1 Supplementary Information

1.1 The Payoff Matrix

The payoff matrix of a Prisoners' Dilemma (PD) has the form:

	C	D
C	R	S
D	T	P

(1)

where C accounts for cooperation, D for defection and $T > R > P > S$. This payoff matrix leads to the dominant strategy of defect because no matter which option the other player chooses it is always more profitable to defect. Evidence shows that subjects often interpret the PD game as Coordination Game [1]. An explanation for this perception is that people have an inequity aversion that reduces the off-diagonal elements of the matrix [2]. Depending on the parameters of this model the PD payoff matrix could turn into a Coordination Game payoff matrix, where the most profitable action depends on the action of the other player. We follow this path and use the following payoff matrix:

$$P = \begin{array}{c|c|c} & C & D \\ \hline C & 1.3 & 0 \\ \hline D & 1.1 & 0.4 \end{array} \tag{2}$$

We keep the C and D notation. While it is true that not all subjects interpret the PD as a Coordination Game, in using this payoff matrix we endowed our model with agents of the more cooperative kind, the ones that are willing to cooperate if the other player will also do it. Precisely for this reason, the agents that we use are the ones that are more resilient to the biases that we study.

With the payoff matrix of Eq. (2) we can compute the action with the highest expected reward given the likelihood of defection of the other player:

$$E(R|C) = P_{CC} + (P_{CD} - P_{CC})\hat{\theta}$$

$$E(R|D) = P_{DC} + (P_{DD} - P_{DC})\hat{\theta}$$

where $E(R|C)$ and $E(R|D)$ are the expected reward of cooperating and the expected reward of defect correspondingly, and $\hat{\theta}$ is the probability that the other player chooses to defect. There is a value of $\hat{\theta}^* = \frac{1}{3}$ for which these two expected rewards are equal. If the estimation of θ is greater than $\hat{\theta}^*$ then $E(R|D) > E(R|C)$ and if it is smaller then $E(R|D) < E(R|C)$.

1.2 The Beta Distribution

The functional form of the probablity density function of the beta distribution for $0 \leq x \leq 1$ is:

$$\beta(x; a, b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a, b)}$$

www.pnas.org/cgi/doi/10.1073/pnas.1803438115

where B stands for the beta function (not the beta distribution) which is a normalization constant, $a > 0$ and $b > 0$.

If both parameters satisfy that $a > 1$ and $b > 1$, then the distribution has a bell shape, as can be seen in Fig. 5, main text. The most important fact of the beta distribution for this work is that is the conjugate prior of the Bernoulli distribution. This fact implies that using the Bayes Theorem to compute a posterior with a Bernoulli likelihood and Beta prior yields a Beta posterior– allowing the use of the posterior as a new prior in an iterative fashion.

1.3 The Bayesian Updating Rule

To estimate the future actions of the other agents, each agent assumes that the other agent will defect with probability θ. This assumption is like thinking that the other agent chooses her options randomly using a Bernoulli distribution with parameter θ. To estimate the parameter of a Bernoulli distribution, we use a $Beta$ distribution. This option is natural because the $Beta$ distribution is the conjugate prior for the Bernoulli distribution, which means that the posterior distribution follows the same parametric form as the prior [3, 4]. The $Beta$ distribution has two parameters a and b which we associate with the number of times an agent experienced that another agent cooperated or defected to him. For example, if an agent has a prior distribution $Beta_\theta(5, 7)$, then it will estimate that the probability that the other agent defect on him is $\frac{7}{12}$, that is the number of defections observed divided the total number of observations. If, when playing, the other agent defect to him, he will update his prior using the Bayes theorem:

$$P(\theta|D) \propto Bi_\theta Beta_\theta(5, 7),$$

where Bi_θ is the Bernoulli distribution, which in this case is the likelihood of observing a defection, and $P(\theta|D)$ is the conditional probability of θ after observing a defection. It turns out that $P(\theta|D) = Beta_\theta(6, 7)$, then the Bayesian way of updating the priors after the observations is incrementing the parameter a or b in one unit depending on if the agent observed a defection or a cooperation action, correspondingly.

A problem that arises with this model is that the agents would have infinite memory because each time they play they update their prior. This is a problem for two reasons: first, the beta approaches a delta function, and numerical problems arise, and second, the memory that people uses in this game is not infinite [5]. To solve this problem, we limit the number of observation to 10 in a First In First Out fashion. That is when a new observation is done the oldest one is deleted. To avoid improper priors, we also add one C and one D observation, in such a way that the parameters of the $Beta$ distribution always sum 12.

1.4 The Empirical Value of *Projection*

In [6] we performed an experiment where subjects, playing a game among each other, reported their belief about other choosing the most selfish action (an

action that is similar to a defect in a PD game). When the subjects were also allowed to choose a selfish action against the other player (similar to a defect in a PD game), then their belief about the other changed significantly in 0.2. This change means that acting in a selfish manner (or defecting) change the belief about the others. The subjects that were allowed to behave more selfishly performed on average 4.47 more selfish actions than the control group, which lead to an average of 0.045 difference per selfish action. Given that we used a *Beta* distribution to describe the belief of agents regarding the probability of others defecting at her we should compute what change in the parameters of the *Beta* will lead to a variation in the mean value of 0.045. That is:

$$\int_0^1 \theta Beta_\theta(a + Projection, b)d\theta = \int_0^1 \theta Beta_\theta(a, b)d\theta + 0.045$$

$$\frac{a + Projection}{a + b} = \frac{a}{a + b} + 0.045$$

$$\frac{Projection}{12} = 0.045$$

$$Projection = 0.54$$

1.5 Properties of the Erdős-Rényi Network

To build the network we used the NetworkX python library [7]. The Erdős-Rényi network was built with 10^5 nodes and a 10^{-5} probability of edge creation, resulting in an average degree of 4. After this, we remove the self-edges, and we kept only the largest connected component. The final network has 97964 nodes and an average degree of 4.07.

1.6 Computation of the Cascade Condition

Given that the estimated probability of defect of another agent, $\hat{\theta}$, is the mean value of a beta-distributed random variable, its value is computed in the following way:

$$\hat{\theta} = \frac{a}{a + b}$$

If we add the *Projection* bias, the value of a is replaced by

$$a \leftarrow \left(a + Projection\frac{a_{own}}{k}\right)\frac{12}{12 + Projection\frac{a_{own}}{k}}$$

$$b \leftarrow b\frac{12}{12 + Projection\frac{a_{own}}{k}}$$

where a_{own} is the number of defection actions done by the agent herself, and k is her number of neighbors. The fraction, $\frac{12}{12+Projection\frac{a_{own}}{k}}$ is a normalization factor to keep $a + b = 12$. Then, the biased estimated probability of defection is computed as:

$$\hat{\theta}^m = \frac{a + Projection\frac{a_{own}}{k}}{12 + Projection\frac{a_{own}}{k}}$$

Finally, setting $a = 1$, which is the best case scenario and $a_{own} = 12$, which is the value you get after enough interactions against an ALLD Agent, this yields the equation:

$$\hat{\theta} = \frac{1 + \frac{12}{k}Projection}{12 + \frac{12}{k}Projection} \tag{3}$$

The requirement for the agents to choose to defect, that is, to change their initial behavior of deciding to cooperate, is that $\hat{\theta} > \frac{1}{3}$. This expression arises from the values of the payoff matrix of the game (see SI 1.1). Then, the condition for an agent to change its behavior toward other BA agents is:

$$k < \frac{8}{3}Projection$$

The agents that are under this condition are called vulnerable, because, one ALLD agent in his neighborhood will change his behavior. Given that the agents are on a random network it is possible to calculate which proportion of them are in this vulnerable condition. If the fraction of vulnerable agents percolates the network (that is that they are a finite fraction), the cascades are possible. In our model, this happens if:

$$Projection > 1.5 \tag{4}$$

If we assume a symmetric *Projection*, its value does not change the vulnerability of the agents. Nevertheless, if the value of the *Paranoia* bias is greater than zero, the computational simulations show a similar behavior of the system.

1.7 Analysis of the Model with Symmetrical Projection

Based on empirical results, we propose a model where selfish action make people think that others are selfish too. However, settings could affect empirical results, and in this section, we analyze the situation where altruistic actions also change the agents' belief so that cooperating make them thing that others are more likely to cooperate. We repeat our analysis but with this symmetric *Projection* bias.

First, we study what happens when only one ALLD agent is introduced. As can be seen in Fig. 1, if the *Paranoia* bias is set to 0, then the effect of the *Projection* bias is not present. However, if the *Paranoia* bias is 0.36 the effect appears. This pattern is due to the interaction effect that is also present in the original model.

Following the same analysis that in the main text, now we study the effect of introducing a fraction of ALLD agents. Given that the *Projection* bias does not produce any effect by itself, but in combination with *Paranoia*, we study

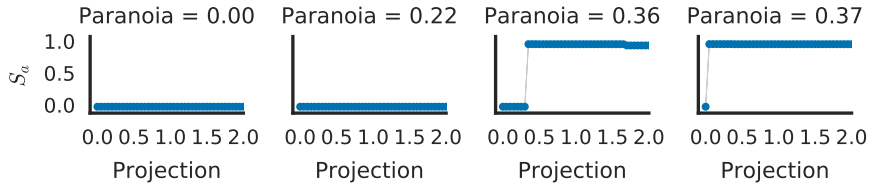


Figure 1: Fraction of the active agents as the *Projection* bias increases for four different values of *Paranoia* when only one ALLD Agent is present in the network.

the case where $Paranoia = 0.22$. In Fig. 2 it can be seen a similar pattern to the one in Fig. 3 in the main text and Fig. 4 in the main text with two kinds of phase transitions.

Finally, and following the analysis of the main text, we simulate the evolution of the system for sixteen combinations of parameters and we classify them in the same three categories of the main text. The results can be seen in Fig. 3. This map is similar to the map in Fig. 6 in the main text, but the regions are shifted towards cooperations as expected due to the symmetry of this new model. Even though the cooperative actions are also affecting the belief of the agents towards cooperation, there are regions of high defection because if the interaction of the *Projection* bias with the *Paranoia* one.

1.8 Percolation Threshold Without Biases

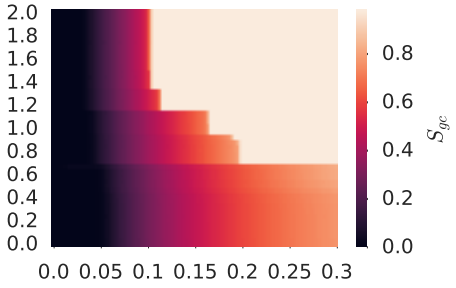
If the agents in the network do not have any bias ($Projection = 0$ and $Paranoia = 0$), the percolation threshold, f_{c1} can be derived analytically. The problem mappable to the the percolation process which states that the percolation threshold for an Erdős-Rényi network is:

$$f_c = \frac{1}{\langle k \rangle}$$

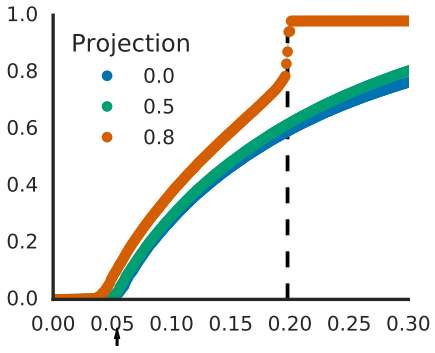
where $\langle k \rangle$ is the average degree of the nodes in the network.

Our system is different because two ALLD agents could be part of the same cluster of defection even if they are not linked by an edge in the network. For example, if two ALLD agents are connected to the same regular agent of the network, they will be part of the same cluster because the three agents would be defecting to each other. Since, in average, each ALLD agents has $\langle k \rangle$ neighbors, each other ALLD has $\langle k \rangle + 1$ (the neighbors and itself) possible ways of forming a cluster with it. Then, the fraction f of ALLD agents should be scale for the factor $\langle k \rangle + 1$, which yields the critical value:

$$f_{c1} = \frac{1}{(\langle k \rangle + 1)\langle k \rangle}$$



(a)



(b)

Figure 2: **(a)** Heat map of Size of the giant component when *Projection* and f vary. The heat map shows two different regions: when *Projection* < 0.7 the value of S_{gc} increases continuously with f, if *Projection* ≥ 0.7 the value of S_{gc} changes discontinuously at a critical value f_{c2} of f. **(b)** 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}.

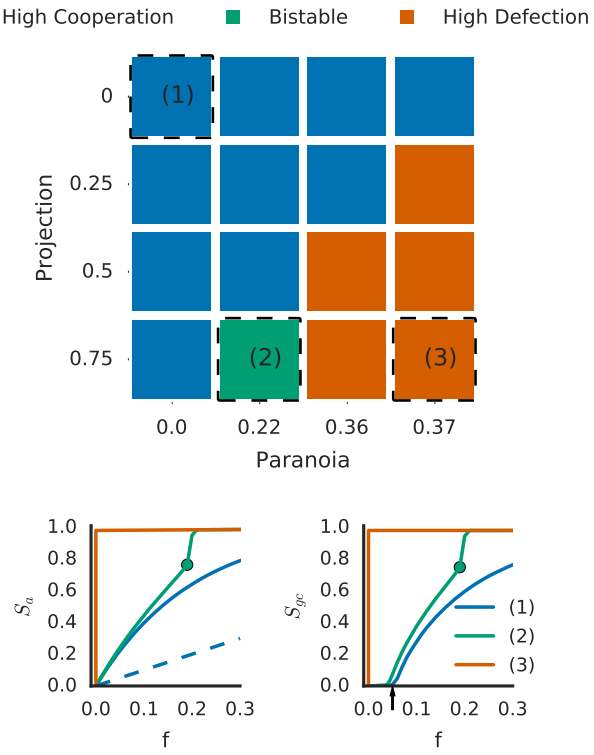


Figure 3: Classification of the network according to its stability for 16 different parameters. The two insets show S_a and S_{gc} as a function of the fraction of ALLD agents. The dashed line in the left inset shows the expected S_a due only to the presence of the ALLD agents. It can be seen that the stable system does not have a sharp transition while the bistable ones do and the unstable one has a large S_a and S_{gc} for any non-zero values of the fraction of ALLD agents.

Given that in our simulations $\langle k \rangle = 4$, the analytical value of the threshold is $f_{c1} = 0.05$.

1.9 Sensitivity to Initial Value of $\hat{\theta}^m$

In the main text, the simulations were performed with a specific initial value of $\hat{\theta}^m$. As can be seen in Fig. 5 main text, the value used in the main text is the maximum $\hat{\theta}^m$ that led to a cooperative behavior. This choice was made to maximize the number of comparisons, but in this section, we show that our conclusions hold even if this value changes.

We explore the results when the initial conditions are such that the initial value of $\hat{\theta}^m$ is lower than the threshold of cooperation. We investigate the

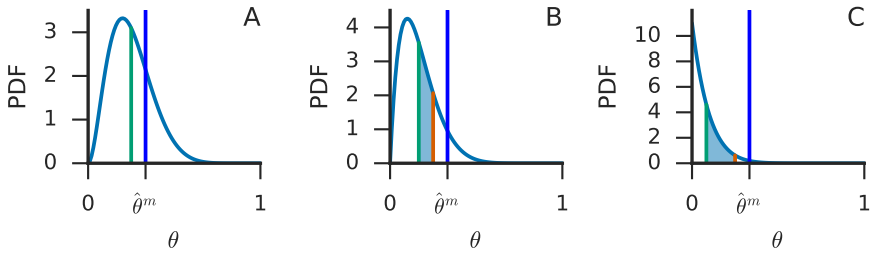


Figure 4: Belief distribution for tree parameter sets of *Paranoia*, a and b. In A we plot a $Beta_\theta(3, 9)$ with $Paranoia = 0$, in B a $Beta_\theta(2, 10)$ with $Paranoia = 0.23$, and in C a $Beta_\theta(1, 11)$ with $Paranoia = 0.34$. The green line shows the mean value of the probability of defection, $\hat{\theta}$, 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 the value of the *Paranoia* parameter. 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. Under these three conditions, the agents, initially, cooperate with each other.

evolution of the system when the initial value of $\hat{\theta}^m = \frac{1}{4}$, $\hat{\theta}^m = \frac{1}{6}$ and $\hat{\theta}^m = \frac{1}{12}$, as can be seen in Fig. 4, Fig. 5 and Fig. 6 accordingly.

As can be seen in Fig. 7, when the initial value of $\hat{\theta}^m = \frac{1}{4}$, the system could behave in the same three ways as we showed in the main text. The main difference between the results of the main text is that now the values of *Projection* that set the system in a High Defection state are greater than the ones in the main text which is in agreement with the fact that now the maximum level of *Paranoia* is smaller. In Fig. 8 and Fig. 9, we show the same results but for the initial value $\hat{\theta}^m = \frac{1}{6}$ and $\hat{\theta}^m = \frac{1}{12}$, accordingly.

We also performed the simulations for the model with symmetric version of the projection bias. The results of for the symmetrical version of the bias, with initial value $\hat{\theta}^m = \frac{1}{4}$, can be seen in Fig. 10 In this case, we find the same three possible states as well. The results of for the symmetrical version of the bias, with initial value $\hat{\theta}^m = \frac{1}{6}$, can be seen in and Fig. 11. In this case, as we already know that the symmetric version of *Projection* does not affect the system if $Paranoia = 0$ we do not show results with initial value $\hat{\theta}^m = \frac{1}{12}$ and $Paranoia = 0$. For the symmetric version of the *Projection* bias, when the initial belief is $\hat{\theta}^m = \frac{1}{6}$, and therefore we can not set the *Paranoia* parameter to a greater value than 0.25, we do not find the High Defection state. This lack of a High Defection state is due to the effect of *Projection* on the cooperative behavior. As stated in the main text, the symmetric version

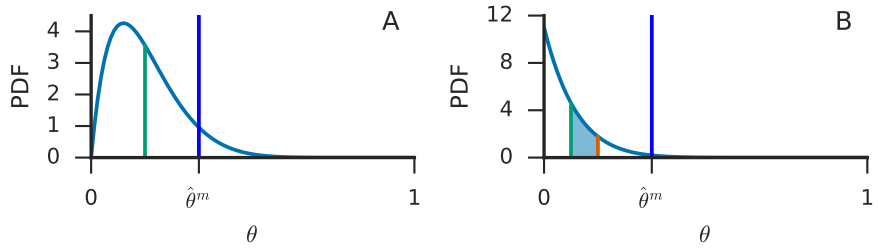


Figure 5: Belief distribution for two parameter sets of *Paranoia*, a and b. In A we plot a $Beta_\theta(2, 10)$ with $Paranoia = 0$, and in B a $Beta_\theta(1, 11)$ with $Paranoia = 0.25$. The green line shows the mean value of the probability of defection, $\hat{\theta}$, while the red line shows the manipulated mean value $\hat{\theta}^m$. The area under the distribution between $\hat{\theta}$ and $\hat{\theta}^m$ is the value of the *Paranoia* parameter. 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. Under these two conditions, the agents, initially, cooperate with each other.

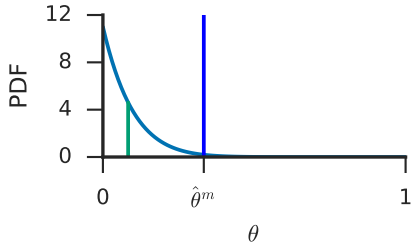


Figure 6: Belief distribution for $Beta_\theta(1, 11)$ with $Paranoia = 0$. The green line shows the mean value of the probability of defection, $\hat{\theta}$. The blue vertical line shows the limit above which the agent believes that the maximum reward is achieved by defecting and below which the agent believes that the maximum reward is obtained by cooperating. Under this condition, the agents, initially, cooperate with each other.

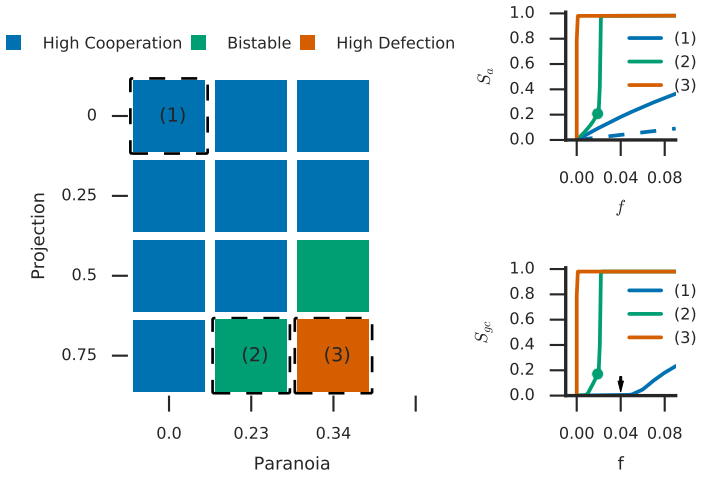


Figure 7: Classification of the network according to its stability for 12 different parameters. The simulations were performed with the asymmetrical version of the *Projection* bias. The two insets show S_a and S_{gc} as a function of the fraction of ALLD agents. The dashed line in the left inset shows the expected S_a due only to the presence of the ALLD agents and assuming that they do not interact with each other. The initial belief of the agents was set such that of $\hat{\theta}^m = \frac{1}{4}$ in every simulation.

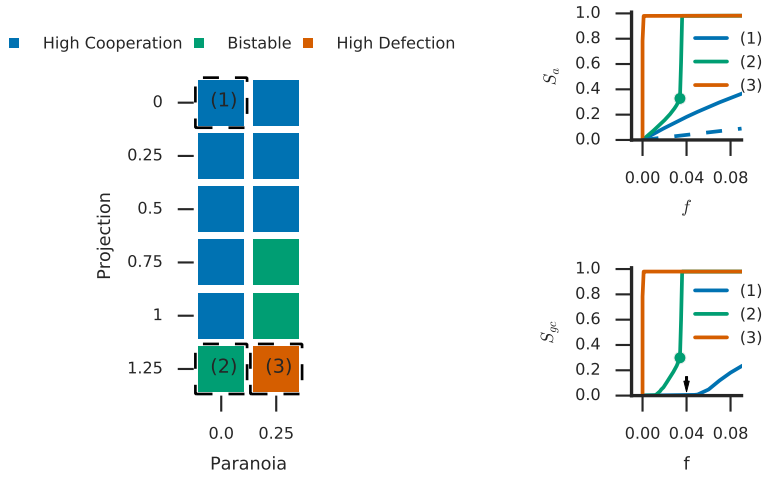


Figure 8: Classification of the network according to its stability for 12 different parameters. The simulations were performed with the asymmetrical version of the *Projection* bias. The two insets show S_a and S_{gc} as a function of the fraction of ALLD agents. The dashed line in the left inset shows the expected S_a due only to the presence of the ALLD agents and assuming that they do not interact with each other. The initial belief of the agents was set such that of $\hat{\theta}^m = \frac{1}{6}$ in every simulation.

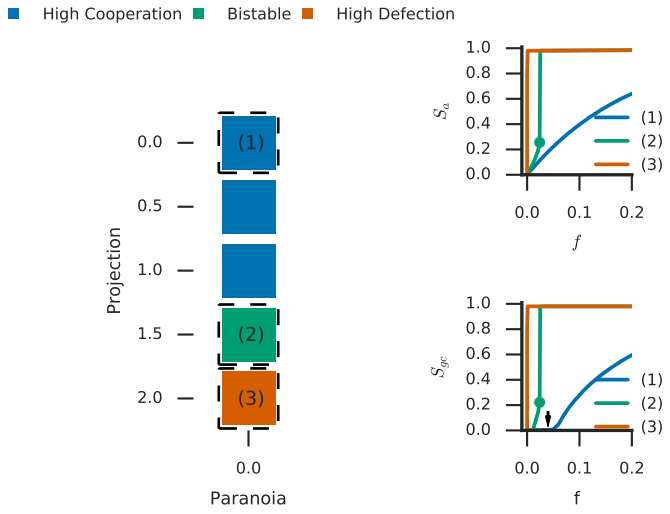


Figure 9: Classification of the network according to its stability for 5 different parameters. The simulations were performed with the asymmetrical version of the *Projection* bias. The two insets show S_a and S_{gc} as a function of the fraction of ALLD agents. The dashed line in the left inset shows the expected S_a due only to the presence of the ALLD agents and assuming that they do not interact with each other. The initial belief of the agents was set such that of $\hat{\theta}^m = \frac{1}{12}$ in every simulation.

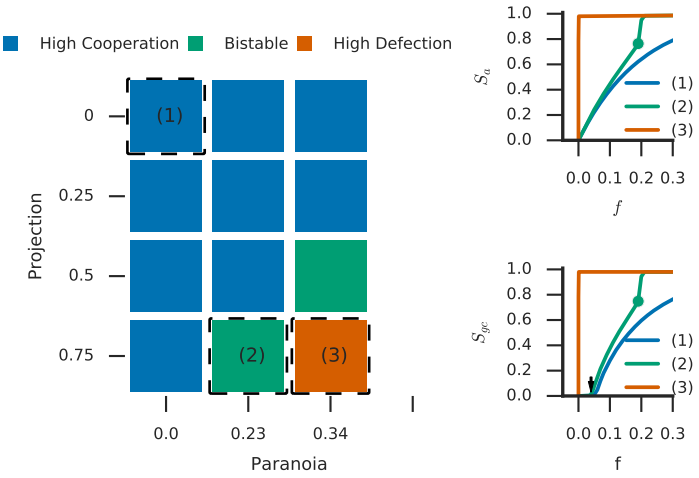


Figure 10: Classification of the network according to its stability for 12 different parameters. The simulations were performed with the symmetrical version of the *Projection* bias. The two insets show S_a and S_{gc} as a function of the fraction of ALLD agents. The dashed line in the left inset shows the expected S_a due only to the presence of the ALLD agents and assuming that they do not interact with each other. The initial belief of the agents was set such that of $\hat{\theta}^m = \frac{1}{4}$ in every simulation.

of *Projection* does not affect the system by itself, but only when *Paranoia* is greater than zero. In this simulation, the value of *Paranoia* is not high enough to, in combination with *Projection*, change the system to a High Defection state. In fact, if the *Projection* bias is great enough, its effect is to isolate the ALLD agents. This isolation happens because the regular agents of the network start with a cooperative behavior, then the *Projection* mechanism biases them even more towards cooperation, counteracting the effect of the observation of the defection actions by the ALLD. But, the values needed to counter the effect of the ALLDare beyond our estimation of its empirical value.

1.10 Sensitivity to the Memory Parameter

In the main text, we set the memory of the agent to 10. This parameter constraints the family of beta distributions that the agents use to estimate θ. The a and b parameters of these beta distributions must be such that $a + b = 12$. Now, we investigate the evolution of the system when we set the memory of the agents to a value of 12. This value constrain impose the contrain $a + b = 14$. These belief distributions are shown in Fig. 12. As the sum of the a and b parameters increases, the variance of the beta distribution decreases, then, we expect the effect of *Paranoia* to be diminished.

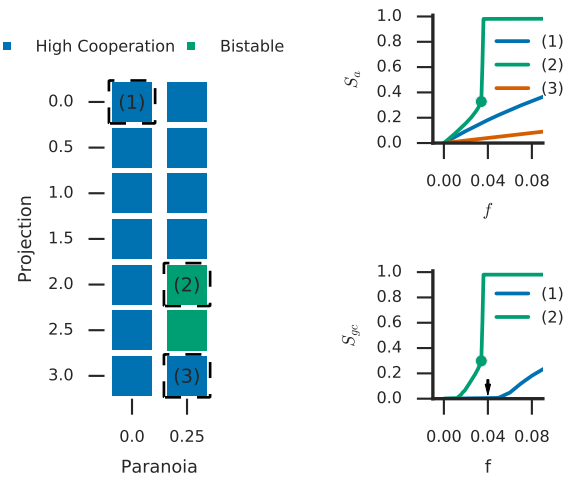


Figure 11: Classification of the network according to its stability for 14 different parameters. The simulations were performed with the symmetrical version of the *Projection* bias. The two insets show S_a and S_{gc} as a function of the fraction of ALLD agents. The initial belief of the agents was set such that of $\hat{\theta}^m = \frac{1}{6}$ in every simulation. The third set of results is not shown in the bottom inset because the non zero values of S_{gc} are in the region $f > 0.25$, as predicted by the standard percolation theory.

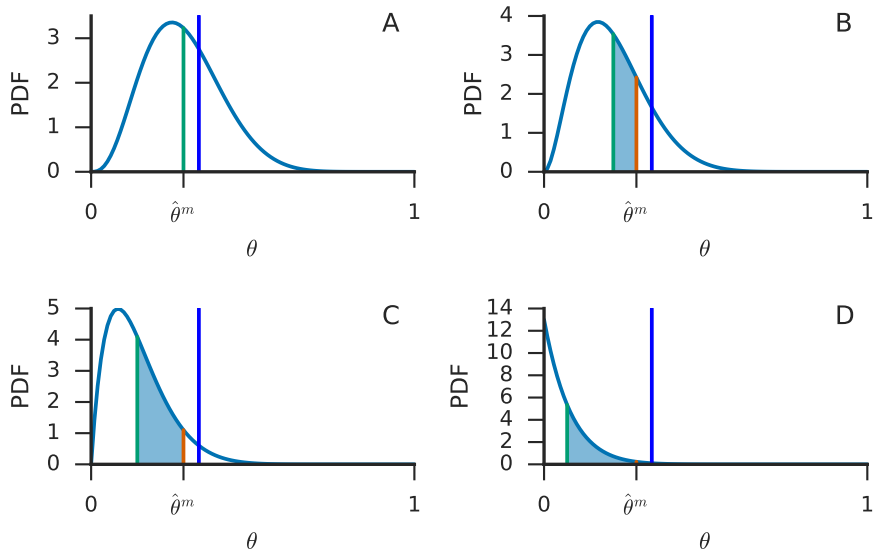


Figure 12: Belief distribution for four parameter sets of *Paranoia*, a and b. In A we plot a $Beta_\theta(4, 10)$ with $Paranoia = 0$, in B a $Beta_\theta(3, 11)$ with $Paranoia = 0.21$, in B a $Beta_\theta(2, 12)$ with $Paranoia = 0.35$, and in D a $Beta_\theta(1, 13)$ with $Paranoia = 0.37$. The green 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 the value of the *Paranoia* parameter. 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 $\hat{\theta}^m = \frac{2}{7}$. 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. Under these four conditions, the agents, initially, cooperate with each other.

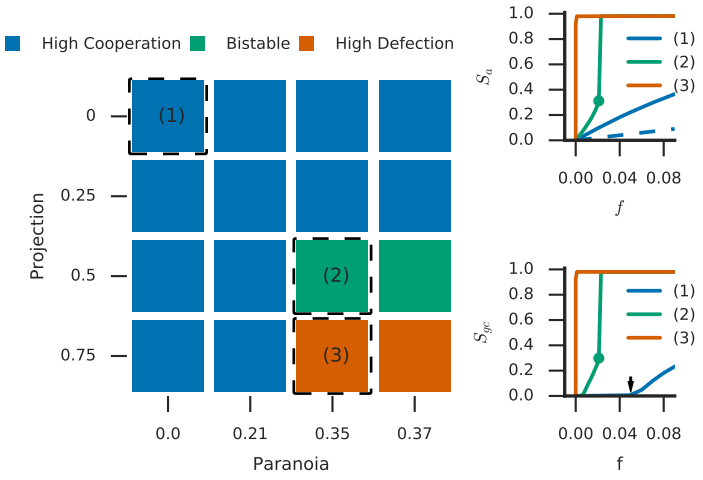


Figure 13: Classification of the network according to its stability for 16 different parameters. The simulations were performed with the asymmetrical version of the *Projection* bias. The two insets show S_a and S_{gc} as a function of the fraction of ALLD agents. The dashed line in the left inset shows the expected S_a due only to the presence of the ALLD agents and assuming that they do not interact with each other. The initial belief of the agents was set such that of $\hat{\theta}^m = \frac{2}{7}$ in every simulation.

In Fig. 13, we show the results using the asymmetric version of the bias, and in Fig. 14, we show the results using the symmetric version of the bias. As in the main text, the three classification of the system appears for both versions of the *Projection* bias.

References

[1] Hayashi N, Ostrom E, Walker J, Yamagishi T (1999) Reciprocity, Trust, and the Sense of Control. *Rationality and Society* 11(1):27–46.

[2] Fehr E, Schmidt KM (1999) A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics* 114(3):817–868.

[3] MacKay DJ (2003) *Information theory, inference and learning algorithms.* (Cambridge university press).

[4] Gelman A, Carlin JB, Stern HS, Rubin DB (2014) *Bayesian data analysis.* (Chapman & Hall/CRC Boca Raton, FL, USA) Vol. 2.

[5] Fudenberg D, Rand DG, Dreber A (2012) Slow to Anger and Fast to Forget: Leniency and Forgiveness in an Uncertain World. *American Economic Review* 102(2):720–749.

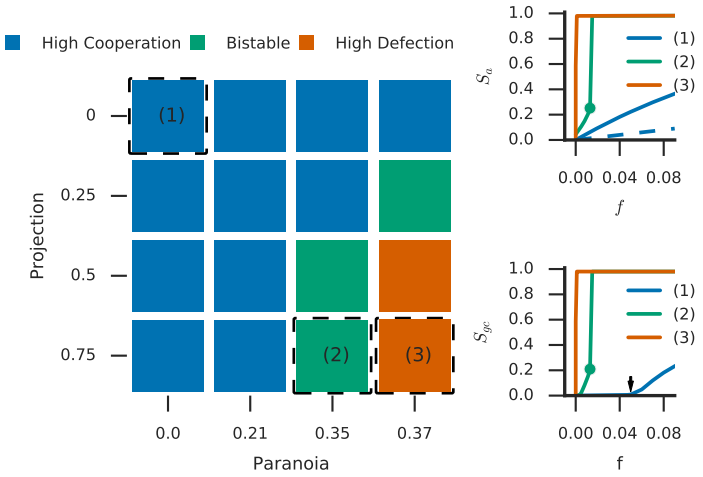


Figure 14: Classification of the network according to its stability for 16 different parameters. The simulations were performed with the symmetrical version of the *Projection* bias. The two insets show S_a and S_{gc} as a function of the fraction of ALLD agents. The dashed line in the left inset shows the expected S_a due only to the presence of the ALLD agents and assuming that they do not interact with each other. The initial belief of the agents was set such that of $\hat{\theta}^m = \frac{2}{7}$ in every simulation.

[6] Di Tella R, Perez-Truglia R, Babino A, Sigman M (2015) Conveniently Upset: Avoiding Altruism by Distorting Beliefs about Others' Altruism. *The American Economic Review* 105(11):3416–3442.

[7] Hagberg AA, Schult DA, Swart PJ (2008) Exploring network structure, dynamics, and function using networkx in *Proceedings of the 7th Python in Science Conference*, eds. Varoquaux G, Vaught T, Millman J. (Pasadena, CA USA), pp. 11 – 15.