et al., 2019; Vasconcelos et al., 2013) with the effect of interventions that alter the information landscape available to individuals, akin to norm-based interventions

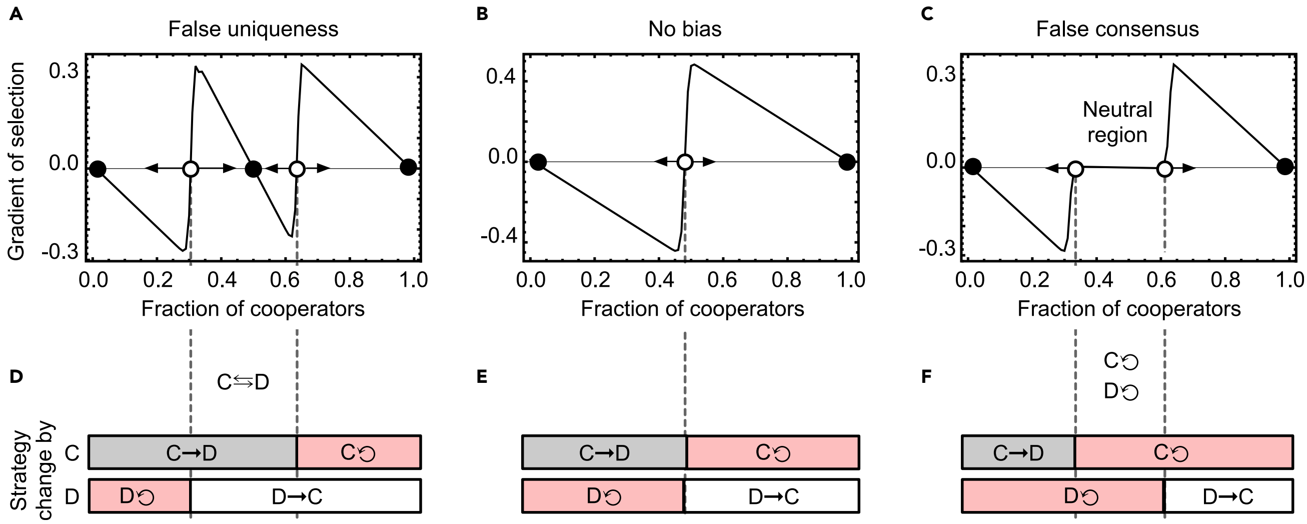


Figure 4. False uniqueness originates a stable cooperator-defector coexistence, whereas false consensus introduces a neutral region on cooperation dynamics

(A–F) The gradient of selection (A–C) measures how likely it is for cooperators to spread in a population, compared with defectors. Positive gradient values mean that cooperators are more likely to spread than defectors. As noted in Figure 3, false uniqueness induces a stable coexistence of cooperators and defectors (A). Further inspection of the strategic dynamics informs that this coexistence is due to a recurring transition of cooperators into defectors and defectors into cooperators (D). Given that individuals adopting a given strategy underestimate the representativeness of that behavior, everyone is inclined to change strategies: cooperators, as they do not believe that a minimal threshold of cooperation can be reached; defectors, as they believe that the threshold was already reached. For reference, we include the gradient corresponding to the no-bias situation (B and E); in that case, stabilizing cooperation requires overcoming one coordination barrier. If false consensus prevails, we note an inactivity area (neutral region, C) where both cooperators and defectors are satisfied with their strategy. Individuals overestimate the representativeness of their strategy in the population; as such, cooperators keep their strategy as they believe that the cooperation threshold was already reached, whereas defectors keep defecting as they believe that the threshold can never be reached (F). We consider $\chi = -\delta = -0.2$ (false uniqueness, A) $\chi = \delta = 0$ (no bias, B), and $\chi = -\delta = 0.2$ (false consensus, C). Same parameters as in Figure 2. See also Figure S1 for analysis of the effects of spontaneous changes and errors.

(Carattini et al., 2019; Miller and Prentice, 2016; Nyborg et al., 2016). Monetary incentives and information (media) campaigns are typical tools to change norms and behaviors (Bicchieri, 2016). We should also note that individuals tend to overestimate the impact of self-interest on the attitudes and behaviors of others (Miller and Ratner, 1998), and this tendency is particularly salient when information is incomplete (Vuolevi and Van Lange, 2010), which again denotes an interplay between monetary incentives (appealing to self-interest) and information incentives (attempting to reduce uncertainty).

Establishing a quantitative link among incentives, bias, and collective success is only possible by considering the combined effect of the different equilibria and dynamical regions identified in Figures 3 and 4. So, now, we turn our attention to identifying the time that a population spends in each state and what the chance is that group success is achieved in those states. This can be accomplished by focusing on a finite population of size Z and analyzing the stochastic, individual decisions. We assume the same process as before but allow for a small probability of not adopting a strategy that is the best response (also called a smooth best response (Fudenberg et al., 1998), which mimics uncertainty in estimating the payoff differences of the order to $1/\beta$) and randomly adopting any possible strategy (with probability μ). Moreover, we alter the game to include punishment applied to defectors (e.g., fines, higher tariffs, or taxes) by an amount ιc, $0 \leq \iota \leq 1$. The value of ι represents how the fines imposed compare with the costs paid by cooperators, with $\iota = 0$ meaning that no punishment is imposed and $\iota = 1$ meaning that all the payoff advantage of defectors, when compared with cooperators, is removed. In Figure 6, we show that increasing the magnitude of punishment has a different effect depending on the nature of bias prevailing in a population. For instance, a lower punishment is necessary to sustain collective success under false consensus, compared with false uniqueness (for the combination of parameters analyzed, in particular, high value $M = 8$). In fact, the prevalent coexistence characterizing false uniqueness and identified in Figure 4A may lead to a fraction of cooperators that remains insufficient to guarantee high average levels of group success; circumventing such stable coexistence of cooperators and defectors proves to be

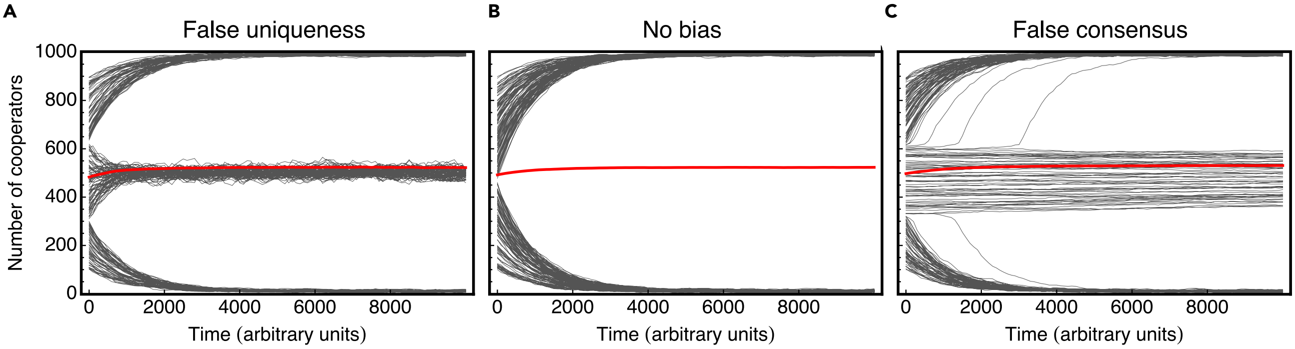


Figure 5. False uniqueness originates a stable cooperator-defector coexistence, whereas false consensus introduces a neutral region on cooperation dynamics (time-series analysis)

(A–C) We simulate the time evolution of strategy adoption in large populations composed of ($Z = 1{,}000$) individuals incurring (A) false uniqueness, (B) no bias, or (C) false consensus. Each gray curve corresponds to a single run starting from a random initial condition (in terms of the initial number of cooperators). The red curve corresponds to the average over all runs. We confirm that false uniqueness originates a prevalent cooperator-defector coexistence, where populations with an intermediate number of initial cooperators get trapped in a deadlock configuration. False consensus, on the other hand, leads to a neutral region where individuals maintain their strategies (eventually approaching the limits of such area and evolving to either full cooperation or full defection). Same parameters as in Figure 2.

harder—requiring extra incentives—than eliminating the neutral region associated with false consensus (Figure 4C). Additionally, we can observe that an effort to reduce individuals' perception biases can render high levels of collective success, even in situations where low incentives (low ι) are not effective—as a baseline, we show the group success characterizing a situation where neither biases nor incentives lead to the coordination in virtually all groups (Figure 6, gray curves).

Here, we assume that incentives are exogenously imposed (Góis et al., 2019) and do not introduce punishment strategies as in, e.g., Couto et al. (2020), Hauert et al. (2007), Quan et al. (2017), Roos et al. (2015), and Vasconcelos et al. (2013). Often, implementing incentives and institutions results in second-order free-riding dilemmas; we argue that, even if such dilemmas are solved, biases can affect the effectiveness of punishment and rewards. Also, we note that a direct comparison of the costs required to implement monetary-based incentives and information-based incentives is case sensitive, and future works can build on the model we propose for that purpose. Our results, however, already show that leveraging cooperation and group success may benefit from explicitly identifying and addressing individuals' social perception biases.

DISCUSSION

Understanding how cooperation can be sustained in public goods dilemmas of different kinds is central to address many of society's current challenges. That endeavor can benefit from recognizing the effect of social perception bias in cooperation dynamics and setting up incentives and interventions that understand and incorporate those dynamics. Here, we show that different types of social perception bias (e.g., false consensus, false uniqueness, over-trust, or under-trust) play a distinct role in the behavioral dynamics associated with non-linear public goods. Over-trust (under-trust) is likely to ease (hinder) the coordination associated with reaching the minimal number of contributors for cooperation to self-organize. False uniqueness leads to a persistent coexistence of cooperators and defectors, which can be insufficient to achieve collective success. Conversely, false consensus originates a neutral region where it is expected that individuals stick with their strategies, possibly changing behaviors only through exploration (Traulsen et al., 2009) and motives extraneous to the game being played. The fact that biases generate new, stable equilibria can have strong implications for the functioning of society. The workings and efficiency of markets and market regulation rely on the bottom-up ability of selfish agents to achieve socially desired outcomes and not get stuck in deadlocks as the ones we identify. Furthermore, these new equilibria are damaging for the possibility to coordinate from unfavorable into highly favorable states. They halt such a transition even in situations when all individuals would personally benefit from it. Besides implying different dynamics, such biases can render incentives less effective: as a prototypical example, false uniqueness requires that additional punishment is imposed on defectors (or, equivalently, rewards on cooperators) to achieve

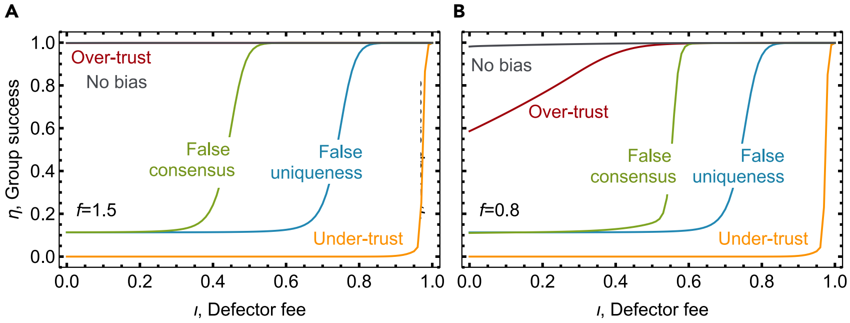


Figure 6. Perception biases affect the effectiveness of monetary incentives (such as a fee to be paid by defectors)

Incentives, like reward or punishment, are often used to move populations from unfavorable to favorable equilibria. The effectiveness of incentives, however, depends on the level and nature of biases existent in a population. Here, we measure group success, i.e., the fraction of groups that, on average, have the necessary number of cooperators to reap the benefits of collective action. We explore a game setting in which unbiased individuals self-organize toward high levels of group success. In a population with individuals that over-trust ($\chi = \delta = 0.6$), extra incentives are unnecessary to achieve group success if full cooperation is an equilibrium (panel A, $f = 1.5$); incentives also improve cooperation when individuals over-trust and if there is no incentive for cooperation above the threshold of group success (panel B, $f = 0.8$). In this case, over-trusting individuals may refrain from cooperating when they erroneously believe that the collective success threshold was already achieved. If individuals incur in false consensus ($\chi = -\delta = 0.6$), a lower punishment on defectors (or conversely, reward to cooperators) is necessary, compared with a scenario of false uniqueness ($\chi = -\delta = -0.6$). Finally, in a population with individuals that under-trust ($\chi = \delta = -0.6$), monetary incentives are ineffective to a large extent. Same parameters as Figure 3. Other parameters: $Z = 100$, $\beta = 10$, $\mu = 0.05$.

See also Figures S2–S5 for an extended exploration of incentives and biases in other games, as well as an exploration of different population sizes, group sizes, and selection intensities. See Figure S6 for heterogeneous, normally distributed biases in a population.

the same levels of group success, when compared with, e.g., populations under the effect of false consensus, and both require severer punishment compared with the absence of biases.

Although, currently, we focus on populations homogeneous in terms of bias and social contacts, the mathematical framework we propose can, in the future, be tuned to explicitly consider differences in biases within the same populations (Pearson et al., 2018) and the extent to which different network topologies may augment the effect of perception bias on cooperation. In fact, some authors suggest that social biases and judgment errors are often contradictory (Krueger and Funder, 2004). In this regard, considering the social network of interacting individuals not only may prove desirable to re-create realistic settings but also can be instrumental in explaining the origin of social perception biases and reconciling the apparently contradictory ones. Lee et al. show that considering homophily and interactions over a social network can help to explain seemingly conflicting biases, such as the overestimation and underestimation of a minority group size (Lee et al., 2019). Similarly, Galesic et al. show that homophily and a sampling process whereby individuals derive their judgments from local information based on their social environment (e.g., family, friends, and acquaintances) can explain when false consensus or false uniqueness is expected to occur (Galesic et al., 2018). Alipourfard et al. further show that individuals' perceptions can be biased as a result of local correlations in a directed social network (Alipourfard et al., 2020), and Lerman et al. show that social network effects can lead individuals to overestimate states that are globally rare, if those are overrepresented in their local neighborhoods—a phenomenon named majority illusion (Lerman et al., 2016). If perception biases result from social network effects rather than cognitive flaws, interventions based on reshaping information flows about global behaviors are possible and can be very impactful.

The analysis performed here is particularly relevant and timely given the growing number of works showing that individuals systematically under- or overestimate the position of others in matters affecting collective action problems (Kjeldahl and Hendricks, 2018; Leviston et al., 2013; Mildenberger and Tingley, 2019; Monin and Norton, 2003; Pearson et al., 2018) (also beyond climate change [Suls et al., 1988]). In fact, such perception biases are only but a subset of cognitive barriers that might affect decision making and impede collective action toward a better future (Weber, 2017). To reason about how those biases come to be and change over time is indispensable for a mechanistic understanding of the feedbacks between interventions and the biases themselves. As mentioned above, the existence of perception bias can be a by-product of individuals' psychological states, as well as the influence of local assortment (Cooney et al., 2016; Lee et al., 2019), specific network topologies (Alipourfard et al., 2020), or information filtering. False consensus, particularly, is likely to emerge if individuals' opinions assort them. Establishing a link between bias and cooperation can further illuminate how cooperation dynamics can depend on factors such as opinion polarization and assortment (McCarty, 2019), echo chambers (Colleoni et al., 2014), information cocoons (Sunstein, 2007), or on decisions about which opinions to share on mass media (Bowen et al., 2021; Boykoff and Boykoff, 2004; Feldman et al., 2012). To realize the emergence and persistence of these biases, one can also focus on the coevolutionary dynamics of strategic behavior at par with the evolutionary dynamics of beliefs (Galesic et al., 2021) and biases (Johnson and Fowler, 2011; Leimar and McNamara, 2019).

Different issues can also be associated with different levels of perception biases. Those levels depend on how visible issues are (Shamir and Shamir, 1997) and how visible the number of individuals supporting them is. Visibility can be a matter of design (e.g., using a COVID-19 tracing app entails the decision to give up privacy and contribute to a public good; informing how many people are using it is a decision of the designer) and policy-making (Nyborg et al., 2016). As Bicchieri puts it, solving *collective action traps* may require a *collective change of expectations* (Bicchieri, 2016). In this regard, our work provides a mechanistic understanding of how norm-based interventions (aiming at changing individuals' perceptions and expectations [Carattini et al., 2019; Miller and Prentice, 2016; Prentice and Paluck, 2020; Tankard and Paluck, 2016]) and information design (Mathevet et al., 2020) can be fundamental tools to trigger and sustain collective action.

Limitations of the study

The current study focuses on dilemmas that consist of the binary decision to cooperate or defect. Furthermore, we do not model explicitly how perception biases evolve. Future studies can address these limitations by extending the proposed model to understand the role of perception biases in dilemmas with continuous contribution decisions (e.g., deciding how much to contribute to collective success from a range of possible contributions), strategies explicitly conditioned on the number of expected cooperators (Ohtsuki 2018) and in contexts where biases can evolve at par with strategies. Bias dynamics can be studied in several ways: On the one hand, as introduced earlier, different biases can emerge in particular network topologies and as a function of individuals' homophily degree (e.g., see Galesic et al., 2018; Lee et al., 2019; Lerman et al., 2016), which calls to consider biases and cooperation dynamics on top of interaction networks. On the other hand, the development of biases can be studied through evolutionary models that explicitly define perception biases' fitness (e.g., as in Johnson and Fowler [2011] where the evolution of overconfidence is studied in the context of conflicts over resources) or through multi-level selection models (Cooney, 2019), where groups with particular sizes, structures, and information dissemination tools can inspire or solve specific perception biases that affect internal cooperation levels and the consequent capacity to outperform other groups. Finally, here we assume that individuals can, at least, track the direction of shifts in cooperation levels correctly. One can argue that biases can also prevent detecting such changes. In this regard, we note that previous works establish a distinction between bias and accuracy (West and Kenny, 2011) such that individuals may systematically misperceive the real cooperation levels due to biases toward their own perspective but accurately track changes in cooperation over time. It would be relevant to investigate, in the future, how accurately perceiving changes can be instrumental in designing incentives for cooperation in the same dilemmas we here study.

Resource availability

Lead contact

Further information and requests for materials should be directed to Fernando P. Santos (fppdsantos@gmail.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The data that support the results of this study are available from the corresponding authors upon request. The figures discussed result directly from the set of equations described in the transparent methods section (Supplemental Information). The code used to implement such equations and generate the figures is available from the corresponding authors upon request.

METHODS

All methods can be found in the accompanying transparent methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102375.

ACKNOWLEDGMENTS

F.P.S. acknowledges support from the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Dynamic and Multi-scale Systems—Postdoctoral Fellowship Award. V.V.V. acknowledges funding from the Princeton Institute for International and Regional Studies, Rapid Switch Community. S.A.L. and V.V.V. acknowledge funding by the Army Research Office grant no. W911NF-18-1-0325. S.A.L. acknowledges support from the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Dynamic and Multi-scale Systems—Collaborative Award, from the National Science Foundation grant no. CCF1917819, and from C3.ai Inc. and Microsoft Corporation. The authors would like to thank members of the Behavioral Science for Policy Lab (Princeton University) and participants of the Theoretical Ecology Lab Tea (EEB, Princeton University) for discussions and very helpful suggestions. The authors thank Cathy H. Teng and Hugues Martin Dit Neuville for discussions and their careful pre-submission internal review.

AUTHOR CONTRIBUTIONS

Conceptualization, F.P.S., S.A.L., and V.V.V.; Methodology, F.P.S. and V.V.V.; Software, F.P.S. and V.V.V.; Formal Analysis, F.P.S. and V.V.V.; Writing – Original Draft, F.P.S. and V.V.V.; Writing – Review & Editing, F.P.S., S.A.L., and V.V.V.; Supervision, S.A.L.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 2, 2020
Revised: February 26, 2021
Accepted: March 25, 2021
Published: April 23, 2021

REFERENCES

Ackermann, K.A., and Murphy, R.O. (2019). Explaining cooperative behavior in public goods games: how preferences and beliefs affect contribution levels. Games *10*, 15.

Alipourfard, N., Nettasinghe, B., Abeliuk, A., Krishnamurthy, V., and Lerman, K. (2020). Friendship paradox biases perceptions in directed networks. Nat. Commun. *11*, 1–9.

Andreoni, J., Harbaugh, W., and Vesterlund, L. (2003). The carrot or the stick: rewards, punishments, and cooperation. Am. Econ. Rev. *93*, 893–902.

Balliet, D., Mulder, L.B., and Van Lange, P.A. (2011). Reward, punishment, and cooperation: a meta-analysis. Psychol. Bull. *137*, 594.

Bicchieri, C. (2016). Norms in the Wild: How to Diagnose, Measure, and Change Social Norms (Oxford University Press).

Bowen, R., Dmitriev, D., and Galperti, S. (2021). Learning from Shared News: When Abundant Information Leads to Belief Polarization (National Bureau of Economic Research).

Boykoff, M.T., and Boykoff, J.M. (2004). Balance as bias: global warming and the US prestige press. Glob. Environ. Change *14*, 125–136.

Carattini, S., Levin, S., and Tavoni, A. (2019). Cooperation in the climate commons. Rev. Environ. Econ. Policy *13*, 227–247.

Colleoni, E., Rozza, A., and Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. J. Commun. *64*, 317–332.

Cooney, D., Allen, B., and Veller, C. (2016). Assortment and the evolution of cooperation in a Moran process with exponential fitness. J. Theor. Biol. *409*, 38–46.

Cooney, D.B. (2019). The replicator dynamics for multilevel selection in evolutionary games. J. Math. Biol. *79*, 101–154.

Couto, M.C., Pacheco, J.M., and Santos, F.C. (2020). Governance of risky public goods under graduated punishment. J. Theor. Biol. *505*, 110423.

Dawes, R.M., McTavish, J., and Shaklee, H. (1977). Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation. J. Personal. Social Psychol. *35*, 1.

Dietz, T., Ostrom, E., and Stern, P.C. (2003). The struggle to govern the commons. Science *302*, 1907–1912.

Dreber, A., Rand, D.G., Fudenberg, D., and Nowak, M.A. (2008). Winners don't punish. Nature *452*, 348–351.

Feldman, L., Maibach, E.W., Roser-Renouf, C., and Leiserowitz, A. (2012). Climate on cable: the nature and impact of global warming coverage on Fox News, CNN, and MSNBC. The Int. J. Press/Politics *17*, 3–31.

Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. Econ. Lett. *71*, 397–404.

Frey, B.S., and Meier, S. (2004). Social comparisons and pro-social behavior: testing" conditional cooperation" in a field experiment. Am. Econ. Rev. *94*, 1717–1722.

Fudenberg, D., Drew, F., Levine, D.K., and Levine, D.K. (1998). The Theory of Learning in Games (MIT press).

Galesic, M., Olsson, H., Dalege, J., van der Does, T., and Stein, D.L. (2021). Integrating social and cognitive aspects of belief dynamics: towards a unifying framework. J. R. Soc. Interface *18*, 20200857.

Galesic, M., Olsson, H., and Rieskamp, J. (2018). A sampling model of social judgment. Psychol. Rev. *125*, 363.

Geiger, N., and Swim, J.K. (2016). Climate of silence: pluralistic ignorance as a barrier to climate change discussion. J. Environ. Psychol. *47*, 79–90.

Goethals, G.R., Messick, D.M., and Allison, S.T. (1991). The uniqueness bias: studies of constructive social comparison. In Social Comparison: Contemporary Theory and Research, pp. 149–176.

Góis, A.R., Santos, F.P., Pacheco, J.M., and Santos, F.C. (2019). Reward and punishment in climate change dilemmas. Sci. Rep. *9*, 1–9.

Hardin, G. (1968). The tragedy of the commons. Science *162*, 1243–1248.

Hauert, C., Traulsen, A., Brandt, H., Nowak, M.A., and Sigmund, K. (2007). Via freedom to coercion: the emergence of costly punishment. Science *316*, 1905–1907.

Jachimowicz, J.M., Hauser, O.P., O'Brien, J.D., Sherman, E., and Galinsky, A.D. (2018). The critical role of second-order normative beliefs in predicting energy conservation. Nat. Hum. Behav. *2*, 757–764.

Johnson, D.D., and Fowler, J.H. (2011). The evolution of overconfidence. Nature *477*, 317–320.

Keohane, R.O., and Victor, D.G. (2016). Cooperation and discord in global climate policy. Nat. Clim. Change *6*, 570–575.

Kjeldahl, E.M., and Hendricks, V.F. (2018). The sense of social influence: pluralistic ignorance in climate change. EMBO Rep. *19*, e47185.

Krueger, J.I., and Funder, D.C. (2004). Towards a balanced social psychology: causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. Behav. Brain Sci. *27*, 313.

Lee, E., Karimi, F., Wagner, C., Jo, H.-H., Strohmaier, M., and Galesic, M. (2019). Homophily and minority-group size explain perception biases in social networks. Nat. Hum. Behav. *3*, 1078–1087.

Leimar, O., and McNamara, J.M. (2019). Learning leads to bounded rationality and the evolution of cognitive bias in public goods games. Sci. Rep. *9*, 1–9.

Lerman, K., Yan, X., and Wu, X.-Z. (2016). The" majority illusion" in social networks. PLoS One *11*, e0147617.

Levin, S.A. (1999). Fragile Dominion: Complexity and the Commons.

Leviston, Z., Walker, I., and Morwinski, S. (2013). Your opinion on climate change might not be as common as you think. Nat. Clim. Change *3*, 334–337.

Mathevet, L., Perego, J., and Taneva, I. (2020). On information design in games. J. Polit. Economy *128*, 1370–1404.

McCarty, N. (2019). Polarization: What Everyone Needs to Know® (Oxford University Press).

Mildenberger, M., and Tingley, D. (2019). Beliefs about climate beliefs: the importance of second-order opinions for climate politics. Br. J. Polit. Sci. *49*, 1279–1307.

Milinski, M., Sommerfeld, R.D., Krambeck, H.-J., Reed, F.A., and Marotzke, J. (2008). The collective-risk social dilemma and the prevention of simulated dangerous climate change. Proc. Natl. Acad. Sci. *105*, 2291–2294.

Miller, D.T., and McFarland, C. (1987). Pluralistic ignorance: when similarity is interpreted as dissimilarity. J. Personal. Social Psychol. *53*, 298.

Miller, D.T., and Prentice, D.A. (2016). Changing norms to change behavior. Annu. Rev. Psychol. *67*, 339–361.

Miller, D.T., and Ratner, R.K. (1998). The disparity between the actual and assumed power of self-interest. J. Personal. Soc. Psychol. *74*, 53.

Monin, B., and Norton, M.I. (2003). Perceptions of a fluid consensus: uniqueness bias, false consensus, false polarization, and pluralistic ignorance in a water conservation crisis. Personal. Social Psychol. Bull. *29*, 559–567.

Noelle-Neumann, E. (1974). The spiral of silence a theory of public opinion. J. Commun. *24*, 43–51.

Nyborg, K., Anderies, J.M., Dannenberg, A., Lindahl, T., Schill, C., Schlüter, M., Adger, W.N., Arrow, K.J., Barrett, S., and Carpenter, S. (2016). Social norms as solutions. Science *354*, 42–43.

Ohtsuki, H. (2018). Evolutionary dynamics of coordinated cooperation. Front. Ecol. Evol. *6*, 62.

Olson, M. (1965). The Logic of Collective Action: Public Goods and the Theory of Groups (Harvard University Press).

Ostrom, E. (2000). Collective action and the evolution of social norms. J. Econ. Perspect. *14*, 137–158.

Pacheco, J.M., Santos, F.C., Souza, M.O., and Skyrms, B. (2009). Evolutionary dynamics of collective action in N-person stag hunt dilemmas. Proc. R. Soc. B *276*, 315–321.

Pearson, A.R., Schuldt, J.P., Romero-Canyas, R., Ballew, M.T., and Larson-Konar, D. (2018). Diverse segments of the US public underestimate the environmental concerns of minority and low-income Americans. Proc. Natl. Acad. Sci. *115*, 12429–12434.

Prentice, D., and Paluck, E.L. (2020). Engineering social change using social norms: lessons from the study of collective action. Curr. Opin. Psychol. *35*, 138–142.

Prentice, D.A., and Miller, D.T. (1993). Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm. J. Personal. Social Psychol. *64*, 243.

Quan, J., Liu, W., Chu, Y., and Wang, X. (2017). Stochastic evolutionary voluntary public goods game with punishment in a Quasi-birth-and-death process. Sci. Rep. *7*, 1–14.

Roos, P., Gelfand, M., Nau, D., and Lun, J. (2015). Societal threat and cultural variation in the strength of social norms: an evolutionary basis. Organizational Behav. Hum. Decis. Process. *129*, 14–23.

Ross, L., Greene, D., and House, P. (1977). The "false consensus effect": an egocentric bias in social perception and attribution processes. J. Exp. Social Psychol. *13*, 279–301.

Santos, F.C., and Pacheco, J.M. (2011). Risk of collective failure provides an escape from the tragedy of the commons. Proc. Natl. Acad. Sci. *108*, 10421–10425.

Santos, F.P., Pacheco, J.M., Santos, F.C., and Levin, S.A. (2021). Dynamics of informal risk sharing in collective index insurance. Nat. Sustainability, 1–7.

Shamir, J., and Shamir, M. (1997). Pluralistic ignorance across issues and over time: information cues and biases. Public Opin. Q. 227–260.

Smith, D.L., Levin, S.A., and Laxminarayan, R. (2005). Strategic interactions in multi-institutional epidemics of antibiotic resistance. Proc. Natl. Acad. Sci. *102*, 3153–3158.

Suls, J., and Wan, C.K. (1987). In search of the false-uniqueness phenomenon: fear and estimates of social consensus. J. Personal. Social Psychol. *52*, 211.

Suls, J., Wan, C.K., and Sanders, G.S. (1988). False consensus and false uniqueness in estimating the prevalence of health-protective behaviors. J. Appl. Social Psychol. *18*, 66–79.

Sunstein, C.R. (2007). Republic.com 2.0 (Princeton University Press).

Taddicken, M., Kohout, S., and Hoppe, I. (2019). How aware are other nations of climate change? Analyzing Germans' second-order climate change beliefs about Chinese, US American and German people. Environ. Commun. *13*, 1024–1040.

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

Tavoni, A., Dannenberg, A., Kallis, G., and Löschel, A. (2011). Inequality, communication, and the avoidance of disastrous climate change in a public goods game. Proc. Natl. Acad. Sci. *108*, 11825–11829.

Traulsen, A., Hauert, C., De Silva, H., Nowak, M.A., and Sigmund, K. (2009). Exploration dynamics in evolutionary games. Proc. Natl. Acad. Sci. *106*, 709–712.

Vasconcelos, V.V., Santos, F.C., and Pacheco, J.M. (2013). A bottom-up institutional approach to cooperative governance of risky commons. Nat. Clim. Change *3*, 797–801.

Vuolevi, J.H., and Van Lange, P.A. (2010). Beyond the information given: the power of a belief in self-interest. Eur. J. Social Psychol. *40*, 26–34.

Weber, E.U. (2017). Breaking cognitive barriers to a sustainable future. Nat. Hum. Behav. *1*, 1–2.

West, T.V., and Kenny, D.A. (2011). The truth and bias model of judgment. Psychol. Rev. *118*, 357.

iScience, Volume *24*

Supplemental information

Biased perceptions explain collective action deadlocks and suggest new mechanisms to prompt cooperation

Fernando P. Santos, Simon A. Levin, and Vítor V. Vasconcelos

Supplemental Figure 1

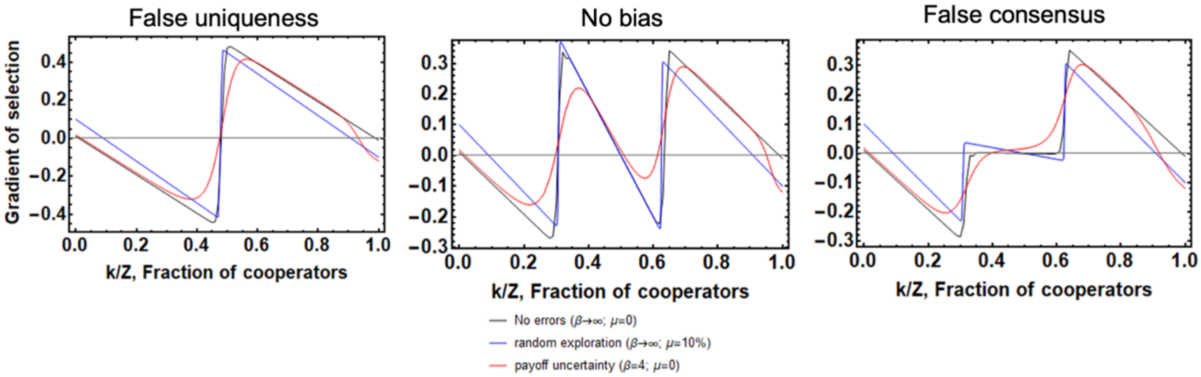


Figure S1. Effect of spontaneous changes (with probability $\boldsymbol{\mu}$, also known as mutation probability) and errors (controlled by the intensity of selection $\boldsymbol{\beta}$, the largest $\boldsymbol{\beta}$, the less errors individuals do when updating strategies) in the infinite population dynamics, Related to Figure 4. Other parameters: $\chi = \pm 0.2$, $\delta = \pm 0.2$, $f = 1.5$, $M = 8$.

Supplemental Figure 2

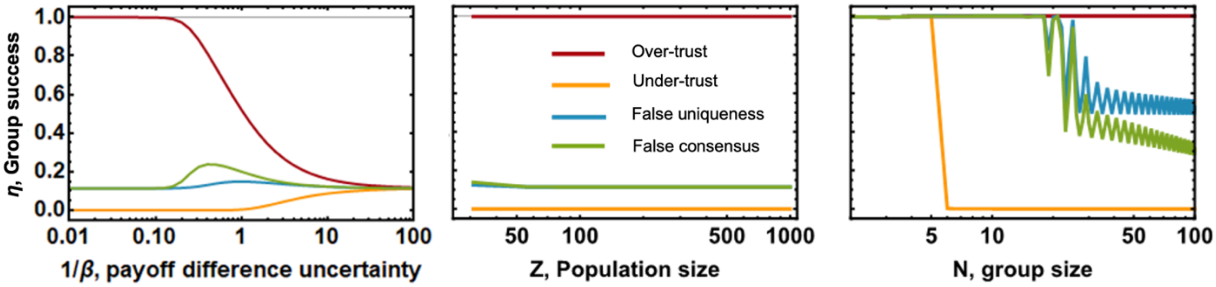


Figure S2. Role of payoff difference uncertainty (the inverse of selection intensity, $\mathbf{1/\beta}$), population size ($\boldsymbol{Z}$) and group size ($\boldsymbol{N}$) in overall group achievement, Related to Figure 6. When fixed, $Z = 100, M = 8, f = 1.5, N = 11, c = 1, b = 10, \beta = 10, \mu = 0.05$. $M =$ rounded$[0.5N]$ when N is varying.

Supplemental Figure 3

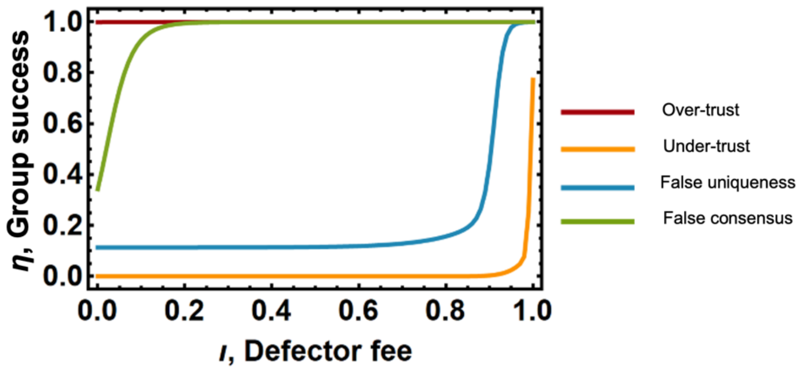


Figure S3. Effect of asymmetric biases, that is, deviation from diagonals in Figure 1, Related to Figure 6. Same parameters as Figure 6 in main text, but with $\chi = \pm 0.7, \delta = \pm 0.5$. Compare with Figure 6 of the main text.

Supplemental Figure 4

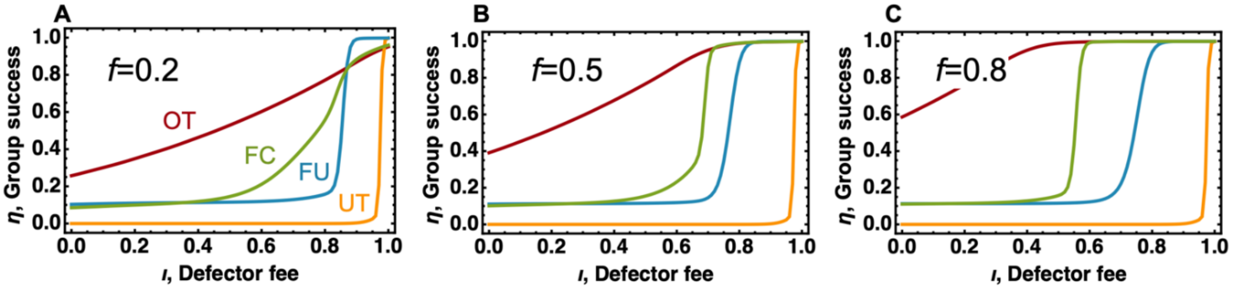


Figure S4. The role of incentives in populations with perception bias: Over-trust (OT), False consensus (FC), False uniqueness (FU) and Under-trust (UT), Related to Figure 6. Results for N-person game with co-existence (f<1), that is, where individuals do not have incentive to contribute further after the group has achieved the collective success threshold. Same parameters as Figure 6 in main text.

Supplemental Figure 5

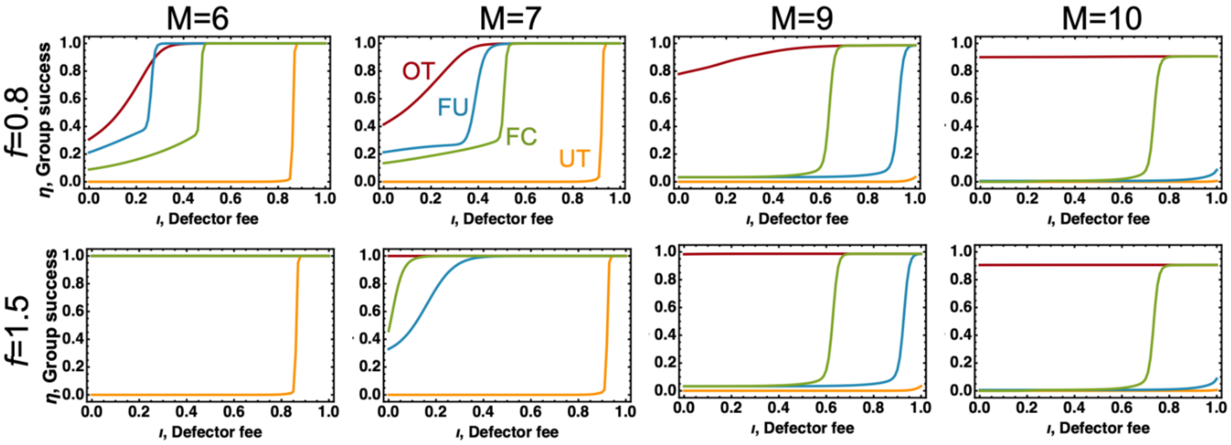


Figure S5. The role of incentives in populations with perception bias playing dillemas with different results from cooperation ($\boldsymbol{f = 0.8}$ and $\boldsymbol{f = 1.5}$) and group success threshold ($\boldsymbol{M = \{6, 7, 9, 10\}}$), Related to Figure 6. Same parameters as Figure 6 in main text. We can observe that, as discussed in the main text, false consensus (FC) requires less punishment to achieve high values of collective success when M and *f* are both high. For low M and *f* we can also observe situations in which false uniqueness (FU) leads to scenarios where it becomes easier to incentivize cooperation. We also represent group success under over-trust (OT) and under-trust (UT).

Supplemental Figure 6

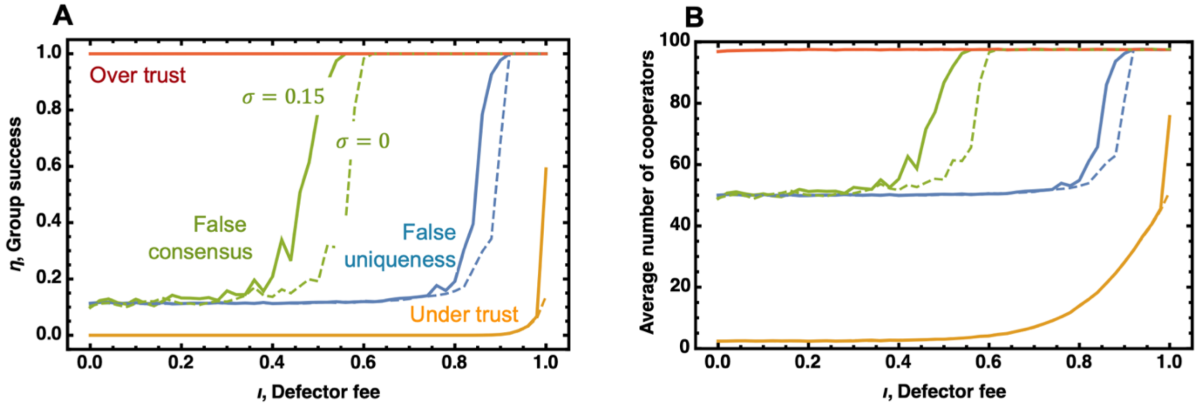


Figure S6. The role of incentives under bias heterogeneity, Related to Figure 6. Contrarily to the homogeneous bias scenario whose results we represent in Figure 6, here we assume that χ and δ convey the mean of a normal distribution with standard deviation σ**.** We compute, numerically, the average fraction of cooperators and average group success for each value of defector fee (ι), taken over 1000 generations (where, at each generation, Z individuals have the possibility of changing strategies). Instead of using the transition probabilities detailed in Equations (9) and (10) below — which preclude any difference among cooperators or among defectors — we explicitly sample individuals, each possibly characterized by a specific value of bias, and compute the probability that this unique individual changes strategy. Assuming that strategy updates follow this stochastic process (also accounting for a mutation probability, $\mu = 0.05$), we keep track of the number of strategies in the population in each generation, which allows us to quantify the average collective success (left panel) and average number of cooperators (right panel) associated with each value of fee ι. This way, we are able to compute the collective success assuming an arbitrary distribution of biases in the population. Each combination of cooperators' and defectors' biases (χ and δ) represents the mean of a Normal distribution with standard deviation σ. For each bias (false uniqueness, false consensus, over-trust, and under-trust), we use the same (χ,δ) as in Figure 6. We plot results for $\sigma = 0$ and $\sigma = 0.15$. Despite the noise associated with the numerical procedure we use here — note that now, on top of bias heterogeneity, to compute a smooth transition probability between states one needs a very large number of samples — we are able to confirm the results of Figure 6 in a scenario of bias heterogeneity. Same parameters as Figure 6. Dashed lines correspond to $\sigma = 0$ (homogeneous bias) and full lines to $\sigma = 0.15$ (heterogeneous bias).

Transparent Methods

Payoff: Players interact in groups of fixed size N to obtain a payoff Π_X that depends on their action, $X = C$ or D, and other players' actions. Action C corresponds to costly cooperation, and action D corresponds to defection. In an interacting group, cooperation costs an amount c, and, if there are less than M cooperators, there is no benefit to any of the group members. Whenever the group reaches a threshold of M cooperators, each individual gets a benefit, bc, plus an additional reward per extra cooperator, fc. Thus, if we let j be the number of cooperators in a group, we can write

$$\Pi_D[j] = \big(bc + fc(j - M)\big)\Theta[j - M] \text{ and} \tag{1}$$

$$\Pi_C[j] = \Pi_D[j] - c, \tag{2}$$

where $\Theta[x]$ is the unit step function, which is 0 for $x < 0$ and 1 for $x \geq 0$. In Figure 6, we alter the game to include punishment applied to defectors (e.g., fines, higher tariffs, or taxes), by an amount ιc, $0 \leq \iota \leq 1$. The value of ι represents how the fines imposed compare with the costs paid by cooperators, with $\iota = 0$ meaning that no punishment is imposed and $\iota = 1$ that all the payoff advantage of defectors, in comparison with cooperators, is removed. Eq. (1) is thereby modified to $\Pi_D[j] = \big(bc + fc(j - M)\big)\Theta[j - M] - \iota c$.

Infinite populations: At each time unit, individuals have the same probability of considering changing their strategy. Changing strategy depends on the outcome they expect to get from their interactions, given the number of cooperators (and defectors) they perceive will be present. An actor playing X will compare the expected payoff of cooperation, $f_C[\tilde{x}^X]$, to the average payoff of defectors, $f_D[\tilde{x}^X]$, when interacting in a group of size N, depending on the perceived fraction of cooperators in the population perceived by that player, $\tilde{x}^X$. Each individual assumes they are equally likely to interact with all others, resulting in expected payoffs of

$$f_C[\tilde{x}^X] = \sum_{k=0}^{N-1} \binom{N-1}{k} (\tilde{x}^X)^k (1 - \tilde{x}^X)^{N-1-k} \Pi_C[k+1] \text{ and} \tag{3}$$

$$f_D[\tilde{x}^X] = \sum_{k=0}^{N-1} \binom{N-1}{k} (\tilde{x}^X)^k (1 - \tilde{x}^X)^{N-1-k} \Pi_D[k]\,. \tag{4}$$

The player with strategy X will change to the strategy Y if the expected average payoff is not worse, with a probability of $p_{X\to Y}[f_Y - f_X] = \Theta[f_Y - f_X]$. Finally, the perceived fraction of cooperators by each strategy, though influenced by the actual number of cooperators in the population, x, is affected by biases, which only act on the strategies of the other players. If χ and δ represent the biases affecting cooperators and defectors, respectively, then a cooperator will estimate a fraction of cooperators $\tilde{x}^C[x] = x^{10^{-\chi}} = \exp\big[10^{-\chi}\ln[x]\big]$ and a defector will estimate a fraction of defectors $\tilde{x}^D[x] = x^{10^{-\delta}} = \exp\Big[10^{-\delta}\ln[x]\Big]$ (where the second equality serves to clarify that $-\chi$ and $-\delta$ are exponents of 10). The previous equalities are also usefull to clarify that the choice of basis 10 is arbitrary and does not affect the generality of our results; different basis can be considered and, by rescaling χ and δ, the same results would follow — for example, we could consider basis e instead of 10, $\tilde{x}^D[x] = x^{e^{-\bar{\delta}}}$ and $\tilde{x}^C[x] = x^{e^{-\bar{\chi}}}$, in which case the results under basis 10 are recovered by equating $\bar{\chi} = \ln[10]\chi$ and $\bar{\delta} = \ln[10]\delta$. This formulation guarantees that positive (negative) values of χ and δ indicate overestimation (underestimation) of cooperation (see Figure 1 in the main text). If $\chi = \delta = 0$ then $\tilde{x}^C[x] = \tilde{x}^D[x] = x$, which recovers the typical no-bias scenario where perceptions match reality. Thus, we can write the probability that the number of Cs increases, and the number of Ds decreases, per time unit as $T^{+}[x] = (1-x)p_{D\to C}\left[f_C\left[\tilde{x}^D[x]\right] - f_D\left[\tilde{x}^D[x]\right]\right]$ and the probability that the number of Cs decreases, and the number of Ds increases, per time unit as $T^{-}[x] = x\,p_{C\to D}\left[f_D\left[\tilde{x}^C[x]\right] - f_C\left[\tilde{x}^C[x]\right]\right]$. The gradient of selection (Figure 4 and Figure S1) indicates the most likely direction of evolution of the population and is given by $g[x] = T^{+}[x] - T^{-}[x]$; when $g[x] > 0$, the number of cooperators is likely to increase and $g[x] < 0$ implies that cooperation is likely to decrease.

Finite populations: Let us now consider a population of size Z. As before, players interact in groups of fixed size $N \leq Z$ to obtain a payoff Π_X that depends on their action, $X = C$ or D, and other players' actions (as detailed above).

Each time unit, individuals have the same probability of considering changing their strategy based on the outcome they expect to get from their interactions, given the number of cooperators (and defectors) they perceive will be present. An actor playing X will compare the average payoff of cooperation, $f_C[\tilde{x}^X]$, to the average payoff of defectors, $f_D[\tilde{x}^X]$, depending on the perceived fraction of cooperators in the population seen by that player, $\tilde{x}^X$. We assume a complete graph of interactions from which the interaction groups are sampled, resulting in average payoffs of

$$f_C[\tilde{x}^X] = \sum_{k=0}^{N-1} \binom{Z-1}{N-1}^{-1} \binom{Z\,\tilde{x}^X - 1}{k} \binom{Z(1-\tilde{x}^X)}{N-1-k} \Pi_C[k+1] \text{ and} \tag{5}$$

$$f_D[\tilde{x}^X] = \sum_{k=0}^{N-1} \binom{Z-1}{N-1}^{-1} \binom{Z\,\tilde{x}^X}{k} \binom{Z(1-\tilde{x}^X)-1}{N-1-k} \Pi_D[k]\,. \tag{6}$$

The player with strategy X will change to strategy Y with a probability that increases with the difference of expected average payoffs. The player can also change spontaneously from one strategy to another at a rate μ, due to some exogenous event. Combining those effects resuls in a probability of changing strategy of $\mu + (1-\mu)\left(1 + e^{-\beta(f_Y - f_X)}\right)^{-1}$. Finally, the actual fraction of cooperators in the population, x, affects the perceived fraction of cooperators by each strategy. However, biases on the strategies of others also affect the latter. If χ and δ represent the bias affecting cooperators and defectors, respectively, then a cooperator will estimate a fraction of cooperators $\tilde{x}^C[x]$ and a defector will estimate a fraction of defectors $\tilde{x}^D[x]$ given by

$$\tilde{x}^C[x] = \left(\frac{Zx-1}{Z-1}\right)^{10^{-\chi}} + \frac{1}{Z} \text{ and} \tag{7}$$

$$\tilde{x}^D[x] = \left(\frac{Zx}{Z-1}\right)^{10^{-\delta}}\,. \tag{8}$$

This formulation guarantees that positive (negative) values of χ and δ indicate overestimation (underestimation) of cooperation. Again, we note that, in Eqs. (7) and (8), $-\chi$ and $-\delta$ are exponents of 10, and basis 10 was chosen without loss of generality. Thus, we can write the probability that the number of Cs increases, and the number of Ds decreases, per time unit, $T^{+}[x]$, and the probability that the number of Cs decreases, and the number of Ds increases, per time unit, $T^{-}[x]$, as

$$T^{+}[x] = (1-x)\left(\mu + (1-\mu)\left(1 + e^{-\beta\left(f_C\left[\tilde{x}^D[x]+\frac{1}{Z}\right] - f_D\left[\tilde{x}^D[x]\right]\right)}\right)^{-1}\right) \text{ and} \tag{9}$$

$$T^{-}[x] = x\left(\mu + (1-\mu)\left(1 + e^{-\beta\left(f_D\left[\tilde{x}^C[x]-\frac{1}{Z}\right] - f_C\left[\tilde{x}^C[x]\right]\right)}\right)^{-1}\right). \tag{10}$$

We note that this update resembles a smooth best-response (Fudenberg et al., 1998) and, in the past, was also used to model so-called counterfactual thinking (Pereira and Santos, 2018). As when considering infinite populations, the gradient of selection indicates the most likely direction of evolution of the population and is, thus, given by $g[x] = T^{+}[x] - T^{-}[x]$. In this case, diffusion indicates the level of noise of the system at any configuration and is given by $d[x] = (T^{+}[x] + T^{-}[x])/Z$.

Analysis of the dynamics for infinite populations

The gradient of selection and diffusion govern the dynamics, which can be written as:

$$\dot{x} = g[x] + \sqrt{d[x]}\,\Gamma[t], \tag{11}$$

where $\Gamma[t]$ is a random variable with gaussian distribution of zero mean and unit variance. Thus, when $g[x]$ is positive, x tends to increase. When $g[x]$ is negative, x tens to decrease. The sign of g alone contains the information of the preferential direction of evolution of the fraction of cooperators in the population.

In the case of an infinite population, $Z \to \infty$, and perfect best response, $\beta \to \infty$, we get

$$\dot{x} = \mu(1-2x) + (1-\mu)\Big((1-x)\,\Theta\big[f_C[\tilde{x}^D[x]] - f_D[\tilde{x}^D[x]]\big] - x\,\Theta\big[f_D[\tilde{x}^C[x]] - f_C[\tilde{x}^C[x]]\big]\Big), \tag{12}$$

and

$$f_C[\tilde{x}^X] = \sum_{k=\mathrm{M}-1}^{N-1} \binom{N-1}{k} (\tilde{x}^X)^k (1-\tilde{x}^X)^{N-1-k} (bc + fc(k-M) + fc) - c \tag{13}$$

$$f_D[\tilde{x}^X] = \sum_{k=\mathrm{M}}^{N-1} \binom{N-1}{k} (\tilde{x}^X)^k (1-\tilde{x}^X)^{N-1-k} \big(bc + fc(k-M)\big) \tag{14}$$

and

$$\begin{aligned}
f_C[\tilde{x}^X] - f_D[\tilde{x}^X] &= \sum_{k=\mathrm{M}-1}^{N-1} \binom{N-1}{k} (\tilde{x}^X)^k (1-\tilde{x}^X)^{N-1-k} (bc + fc(k-M) + fc) - c \\
&\quad - \sum_{k=\mathrm{M}}^{N-1} \binom{N-1}{k} (\tilde{x}^X)^k (1-\tilde{x}^X)^{N-1-k} \big(bc + fc(k-M)\big) \\
&= bc\binom{N-1}{\mathrm{M}-1} (\tilde{x}^X)^{\mathrm{M}-1} (1-\tilde{x}^X)^{N-M} + fc\sum_{k=\mathrm{M}}^{N-1} \binom{N-1}{k} (\tilde{x}^X)^k (1-\tilde{x}^X)^{N-1-k} - c \\
&= fc(1 - \mathrm{CDF}[\mathrm{Binomial}[N-1, \tilde{x}^X], M-1]) + \binom{N-1}{M} (\tilde{x}^X)^{M-1} (1-\tilde{x}^X)^{N-M} bc - c.
\end{aligned}$$

Notice that $f_C[0] - f_D[0] = -c < 0$ and, when $f > 1$, $f_C[1] - f_D[1] = fc - c > 0$. If $X = D$, this change of signs guarantees that, from the perspective of a defector, there is at least one coordination dilemma, i.e., there is no incentive to change strategy if there are too few cooperators, and there is an incentive to become a cooperator if there are enough cooperators. Identically for the perspective of cooperators, when $X = C$, irrespectively of the bias function.

Analysis of the dynamics for finite populations

Recovering Eqs.(5-6) and Eqs.(1-2), we can write

$$\begin{aligned}
f_D[\tilde{x}^X] &= \sum_{k=0}^{N-1} \binom{Z-1}{N-1}^{-1} \binom{Z\,\tilde{x}^X}{k} \binom{Z(1-\tilde{x}^X)-1}{N-1-k} \Pi_D[k] \\
&= \sum_{k=0}^{N-1} \binom{Z-1}{N-1}^{-1} \binom{Z\,\tilde{x}^X}{k} \binom{Z(1-\tilde{x}^X)-1}{N-1-k} \big(bc + fc(k-M)\big)\Theta[k-M] \\
&= \sum_{k=M}^{N-1} \binom{Z-1}{N-1}^{-1} \binom{Z\,\tilde{x}^X}{k} \binom{Z(1-\tilde{x}^X)-1}{N-1-k} \big(bc + fc(k-M)\big)
\end{aligned} \tag{15}$$

$$\begin{aligned}
f_C[\tilde{x}^X] &= \sum_{k=0}^{N-1} \binom{Z-1}{N-1}^{-1} \binom{Z\,\tilde{x}^X-1}{k} \binom{Z(1-\tilde{x}^X)-\delta_{XD}}{N-1-k} \Pi_C[k+1] \\
&= \sum_{k=0}^{N-1} \binom{Z-1}{N-1}^{-1} \binom{Z\,\tilde{x}^X-1}{k} \binom{Z(1-\tilde{x}^X)}{N-1-k} \big(bc + fc(k+1-M)\big)\Theta[k+1-M] - c \\
&= \sum_{k=\mathrm{M}-1}^{N-1} \binom{Z-1}{N-1}^{-1} \binom{Z\,\tilde{x}^X-1}{k} \binom{Z(1-\tilde{x}^X)}{N-1-k} \big(bc + fc(k+1-M)\big) - c \\
&= \binom{Z-1}{N-1}^{-1} \binom{Z\,\tilde{x}^X-1}{M-1} \binom{Z(1-\tilde{x}^X)}{N-\mathrm{M}} bc + \\
&\quad + \sum_{k=\mathrm{M}}^{N-1} \binom{Z-1}{N-1}^{-1} \binom{Z\,\tilde{x}^X-1}{k} \binom{Z(1-\tilde{x}^X)}{N-1-k} \big(bc + fc(k+1-M)\big) - c
\end{aligned} \tag{16}$$

To compute T^+ we need

$$\begin{aligned}
& f_C\left[\tilde{x}^D[x] + \frac{1}{Z}\right] - f_D\big[\tilde{x}^D[x]\big] = \\
&\qquad = \binom{Z-1}{N-1}^{-1} \binom{Z\,\tilde{x}^D}{M-1} \binom{Z(1-\tilde{x}^D)-1}{N-\mathrm{M}} bc + \\
&\qquad\quad + fc \sum_{k=\mathrm{M}}^{N-1} \binom{Z-1}{N-1}^{-1} \binom{Z\,\tilde{x}^D}{k} \binom{Z(1-\tilde{x}^D)-1}{N-1-k} - c \\
&\qquad = fc\big(1 - \mathrm{CDF}\big[\mathrm{HyperGeo}[Z-1, N-1, Z\tilde{x}^D], M-1\big]\big) + \\
&\qquad\quad + \binom{Z-1}{N-1}^{-1} \binom{Z\,\tilde{x}^D}{M-1} \binom{Z(1-\tilde{x}^D)-1}{N-\mathrm{M}} bc - c.
\end{aligned} \tag{17}$$

To compute T^- we need

$$\begin{aligned}
& f_D\left[\tilde{x}^C[x] - \frac{1}{Z}\right] - f_C\big[\tilde{x}^C[x]\big] = \\
&\qquad = -\binom{Z-1}{N-1}^{-1} \binom{Z\,\tilde{x}^C-1}{M-1} \binom{Z(1-\tilde{x}^C)}{N-M} bc - \\
&\qquad\quad - fc \sum_{k=\mathrm{M}}^{N-1} \binom{Z-1}{N-1}^{-1} \binom{Z\,\tilde{x}^C-1}{k} \binom{Z(1-\tilde{x}^C)}{N-1-k} + c \\
&\qquad = -fc\big(1 - \mathrm{CDF}\big[\mathrm{HyperGeo}[Z-1, N-1, Z\tilde{x}^C-1], M-1\big]\big) - \\
&\qquad\quad - \binom{Z-1}{N-1}^{-1} \binom{Z\,\tilde{x}^C-1}{M-1} \binom{Z(1-\tilde{x}^C)}{N-M} bc + c.
\end{aligned} \tag{18}$$

Importantly, we can write

$$\tilde{x}^X[x] = \left(\frac{Zx - \delta_{XC}}{Z - 1}\right)^{10^{-b_X}} + \frac{\delta_{XC}}{Z}, \tag{19}$$

with $b_X = \chi\delta_{XC} + \delta\,\delta_{XD}$. Here, δ_{XY} represents the Kroneker delta and should not be confused with δ (without subscripts and representing defectors' bias): $\delta_{XY} = 1$ if $X = Y$ and $\delta_{XY} = 0$ otherwise. For any $\mu > 0$ and finite β we can define a Markov chain of the number of cooperators over time, i, using the probabilities of increasing and decreasing i by one unit as $T^+[i/Z]$ and $T^-[i/Z]$, respectively. The evolution of i is governed by a Master-equation of the form

$$\frac{dp_i[t]}{dt} = p_{i-1}[t]T^+\left[\frac{i-1}{Z}\right] + p_{i+1}[t]T^-\left[\frac{i+1}{Z}\right] - p_i[t]\left(T^+\left[\frac{i}{Z}\right] + T^-\left[\frac{i}{Z}\right]\right), \tag{20}$$

where $p_i[t]$ is the probability of finding the system in configuration i after a period t in which the system was in some configuration i_0, $p_i[0] = \delta_{ii_0}$. The solution will converge to a stationary solution, p_i^*, which is independent of the initial condition i_0. Thus, p_i^* reflects the probability of finding the system with i cooperators a longer time after we observe i_0 cooperators (which is our best bet if there are no observations at all).

With it, we can compute the expected number of groups that reach the threshold, which we call group achievement, η. The group achievement is computed as

$$\eta = \sum_{i=0}^{Z} p_i^* \sum_{k=M}^{N} \binom{Z}{N}^{-1} \binom{i}{k} \binom{Z-i}{N-k}. \tag{21}$$

We can also compute the average level of cooperation simply as

$$\langle\frac{i}{Z}\rangle = \sum_{i=0}^{Z} p_i^* \frac{i}{Z}. \tag{22}$$

A note on the definition of false uniqueness

We note that the operationalization of false uniqueness that we use throughout our main text does not perfectly match all previous definitions of this social perception bias: false uniqueness was referred to as the tendency for individuals to underestimate the proportion of those sharing their desirable attributes (Baumeister and Vohs, 2007; Suls et al., 1988); an alternative would be using pluralistic ignorance, previously defined as the tendency for individuals to wrongly assume that their behaviors differ from everybody else's, which happens as public actions can differ from private beliefs and opinions (Baumeister and Vohs, 2007; Miller and McFarland, 1987). A completely accurate implementation of false uniqueness would require defining desirability, while a completely accurate implementation of pluralistic ignorance would require distinguishing public and private strategies in or model. As explicitly introducing desirability or private behaviors would increase the complexity of our model beyond the scope of the analysis we intend to perform (for the complexity of modeling desirability and private information associated with cooperation see, respectively, (Ohtsuki and Iwasa, 2004; Santos et al., 2018) and (Hilbe et al., 2018; Ohtsuki et al., 2015)), we opted to use false uniqueness to simply denote the tendency for individuals to underestimate the representativeness of their own strategy in the population, following works such as (Galesic et al., 2018; Krueger, 2000; Lee et al., 2019).

Supplemental References

Baumeister, R.F., Vohs, K.D., 2007. Encyclopedia of social psychology. Sage.

Fudenberg, D., Drew, F., Levine, D.K., Levine, D.K., 1998. The theory of learning in games. MIT press.

Galesic, M., Olsson, H., Rieskamp, J., 2018. A sampling model of social judgment. Psychological review 125, 363.

Hilbe, C., Schmid, L., Tkadlec, J., Chatterjee, K., Nowak, M.A., 2018. Indirect reciprocity with private, noisy, and incomplete information. Proceedings of the National Academy of Sciences 115, 12241–12246.

Krueger, J., 2000. The projective perception of the social world, in: Handbook of Social Comparison. Springer, pp. 323–351.

Lee, E., Karimi, F., Wagner, C., Jo, H.-H., Strohmaier, M., Galesic, M., 2019. Homophily and minority-group size explain perception biases in social networks. Nature Human Behaviour 3, 1078–1087.

Miller, D.T., McFarland, C., 1987. Pluralistic ignorance: When similarity is interpreted as dissimilarity. Journal of Personality and social Psychology 53, 298.

Ohtsuki, H., Iwasa, Y., 2004. How should we define goodness?—reputation dynamics in indirect reciprocity. Journal of theoretical biology 231, 107–120.

Ohtsuki, H., Iwasa, Y., Nowak, M.A., 2015. Reputation effects in public and private interactions. PLoS Comput Biol 11, e1004527.

Pereira, L.M., Santos, F.C., 2018. Counterfactual thinking in cooperation dynamics. Presented at the International conference on Model-Based Reasoning, Springer, pp. 69–82.

Santos, F.P., Santos, F.C., Pacheco, J.M., 2018. Social norm complexity and past reputations in the evolution of cooperation. Nature 555, 242–245.

Suls, J., Wan, C.K., Sanders, G.S., 1988. False consensus and false uniqueness in estimating the prevalence of health-protective behaviors. Journal of Applied Social Psychology 18, 66–79.