Maintaining trust when agents can engage in self-deception

Andrés Babino[a,b,1,2], Hernán A. Makse[c,d,1], Rafael DiTella[e,f,g,1], and Mariano Sigman[h,1]

[a]Departamento de Física J.J. Giambiagi, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires 1428, Argentina; [b]Instituto de Física de Buenos Aires, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires 1428, Argentina; [c]Levich Institute, City College of New York, New York, NY 10031; [d]Physics Department, City College of New York, New York, NY 10031; [e]Political Economy Group, National Bureau of Economic Research, Cambridge, MA 02138; [f]Social Interactions, Identity and Well-being Program, Canadian Institute for Advanced Research, Toronto, ON M5G 1M1, Canada; [g]Government and the International Economy Unit, Harvard Business School, Boston, MA 02163; and [h]Laboratorio de Neurociencia, CONICET, Universidad Torcuato Di Tella, C1428BIJ Buenos Aires, Argentina

Edited by Albert-László Barabási, Northeastern University, Boston, MA, and accepted by Editorial Board Member David A. Weitz July 17, 2018 (received for review February 28, 2018)

The coexistence of cooperation and selfish instincts is a remarkable characteristic of humans. Psychological research has unveiled the cognitive mechanisms behind self-deception. Two important findings are that a higher ambiguity about others' social preferences leads to a higher likelihood of acting selfishly and that agents acting selfishly will increase their belief that others are also selfish. In this work, we posit a mathematical model of these mechanisms and explain their impact on the undermining of a global cooperative society. We simulate the behavior of agents playing a prisoner's dilemma game in a random network of contacts. We endow each agent with these two self-deception mechanisms which bias her toward thinking that the other agent will defect. We study behavior when a fraction of agents with the "always defect" strategy is introduced in the network. Depending on the magnitude of the biases the players could start a cascade of defection or isolate the defectors. We find that there are thresholds above which the system approaches a state of complete distrust.

behavioral economics | cognitive neuroscience | corruption | cooperation | self-deception

Individuals often deviate from the behavior that maximizes their material reward (1, 2). For example, in the ultimatum game, people prefer to reject profitable offers that they consider unfair (3). This behavior, and other phenomena such as fairness or cooperation (2, 4), can be accounted for within a rational model that includes broader objectives or "social preferences" (altruism, fairness concerns, etc.) as part of the function which agents seek to optimize.

Naturally, agents seek to reduce the problems that arise when material rewards collide with social preferences. For example, believing that others are altruistic may make it more difficult for an agent to act selfishly which, in turn, may reduce its monetary payoff. A way of solving this tension is to develop a self-serving bias: that is, to believe that others are not altruistic to "justify" a selfish act. Cognitive dissonance theory (5, 6) aims to explain the emergence of belief with self-serving biases. The idea is that *dissonance* (contradiction) between cognitions is psychologically uncomfortable, and so it triggers mechanisms of dissonance reduction—and one way of doing so is by altering beliefs (7, 8).

Self-deception mechanisms have been broadly studied in economics (2, 9). Recently, using an experimental design called "The Corruption Game," we demonstrated two of these principles (10):

Principle 1 (P1) Selfish action alters beliefs about others' social-preferences.

P2 Ambiguity regarding others' social preferences increases the likelihood of acting selfishly.

We use the term *Projection* to refer to P1, which is a trait that describes how people blame others for their actions. The notion of *Projection* by which our actions affect how we think of others (11, 12) is at the same time intuitive and paradoxical. From a rational perspective, beliefs about others should be based on what they have done, not on what we have done to them. However, it has been observed that subjects in economic games not only take into account the previous actions of other players, but also their past actions (13, 14). Additionally, people's beliefs also depend on their own previous actions (10).

Here, we use the colloquial term *Paranoia* to refer to P2 (the idea that if there is ambiguity about how another person may act, an agent will sample the distribution biased for the worse outcomes). Closely related to P2 is the mechanism of "categorization" and "malleability" (15). For example, stealing a pen is more malleable than stealing the money needed to buy the pen. Similarly, the distribution of beliefs on the moral judgment of the malleable case (stealing the pen) is ambiguous, and hence people may use this ambiguity in their favor to act more selfishly.

The aim of this work is twofold: first, to provide a mathematical description of these self-deception mechanisms (*Paranoia* and *Projection*); and second, based on this mathematical description, to investigate the impact that they may have on the evolution of trust among the agents of a society.

The Model

We study a set of 10^5 interacting agents that play a modified Prisoner's Dilemma (*SI Appendix*, section 1.1) game against each other in a static random network. The main difference from other similar approaches investigating networks and evolution of

Significance

"He who wants to kill his dog accuses it of having rabies," the French proverb says. The fact that we alter our beliefs about others to act selfishly and, at the same time, keep a positive self-view has been widely studied by behavioral sciences. Here, we propose a mathematical description of two of these mechanisms of altering beliefs and study a simulation of a society of agents provided with these biases. We find that there are sets of parameters that make societies propagate defection actions and others that protect them from spreading malicious behavior.

Author contributions: A.B., H.A.M., R.D., and M.S. designed research; A.B. performed research; A.B. analyzed data; and A.B., H.A.M., R.D., and M.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. A.-L.B. is a guest editor invited by the Editorial Board.

[1]A.B., H.A.M., R.D., and M.S. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: ababino@df.uba.ar.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1803438115/-/DCSupplemental.

Published online August 13, 2018.

cooperation (or corruption or reputation) (16–19) is that here, we used a Bayesian updating rule and inference process (similar to ref. 20). This rule was necessary to generate a mathematical model of cognitive biases *Paranoia* and *Projection* and study their impact on the propagation of strategies.

For clarity, we divide the strategy of the agents into three stages (Fig. 1): observation, inference, and decision. Observation is the process of accumulation of information about other agent's actions. The inference process uses observed information and combines it with priors to generate—using a Bayesian model—a belief about other agents' behavior. This stage is modeled as a beta-binomial process (*SI Appendix*, section 1.3). The output of the inference process is the expected reward for each possible action. Finally, in the decision stage, the agent chooses the option that maximizes her expected reward.

Under these settings, the agents in the network will end up defecting or cooperating with each other depending on the initial conditions. Our primary goal is to investigate how incorporating the cognitive biases described in the introduction (*Paranoia* and *Projection*) affect the evolution of cooperation or defection in the network. In the next subsection, we explain how these cognitive biases can be incorporated into a Bayesian inference process.

The essential step in the inferential process in our model is the estimation of an agent's probability of defection, θ, in a given interaction—or equivalently, the probability of cooperation $p_c = 1 - \theta$. If the estimation of θ is small enough, the agent will trust the other player and will cooperate; if not, the agent will choose to defect. An agent estimates θ based on her previous observations of the other agent.

In the beta-binomial model, the $Beta$ distribution (*SI Appendix*, section 1.2) is used to describe the prior belief distribution of this variable. Agents use the mean value of their belief as an estimation of this parameter:

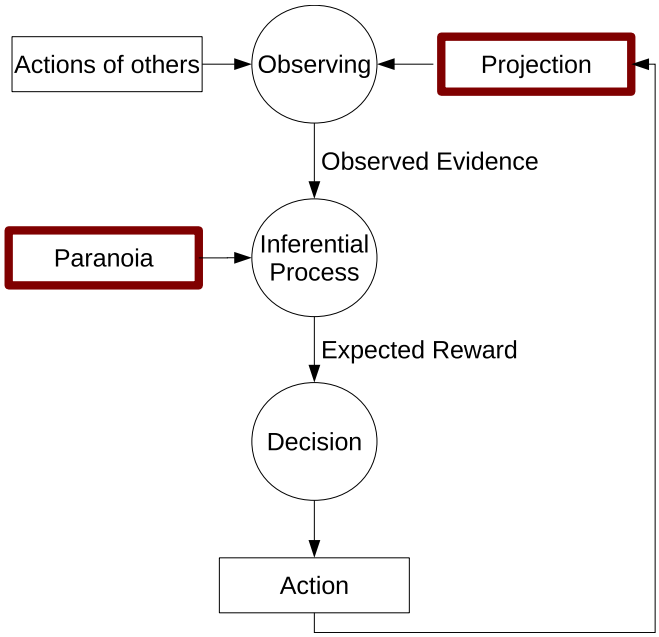

Fig. 1. Sketch of the decision-making process of each agent. The observing stage is fed with the actions of other agents, but also from the decisions of the own agent as a result of the *Projection* bias. The observed evidence (the number of defection and cooperation actions of the other agents) is used in the inferential process to estimate the expected reward of each possible action. This expected reward could be biased by *Paranoia*. Finally, based on the expected reward of each action, in the decision stage, the agent chooses among her options.

$$\hat{\theta} = \int_0^1 \theta \, Beta_\theta(a, b) \, d\theta,$$

where a and b are, respectively, the numbers of observed defections and cooperations. As a increases, respect to b, $\hat{\theta}$ approaches 1, indicating that the agent believes that the other agent is likely to choose to defect. In this model, each agent has a specific belief distribution for each other agent she interacts with.

Paranoia and *Projection* have a different effect on the inference process. Broadly, *Projection* changes the beta distribution (as if own actions were fragments of observed actions), and *Paranoia* results in sampling unevenly (focusing on the worse outcomes) of the beta distribution.

Projection is a trait that describes how people blame others for their actions. Although an ideal observer constructs this distribution only from priors and observations, to model this characteristic, each time an agent defects, she modifies her beta distribution of beliefs. With *Projection* the actions of the agent impact on the resulting $Beta$ distribution, which she then uses to estimate the probability of defection or cooperation.

Specifically, this is done by changing, whenever the agent defects, the a parameter (which measures the number of observed defections) of the $Beta$ distribution. How much a is changed each time the agent defects is scaled by the parameter *Projection* in such a way that if *Projection* $= 1$ defecting on another agent has the same impact on the $Beta$ distribution than if the other agent defects. If *Projection* $= 0.1$, 10 defections of an agent would have the same effect on its estimated $Beta$ distribution than a defection by the other player, and so on. This same mechanism could be used when the agent cooperates—namely, when an agent cooperates, it changes its belief about others to think that it is more likely that others cooperate too. In this case, the value of *Projection* determines the variation on the b parameter of the $Beta$ distribution after a cooperation.

We explore two different variations of the *Projection* bias. First, when the *Projection* affects only the a parameter if the agent defects. We call this asymmetric *Projection*. Second, when the *Projection* affects both, the a and the b, parameters after the agent defects or cooperates, correspondingly. We call this the symmetric *Projection*.

In both cases, the *Projection* bias changes the $Beta$ parameters that describe the belief about all other agents, not only the one involved in the specific interaction. For example, if agent A defects in a given interaction with B, due to the effect of *Projection*, A will believe that B is more likely to defect. Similarly, A will believe that all other agents are more likely to defect. His own defection has changed his beliefs regarding how all of the other agents he will interact with will behave.

It is important to highlight that, even though this bias yields a wrong value for $\hat{\theta}$, that does not imply that this kind of computation is not optimal. In fact, in models where fairness and others' social preferences are incorporated into a more general utility function, beliefs and action correlate in the same way as in our model (10). That is, the selfish acts are associated with the belief that others are selfish too. The main difference with our model is that, as has been shown experimentally, there is not only correlation but causation. Surprisingly, this causation is in both ways: Belief alters actions, and actions alter belief. Our model explicitly describes this two-way causation using a Bayesian estimation (from beliefs to actions) and by the *Projection* bias (from action to beliefs).

In ref. 10, we showed that the effect of a defecting action produces an average variation of 0.2 in the belief. From this, we can have an estimate of the experimental value of *Projection* of 0.54 (see *SI Appendix*, section 1.4 for details). This result serves as an order of magnitude of realistic values of *Projection* when we

inquire about the impact of this parameter on the propagation of cooperation and defection.

The other bias that we explore, *Paranoia* , acts on how $\hat{\theta}$ is estimated from a distribution. If *Paranoia* $= 0$, the estimation of θ is the mean value of the variable, as it would be in an optimal inference process. When *Paranoia* assumes a nonzero value (here, we only investigate positive values), the estimation is given by this implicit equation.

$$\int_{\hat{\theta}}^{\hat{\theta}^m} Beta_\theta(a, b)\, d\theta = Paranoia. \quad [1]$$

This equation means that the agent shifts its estimate from $\hat{\theta}$ to $\hat{\theta}^m$. The value of *Paranoia* measures the total area under the probability distribution between the optimal and biased estimates $\hat{\theta}$ and $\hat{\theta}^m$. This equation also implies that the impact of *Paranoia* on the estimation of $\hat{\theta}^m$ depends on the shape of the probability distribution. If there is substantial evidence of something, that will be reflected in a narrow and peaked distribution shape, and the effect of *Paranoia* on $\hat{\theta}^m$ will be weak. Instead, if the distribution is wide, meaning that there is insufficient evidence, this same amount of probability will cause a higher distortion in the estimated belief. This mechanism, then, reproduces the empirical fact that higher ambiguity increases the likelihood of being selfish (P2). Note also that *Paranoia* is measured in units of the probability distribution (and hence has one as absolute maximum).

This bias is closely related to the models of reciprocal altruism (10, 21) where belief may be altered too. In these models, changing the belief has a cost that increases as the difference between the unbiased and the biased belief increases. The *Paranoia* bias could be thought of as step function where there is no cost in changing a belief up to some point where the cost, suddenly, approaches infinity. The explicit description of the belief as the mean of a distribution allows us to model P2.

Results

We study how the network evolves when it is contaminated with a fraction of agents with the strategy ALLD (always defect). These agents do not learn or change their behavior in any way; they stubbornly defect independently of the history of actions.

All simulations begin with a network in which agents trust each other. That is, they believe that the expected reward of cooperating is higher than the expected reward for defecting (Fig. 5). Then, we replace a fraction of the regular agents of the network by ALLD agents. These replacements are distributed at random in the sites of the network. Specifically, the question we ask here is how the network parameters convey more resistance or vulnerability to this "infection" process. To do so, we let the network evolve under the influence of ALLD agents and study whether the defection policy extends over the network.

Cascades. First, we study how the system evolves when one ALLD agent is introduced in the network. Under this condition, we measure the fraction of agents that are defecting to each other, which is called the active fraction, S_a. Fig. 2 shows how the S_a changes as *Projection* changes when we use the asymmetric version of the bias. There is a transition in the value *Projection* $= 1.5$ (dashed line in Fig. 2). At this point, one ALLD agent is capable of changing the behavior of a finite fraction of all the other agents that start cooperating and end up defecting with each other. Or, in cascades jargon, it produces a "global cascade" of defection. If the symmetric version of the bias is used, the value of *Projection* changes the vulnerability of the agents only if the value of the *Paranoia* is greater than zero. Then, a similar behavior is observed (*SI Appendix*, section 1.7).

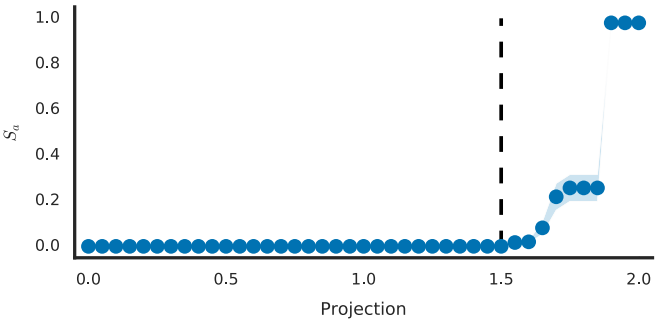

Fig. 2. The active fraction changes as the *Projection* bias increases when only one ALLD agent is present in the network. The point marked with a dashed line is the theoretical value at which the biggest cluster of vulnerable nodes percolates the network. As predicted, this value also indicates the transition from zero to positive values of S_a.

In the case of the asymmetric bias, this value, *Projection* $= 1.5$, can be derived analytically following the method of Watts (22), and the result is in agreement with the numerical simulation (*SI Appendix*, section 1.6).

Percolation and Phase Transitions. Now, we generalize this analysis to a broader situation, where not only one agent, but a fraction, f, of ALLD agents are present in the network. We examine the robustness of the system by analyzing its evolution as two parameters are changed: the fraction, f, and the value of the *Projection* parameter. We calculate the robustness of the network measuring S_a. Here, because there is more than one ALLD, we also incorporate a more refined measure of phase transition referred to as the size of the giant active component, S_{gc}. S_{gc} measures the size of the largest cluster of agents that are defecting to each other. If there is only one agent, S_a and S_{gc} are equivalent.

Fig 3 shows S_{gc} in the $(Projection, f)$ map for the asymmetric version of the bias. It can be seen that the value of S_{gc} increases as the value of f or *Projection* increases. To better understand the transitions in this map, Fig. 4 shows the S_{gc} as a function of f for three values of *Projection* for the asymmetric bias.

When *Projection* $= 0$, the ALLD agents do not change the behavior of the regular agents of the network toward other agents. Then, the change in S_{gc} is due only to the fact that more ALLD agents are defecting in the network. Marked with an arrow in Fig. 4, for this case, a continuous transition can be seen at $f = f_{c1}$. If $f < f_{c1}$, the value of S_{gc} is zero and if $f > f_{c1}$ the value of S_{gc} take positive values. In this situation, the process can be mapped into a standard percolation (*SI Appendix*, section 1.8). The analytical result yields the value $f_{c1} = 0.05$, which is in agreement with the numerical simulations, as can be seen in Fig. 4.

If the value of *Projection* is greater than zero, the actions of the ALLD agents change the belief of the regular agents in the network. In Fig. 4, it can be seen that when *Projection* increases, the value of S_{gc} increases and the value of the critical point f_{c1} moves toward zero.

Interestingly, if the value of *Projection* ≥ 0.8, another kind of transition appears in the system. This second transition at $f = f_{c2}$ is not continuous.

If the symmetric version of the *Projection* bias is used, the results remain equivalent only if the *Paranoia* bias is set to a value higher than 0 (*SI Appendix*, section 1.7). Next, we study more generally for all models, how *Projection* and *Paranoia* can interact to affect the robustness of the network.

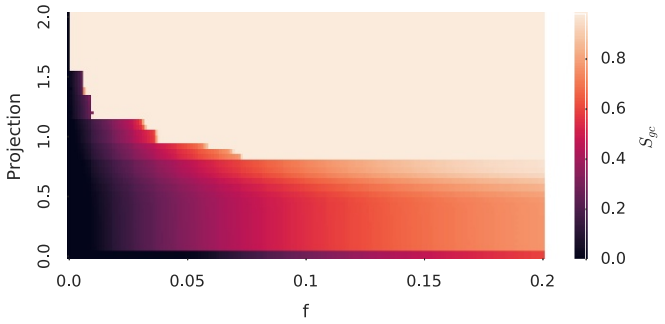

Fig. 3. Heat map of size of the giant component when *Projection* and *f* vary. The heat map shows two different regions: When *Projection* < 0.8, the value of S_{gc} increases continuously with *f*; if *Projection* ≥ 0.8, the value of S_{gc} changes discontinuously at a critical value f_{c2} of *f*.

***Projection* and *Paranoia* Interaction.** The *Paranoia* bias not only affects the dynamics, but the initial condition of the network as well. An agent may believe that others are likely to defect because her internal distribution (based only on the observation of the actions of other agents) has shifted toward defection or because her *Paranoia* bias parameter is greater than zero and her estimation $\hat{\theta}^m$ is higher than what the evidence shows her. Our analytic approach is to investigate the dynamics of a set of networks in which *Paranoia* and the mean of the $Beta$ are covaried, maintaining the same initial $\hat{\theta}^m$, and, hence, we combine priors and *Paranoia* parameters in such a way that they yield the same value of $\hat{\theta}^m$, as shown in Fig. 5. In this way, we can be sure that any difference observed among the simulations is due to a change in the dynamics of the system and not to a change in the initial estimation $\hat{\theta}^m$. Since the parameters of the $Beta$ distribution are integers, this approach imposes a constraint on the values that the *Paranoia* bias can take to keep the initial estimation $\hat{\theta}^m$ unchanged. That is why we investigate only four values of *Paranoia* and not a more fine-grained set of values as we do with the *Projection* bias and the fraction of ALLD agents. But, we also explore the results using different initial values of $\hat{\theta}^m$ and different values of agents' memory in *SI Appendix*, sections 1.9 and 1.10.

The four values of *Paranoia* depicted in Fig. 5 in combination with four values of *Projection* yield a matrix of 16 sets of parameters. As depicted in Fig. 6*A*, for the symmetric version of the *Projection* bias, the evolution of all networks could be grouped into three classes. For some parameters, marked in blue in Fig. 6*B*, the networks were highly cooperative, showing a smooth progression of S_a as a function of f. Interestingly, this is not the case for S_{gc}, which shows a sharp, but continuous, transition at $f = f_{c1}$, as can be seen in Fig. 6*C*. This is evidence of a hidden phase transition, which is not visible in the overall activation, S_a. A second class, colored green in Fig. 6, revealed a discontinuous transition in both, S_a and S_{gc}. For $f < f_{c2}$, the network remains mostly cooperative and transitions abruptly to defection for $f > f_{c2}$. For other parameters, instead, the network ends in a high defection state. Introducing a single ALLD agent is enough to shift the network toward total defection (red regions in Fig. 6*A*). We refer to these three classes as (*i*) high cooperation, (*ii*) bistable, and (*iii*) high defection, respectively. These classes are also present when the initial value of $\hat{\theta}^m$ is below the threshold cooperation (*SI Appendix*, section 1.9), and then these results are robust and do not depend on this specific initial value. As in the previous simulations, a similar map with the same type of states and transitions is found if we use the symmetric version of the *Projection* bias (*SI Appendix*, section 1.7).

A specific analysis of which priors yield to different regimes of stability (Fig. 6*A*, first row) indicates that when the *Projection* parameter is set to 0, regardless of the value of paranoia, the networks belong to the high-cooperation class. This means that the society is robust under the inception of a fraction of ALLD agents. For moderate values of *Projection* (0.25) which are lower than the estimated experimentally from ref. 10, the network is in high cooperation for low values of *Paranoia* and shifts to bistability for values of *Paranoia* of 0.36. For this level of *Projection*, even for maximum values of *Paranoia* (this is very close to the strict maximum since greater values of *Paranoia* are incompatible with a network that begins in full cooperation), the network never is in the high-defection class. For higher values of *Projection* (0.75), which are slightly above experimental estimates, the network displays the three different behaviors, depending on the value of *Paranoia*. If it is zero, then the network is in high cooperation, and, as the value of *Paranoia* increases, the network behaves as a bistable system and, finally, is in a high defection state.

Discussion

By using a Bayesian updating rule in an agent-based simulation, we were able to model how two specific cognitive biases, *Projection* and *Paranoia*, impact the decision-making process. Then, we use this to inquire how these parameters affect the propagation of defection started by a set of ALLD agents.

Our first result is that if only one ALLD agent is introduced in the network, there is a threshold in the value of *Projection* up to which the agents keep cooperating with each other. If the value of *Projection* is higher than the threshold, then a positive fraction of the agents start to defect. In the case of the asymmetric version of the *Projection* bias, the threshold could be deduced analytically and coincides with the value found in our simulations.

Then, we find two kinds of transitions when a fraction of ALLD agents is introduced. If the value of *Projection* is low enough, there is only one transition at f_{c1} where the size of the giant component of defecting agents goes from being zero to being greater than zero, a continuous transition. If the value of *Projection* is >0.8, we find another transition at f_{c2}, where the fraction of agents that are defecting to each other jumps in a noncontinuous manner. These two types of transitions resemble the transitions present in the bootstrap percolation process (23). Even though there are differences between the two processes, the appearance of the same two kinds of transitions suggests that there is a connection between them. These same kinds of transitions have been observed in the spread of extreme opinions (24). In ref. 24, Ramos et al. used a k-core model to explain their results which are closely related to bootstrap percolation (23).

Additionally, we study how the effect of *Paranoia* and *Projection* interact in the spreading of the defections produced by fraction ALLD agents. The main result is that, if the *Projection* bias is set to zero, the *Paranoia* bias does not make the network

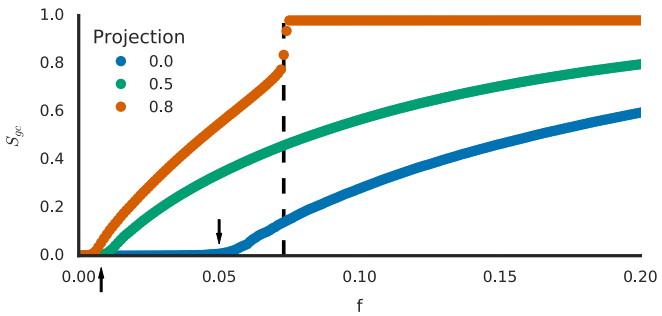

Fig. 4. Examples of *f* vs. S_{gc} for three values of *Projection*. The arrows show the point f_{c1} at which S_{gc} changes from zero to positive values and the dashed line indicates the discontinuous transition at f_{c2}.

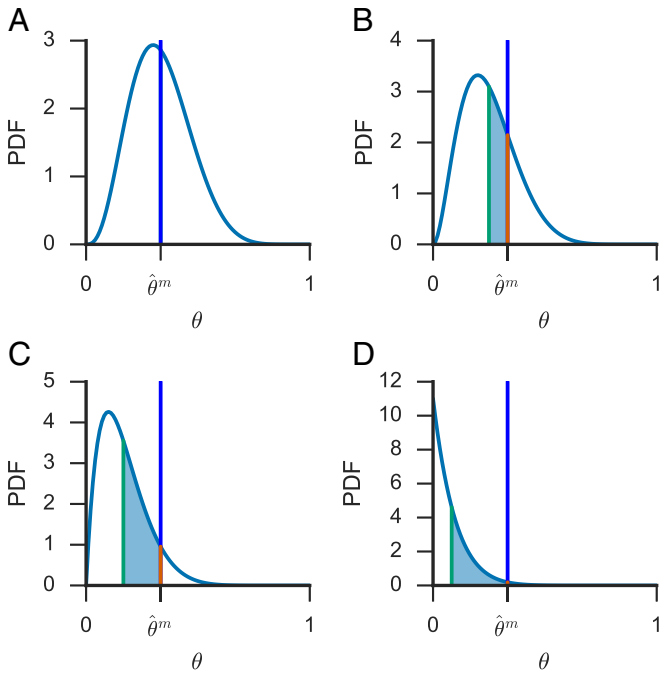

Fig. 5. Belief distribution for four parameter sets of *Paranoia*, *a* and *b*. In *A*, we plot a $Beta_\theta(4, 8)$ with *Paranoia* = 0; in *B*, a $Beta_\theta(3, 9)$ with *Paranoia* = 0.22; in *C*, a $Beta_\theta(2, 10)$ with *Paranoia* = 0.36; and in *D*, a $Beta_\theta(1, 11)$ with *Paranoia* = 0.37. The green line shows the mean value of the probability of defection, $\hat{\theta}$, while the red line shows the manipulated mean value $\hat{\theta}^m$. The area under the distribution between $\hat{\theta}$ and $\hat{\theta}^m$ is, by definition, the value of the *Paranoia* parameter (Eq. **1**). The mean of the distribution (or, equivalently, the values of *a* and *b*) and the *Paranoia* parameters have been chosen in such a way that they compensate each other and lead to the same manipulated mean $\hat{\theta}^m$. The blue vertical line shows the limit above which the agent believes that the maximum reward is achieved by defecting and below which the agent believes that the maximum reward is obtained by cooperating. If $\hat{\theta}^m$ is higher than this value, then the agent decides to defect; if, on the contrary, $\hat{\theta}^m$ lays to the left of the blue vertical line the agent chooses to cooperate. Under these four conditions, the agents, initially, cooperate with each other. PDF, probability density function.

less robust to the infection of ALLD agents. Only if it is combined with the *Projection* bias does it weaken the network.

A distinguishing characteristic of our model is that it is built upon a network, and the agents can choose a different action for different contacts. In their seminal work, Nowak and May (25) use a regular lattice where each agent plays only one action in each step against their contacts. This model is suitable to study the evolution of cooperation, but we believe that, to investigate cooperation at a cultural level, we have to add the possibility of acting differently with different contacts. We used a static random network, but this is just another parameter of the model that can be modified. For example, a network whose degree distribution follows a power law, or a small-world network, could be used (26, 27). It is also possible to add dynamics to the network, allowing the interaction to change iteration after iteration. Additionally, the ALLD agents could be placed, not randomly, but rather in high-centrality nodes or in given k-shell within the network. This variation could be used to investigate the effect that highly connected people's behavior might have on the rest of the community.

The long-term motivation for this study is to understand why different societies may converge to different policies of cooperation and defection. One particular case which we are interested in, and which was the motivation for the work on the corruption game (10), was how this might impact in different degrees of corruption. Recently, Gächter and Schulz (28) using an anonymous die-rolling experiment (29) showed that there is a correlation between individual traits and global corruption. The corruption game (10) was measured in the United States and Argentina, where there are very different indices of corruption and results were not very different. This suggests that *Projection* is not the most likely psychological bias to account for. This is consistent with our findings that a wide range of behaviors is only observed for experimentally observed values of *Projection*. Within this range, variability in *Paranoia* may explain bifurcations between societies that (with similar initial states) converge to defection or cooperation, or similarly in our interest to a high or low level of corruption. Another source of variability could be the effective impact of the ALLD agents in a society. Some governments have more efficient institutions which control the acts of the ALLD agents more effectively. Small changes in this control could lead to a different effective fraction of ALLD agents which, in a society that is in a bistable state, will lead to convergence to cooperation or defection.

Our work builds on, and links, two different fields of behavioral sciences. On the one hand, a tradition that has studied agents with simple, yet effective, strategies, and which of these strategies prevails in different contexts. Under this approach, using unbiased agents, the appearance and prevalence of corruption have been investigated (17, 18, 20). The conundrum of the emergence of cooperation has been addressed under this framework, too (30, 31). On the other hand, our work builds on a Bayesian approach to decision making that has sought to inquire how priors and evidence are used to generate beliefs and guide actions (32, 33). This work can be seen as a mixture of these two traditions, where we can then ask how low-level psychological constructs which affect the inferential process (micro) have an impact on the large-scale organization of societies (macro). To build this bridge between these traditions, we use network theory, and the process that emerged under this framework was very similar to an already-known network process: bootstrap percolation. We did not impose this process, but it was the result of the cognitive model of the agents. Under this framework, we find that small variations on cognitive biases could have a significant impact on the average behavior of all of society. According to our model, society-level phenomena, including the broken window theory (34) and the broad range in the degree

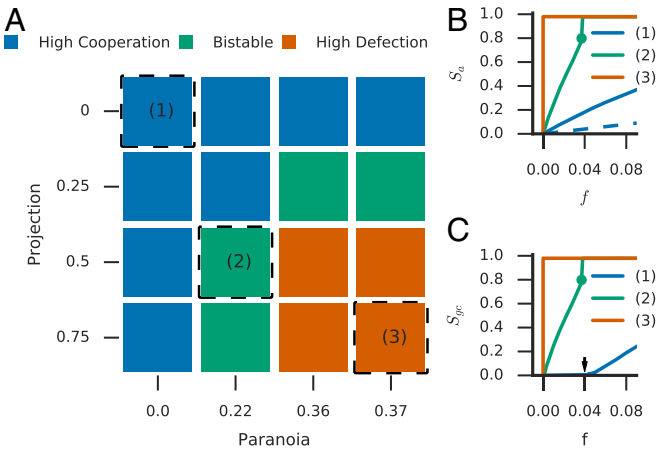

Fig. 6. (*A*) Classification of the network according to its stability for 16 different parameters. (*B* and *C*) S_a and S_{gc} as a function of the fraction of ALLD agents. The dashed line in *B* shows the expected S_a due only to the presence of the ALLD agents and assuming that they do not interact with each other. It can be seen that the high-cooperation region does not have a sharp transition, while the bistable ones do, and the high-defection one has a large S_a and S_{gc} for any nonzero values of the fraction of ALLD agents.

of corruption among countries, is the consequence of cognitive dissonance reduction mechanisms such as *Projection* and *Paranoia*.

ACKNOWLEDGMENTS. We thank Nicolas E. Stier-Moses for his helpful comments. This research was supported by Consejo Nacional de Investigaciones Científicas y Técnicas and Fondo para la Investigación Científica y Tecnológica Grant PICT 2013 N°1653. M.S. is sponsored by the James McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition-Scholar Award. H.A.M. is sponsored by NIH National Institute of Biomedical Imaging and Bioengineering Grant R01EB022720 and NSF Information and Intelligent Systems Grant 1515022.

1. Kahneman D, Knetsch JL, Thaler RH (1986) Fairness and the assumptions of economics. *J Bus* 59:S285–S300.
2. Konow J (2000) Fair shares: Accountability and cognitive dissonance in allocation. *Am Econ Rev* 90:1072–1091.
3. Güth W, Schmittberger R, Schwarze B (1982) An experimental analysis of ultimatum bargaining. *J Econ Behav Organ* 3:367–388.
4. Fehr E, Schmidt KM (1999) A theory of fairness, competition, and cooperation. *Q J Econ* 114:817–868.
5. Festinger L (1962) *A Theory of Cognitive Dissonance* (Stanford Univ Press, Stanford, CA), Vol 2.
6. Festinger L, Carlsmith JM (1959) Cognitive consequences of forced compliance. *J Abnorm Psychol* 58:203–210.
7. van Veen V, Krug MK, Schooler JW, Carter CS (2009) Neural activity predicts attitude change in cognitive dissonance. *Nat Neurosci* 12:1469–1474.
8. Izuma K, et al. (2010) Neural correlates of cognitive dissonance and choice-induced preference change. *Proc Natl Acad Sci USA* 107:22014–22019.
9. Rabin M (1995) *Moral Preferences, Moral Constraints, and Self-Serving Biases* (*Department of Economics, Univ of California, Berkeley*).
10. Di Tella R, Perez-Truglia R, Babino A, Sigman M (2015) Conveniently upset: Avoiding altruism by distorting beliefs about others' altruism. *Am Econ Rev* 105:3416–3442.
11. Ariely D, Loewenstein G, Prelec D (2003) Coherent arbitrariness: Stable demand curves without stable preferences. *Q J Econ* 118:73–105.
12. Ariely D, Norton MI (2008) How actions create—not just reveal—preferences. *Trends Cogn Sci* 12:13–16.
13. Grujic J, Fosco C, Araujo L, Cuesta JA, Sanchez A (2010) Social experiments in the mesoscale: Humans playing a spatial prisoner's dilemma. *PLoS ONE* 5:e13749.
14. Gracia-Lazaro C, et al. (2012) Heterogeneous networks do not promote cooperation when humans play a Prisoner's dilemma. *Proc Natl Acad Sci USA* 109:12922–12926.
15. Mazar N, Amir O, Ariely D (2008) The dishonesty of honest people: A theory of self-concept maintenance. *J Mark Res* 45:633–644.
16. Bear A, Rand DG (2016) Intuition, deliberation, and the evolution of cooperation. *Proc Natl Acad Sci USA* 113:936–941.
17. Hauk E, Saez-Marti M (2002) On the cultural transmission of corruption. *J Econ Theor* 107:311–335.
18. Tirole J (1996) A theory of collective reputations (with applications to the persistence of corruption and to firm quality). *Rev Econ Stud* 63:1–22.
19. Bisin A, Verdier T (2001) The economics of cultural transmission and the dynamics of preferences. *J Econ Theor* 97:298–319.
20. Sah RK (2007) Corruption across countries and regions: Some consequences of local osmosis. *J Econ Dyn Control* 31:2573–2598.
21. Levine DK (1998) Modeling altruism and spitefulness in experiments. *Rev Econ Dyn* 1:593–622.
22. Watts D (2002) A simple model of global cascades on random networks. *Proc Natl Acad Sci USA* 99:5766–5771.
23. Baxter GJ, Dorogovtsev SN, Goltsev AV, Mendes JFF (2010) Bootstrap percolation on complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 82:1–9.
24. Ramos M, et al. (2015) How does public opinion become extreme? *Sci Rep* 5:10032.
25. Nowak MA, May RM (1992) Evolutionary games and spatial chaos. *Nature* 359: 826–829.
26. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* 393:440–442.
27. Newman ME, Strogatz SH, Watts DJ (2001) Random graphs with arbitrary degree distributions and their applications. *Phys Rev E Stat Nonlin Soft Matter Phys* 64:026118.
28. Gächter S, Schulz JF (2016) Intrinsic honesty and the prevalence of rule violations across societies. *Nature* 531:496–499.
29. Fischbacher U, Föllmi-Heusi F (2013) Lies in disguise-an experimental study on cheating. *J Eur Econ Assoc* 11:525–547.
30. Nowak MA, Sasaki A, Taylor C, Fudenberg D (2004) Emergence of cooperation and evolutionary stability in finite populations. *Nature* 428:646–650.
31. Rand DG, Tarnita CE, Ohtsuki H, Nowak MA (2013) Evolution of fairness in the one-shot anonymous ultimatum game. *Proc Natl Acad Sci USA* 110:2581–2586.
32. Griffiths TL, Tenenbaum JB (2006) Optimal predictions in everyday cognition. *Psychol Sci* 17:767–773.
33. Austerweil JL, Gershman SJ, Tenenbaum JB, Griffiths TL (2015) Structure and flexibility in Bayesian models of cognition. *Oxford Handbook of Computational and Mathematical Psychology* (Oxford Univ Press, New York), pp 187–208.
34. Keizer K (2008) The spreading of disorder. *Science* 322:1681–1685.