The evolution of trust and trustworthiness

Aanjaneya Kumar,[1] Valerio Capraro,[2,*] and Matjaž Perc[3,4,5,†]

[1]*Department of Physics, Indian Institute of Science Education and Research, Dr. Homi Bhabha Road, Pune 411008, India*
[2]*Department of Economics, Middlesex University, The Burroughs, London NW4 4BT, U.K.*
[3]*Faculty of Natural Sciences and Mathematics, University of Maribor, Koroška cesta 160, 2000 Maribor, Slovenia*
[4]*Department of Medical Research, China Medical University Hospital, China Medical University, Taichung 404, Taiwan*
[5]*Complexity Science Hub Vienna, Josefstädterstraße 39, 1080 Vienna, Austria*

Trust and trustworthiness form the basis for continued social and economic interactions, and they are also fundamental for cooperation, fairness, honesty, and indeed for many other forms of prosocial and moral behavior. However, trust entails risks, and building a trustworthy reputation requires effort. So how did trust and trustworthiness evolve, and under which conditions do they thrive? To find answers, we operationalize trust and trustworthiness using the trust game with the trustor's investment and the trustee's return of the investment as the two key parameters. We study this game on different networks, including the complete network, random and scale-free networks, and in the well-mixed limit. We show that in all but one case the network structure has little effect on the evolution of trust and trustworthiness. Specifically, for well-mixed populations, lattices, random and scale-free networks, we find that trust never evolves, while trustworthiness evolves with some probability depending on the game parameters and the updating dynamics. Only for the scale-free network with degree non-normalized dynamics, we find parameter values for which trust evolves but trustworthiness does not, as well as values for which both trust and trustworthiness evolve. We conclude with a discussion about mechanisms that could lead to the evolution of trust and outline directions for future work.

I. INTRODUCTION

While we live in a time where the average individual is much healthier and safer than ever before [1, 2], we are also daunted by several political conflicts, health threats, and extreme poverty in many parts of the world. New innovations and technological breakthroughs often seem to promise a better tomorrow, but the privileges remain restricted to only a tiny fraction of the population. While the issues of equality and egalitarianism are certainly multi-faceted, it is clear that solutions would require us to act prosocially, giving up parts of our personal benefits to help others. But the caveat is that behaving prosocially is costly and not optimal for the individual, and thus will not present itself unless additional mechanisms are at play. No wonder, thus, that understanding the mechanisms that favor the evolution of prosocial behavior has been declared one of the greatest challenges of the 21st century, and that scholars from disciplines as diverse as sociology, psychology, anthropology, economics, biology, and physics have tried to solve the puzzle [3–12].

An exciting development during the past two decades has been the coming of age of network science [13–16], which combined with other methods of statistical physics [17–19], has reached a level of maturity that allows us to tackle some of the greatest challenges of our time. The study of social dynamics [20], traffic [21], crime [22], epidemic processes [23], climate inaction [24], and vaccination [25] are all examples of this exciting development, which can be put under the umbrella of social physics [26]. Prosocial behavior is no exception either and, in particular, the Monte Carlo method for the simulation of evolutionary games and related models on networks has been used prolifically to shed light on the mechanism that may promote it. Most previous works have focussed on three kinds of prosocial behavior, namely cooperation [27–36], strategic fairness [37–44], and altruistic punishment [45–47].

However, recent empirical research in experimental economics and psychology suggests that two of these behaviors – cooperation and altruistic punishment – can be seen as a special form of a more general class of behavior, namely moral behavior [48–53]. This observation opens up the possibility of using the same methods that have been used to study the evolution of cooperation and altruistic punishment to effectively study also the evolution of other types of moral behavior [54]. Following this idea, recent work has explored the evolution of lying and found a number of intriguing conditions for the evolution of truth-telling [55, 56]. Therefore, motivated by the success of this new line of work, we here apply the same methods to study the evolution of trust and trustworthiness.

While the precise definition of trust and trustworthiness depends upon the specific context in which it is being used, a general feature that it exhibits is the willingness of an agent – the trustor – to act in such a way that she is placed in a vulnerable situation with respect to another agent – the trustee, especially when the trustor has no direct ability to monitor the trustee's actions. Thus, trust invariably involves putting oneself in a vulnerable situation in the hope of high returns. High returns that, in the absence of any mechanism to enforce the reciprocation of the trust, might never come, because the trustee can maximize his gain by simply walking away with the profit obtained by betraying the trustor. Knowing this, the trustor should not trust in the first place. Therefore, both trust and trustworthiness go against the assumptions of narrow self-interest. They in fact correlate with several measures of morality, including cooperation and altruism [57]. Moreover, trustworthiness, in the form of 'returning favors', has been recently found to be a universal moral rule across 60 societies

[*] Electronic address: v.capraro@mdx.ac.uk
[†] Electronic address: matjaz.perc@gmail.com

around the world [58]. Yet, despite the fact that both trust and trustworthiness go against the assumption of narrow self-interest, we see them in action everyday – from travelers preferring to look for accommodation through Airbnb rather than spending money on booking a hotel room, to computers in a network deciding to receive information from a source outside of their network. Trust and trustworthiness are ubiquitous in our society, which suggests that, in reality, some mechanisms that favor the evolution of trust and trustworthiness must be at play. The question is which are these mechanisms?

Real interactions do not happen in a vacuum, nor are they random. They are inherently limited to a subset of the population. Some interactions are more frequent than others, and some individuals have many more contacts than others. We are far more likely to interact with friends, family members, and co-workers, than we are with random people. The very fact that interactions are structured has been shown to promote cooperation, along the logic of network reciprocity: cooperators can form clusters to protect themselves from the invasion of defectors [8]. Similarly, it has been shown that spatial structure favors the evolution of fairness and altruistic punishment [38, 45], as well as the evolution of truth-telling, at least in some cases [56]. In this paper, we take inspiration from this line of research and we ask whether network reciprocity promotes also the evolution of trust and trustworthiness.

The plan of the paper is as follows: in Section II A, we will describe the trust game and give a brief overview of the research attention that it has attracted since its introduction. In Section II B, we will provide a description of the Monte Carlo method used to numerically evaluate the stationary state frequencies of different strategies in the trust game, played in well-mixed as well as in networked populations. We will present our results in Section III, and we will end with a summary and outlook for future research in Section IV.

II. METHODS

A. The trust game

Berg, Dickhaut and McCabe [59] proposed the trust game in 1995 as an elegant way to measure trust and trustworthiness between two agents. Player A (the trustor) is initially given some amount of money, normalized to 1. In the first step of the game, player A can choose to *trust* player B (the trustee) and transfer a proportion $x \in [0, 1]$ of her endowment to player B. A transfer of $x = 0$ corresponds to player A choosing to not trust B and to walk away with her money; in this case, the game ends. Instead, if A transfers some amount $x > 0$ to B, the amount of money transferred to player B is tripled (i.e. B gets $3x$ units of money while A is left with $1 - x$) and the game continues. In the second step, player B chooses a fraction $r \in [0, 1]$ of the money he possesses to return to player A. This marks the end of the game. Therefore, the final payoffs of player A and player B are, respectively, $1-x+3xr$ and $3x(1 - r)$ units of money.

In a one-shot anonymous trust game, it is clear that a self-interested player B has no incentive to return any amount of money to player A. This backward induction argument suggests that the best strategy for player A would be to not trust player B. However, experimental research has repeatedly reported that a significant proportion of people choose to transfer a non-zero amount of money to their co-player and a substantial amount of money is also returned [59–65]. Importantly, this behaviour cannot be explained by lack of comprehension [66] or risk aversion [67–70]. Specifically, one observes trust and trustworthiness also among experimental participants who have a clear understanding of what their payoff-maximising strategy is. And trust does not seem to be driven by risk seeking: many individuals who choose to trust in the trust game are averse to taking the risk in an equivalent lottery. In summary, the empirical literature on the trust game provides a clear indication that, while trust and trustworthiness go against monetary payoff maximisation, they often emerge. In order to better comprehend the origin and evolution of trust and trustworthiness, experimental studies need to be complemented with extensive numerical simulations which can help us shed light on when and how trust can be selected in a population and what role does the structure of the population has on its evolution.

At this stage, one might wonder whether trust and trustworthiness are fundamentally different from other forms of social behaviour that have been studied with methods of statistical physics. The answer is positive. This is easy to see in the case of the ultimatum game (used to measure strategic fairness and altruistic punishment) and the sender-receiver game or other deception games (used to measure lying), because they, compared to the trust game, have a completely different payoff structure and set of Nash equilibria. If anything, the trust game *looks* similar to the prisoner's dilemma, the symmetric game in which both players have to decide whether to cooperate or defect: cooperation means paying a cost to give a greater benefit to the other player; defecting means doing nothing. Although the trust game and the prisoner's dilemma might superficially look similar, they are actually fundamentally different. Not only the trust game differs from the prisoner's dilemma on the technical fact that the latter is symmetric, while the former is not, but, more crucially, it fundamentally differs in the evolutionary patterns that it generates, as we will now show.

B. The Monte Carlo method

In the trust game, the amount x that player A transfers to player B is considered as an individual measure of trust, whereas the fraction r that player B returns to player A is taken as a measure of trustworthiness. While in theory any amount of trust $x \in [0, 1]$ and any amount of trustworthiness $r \in [0, 1]$ can be possible, in practice, people in the position of player A often have a binary decision to make, whether to trust or not to trust people in the position of Player B; similarly, people in the position of player B often have a binary decision to make, whether to return a previously agreed amount of money or not [71, 72]. We follow this line of work and we also consider a binary version of the trust game, in which player A

can either choose to trust (T) or not trust (N), whereas player B can either reciprocate (R) player A's trust or betray (B). This yields a payoff bimatrix:

	T	N
R	$1+(3r-1)x, 3(1-r)x$	$1, 0$
B	$1-x, 3x$	$1, 0$

A particularly interesting case is when $x = 1$ and $r = 0.5$, corresponding to the case in which the trustor invests all her money which is normalized to 1 and the amount that the trustee can return corresponds to an equal split between them.

We carried out simulations of the trust game for well mixed populations as well as several network structures (hexagonal, square, and triangular lattices, as well as random networks and scale-free networks) using the Monte Carlo (MC) method. For a well-mixed population with N players, the following are the elementary steps: The simulation starts by randomly distributing the four strategies (T,R), (T,B), (N,R) and (N,B) among N agents. Two players P_1 and P_2 are then randomly picked and they play the trust game with four randomly chosen neighbors. We note that since these players are picked randomly without restricting the selection to nearest neighbors or linked players in a network, the procedure thus yields well-mixed conditions. In each of the eight games, the roles of players are assigned randomly. P_1 and P_2 collect payoffs Π_{P_1} and Π_{P_2}, respectively. Then Player P_2 copies the strategy of player P_1 with probability

$$w = 1/(1 + \exp(\Pi_{P_2} - \Pi_{P_1})/K) \qquad (1)$$

where we choose $K = 0.1$. This step is repeated N times, which by definition completes one full MC step [73]. During the repetition of many full MC steps, every player will thus (since N is also the population size) have a chance once, on average, to change its strategy for each full Monte Carlo step that is made. Indeed, in our simulations, we have performed the MC method for up to 10000 full MC steps. This completes one realization. We conducted 5000 realizations, using randomized initial conditions and the evolution described in the Results section is obtained by averaging over these realizations.

For structured populations, we introduce the constraint in the above mentioned elementary steps that P_1 and P_2 must necessarily be neighbours, or directly linked players in a network. In the case of heterogeneous networks, we use two different imitation rules – the normalized and the unnormalized replicator dynamics. In the unnormalized dynamics, the probability with which P_2 replicates the strategy of P_1 is as before:

$$w = 1/(1 + \exp(\Pi_{P_2} - \Pi_{P_1})/K). \qquad (2)$$

The normalized replication probability differs from the unnormalized one in that payoffs are scaled down by the degree of the players, that is,

$$w = 1/(1 + \exp(\Pi'_{P_2} - \Pi'_{P_1})/K), \qquad (3)$$

where $\Pi'_{P_i} = \frac{\Pi_{P_i}}{k_i}$ and k_i is the degree of player i. In practice, this means that, in the normalized replicator dynamics, players take into account their degree and the degree of other players; whereas, this does not happen in the unnormalized dynamics. Both dynamics are useful in their domain of applicability. The normalized replicator dynamics is useful in situations in which the degrees are visible, and the individuals make a fair comparison with their neighbours – this often happens online, as social media allows people to visualize the connections of an agent with other agents. The unnormalized dynamics is useful in situations in which the imitation of strategies happens on the basis of how well the other player is doing, without taking into account the number of connections they have – for example, people trying to mimic the habits of successful people without accounting for the number of resources that they have at their disposal compared to the person they are copying.

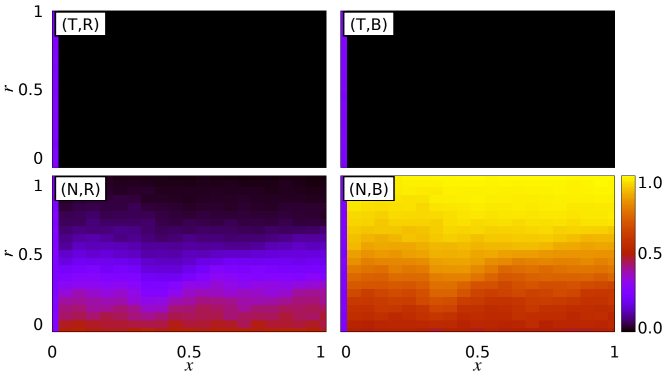

FIG. 1: The stationary density of the four strategies plotted on a 20×20 grid of (x, r) values with both x and r ranging from 0 to 1. Well-mixed population.

III. RESULTS

A. Well-mixed populations

We first report the final densities of the four strategies (T,R), (T,B), (N,R) and (N,B), as a function of the parameters x and r, in a well-mixed population consisting of 500 agents. Figure 1 highlights that, in this case, trust does not evolve, as both the strategies (T,R) and (T,B) appear with density 0 at the steady state, irrespective of the parameters x and r. By contrast, the final density of trustworthiness highly depends on the parameter r, while being insensitive to the parameter x. Specifically, for each x, the prevalence of trustworthiness is about 50% for very small values of r, and then monotonically decreases as r increases.

In order to gain a better understanding of the evolution of trust and trustworthiness, we also conducted several simulations to study the time evolution of the frequencies. We do not report the outputs in the figures, as they are all very similar and certainly not surprising. For example, for $x = 1$ and $r = 0.5$, consistent with Figure 1, we found that the frequencies of the strategies (T,R) and (T,B) go to zero after about 40 MC steps. On the other hand, the strategy (N,B) survives with

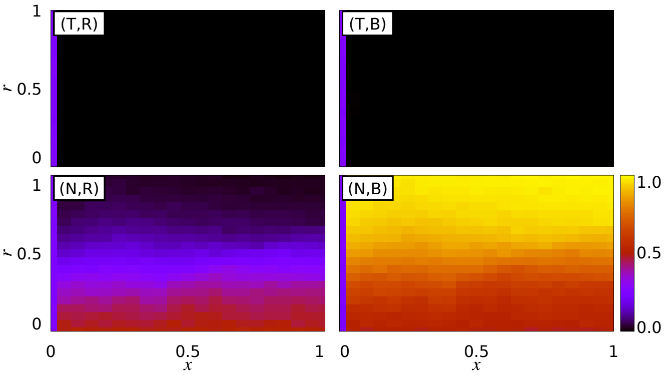

FIG. 2: The stationary density of the four strategies plotted on a 20×20 grid of (x, r) values with both x and r ranging from 0 to 1. Square lattice.

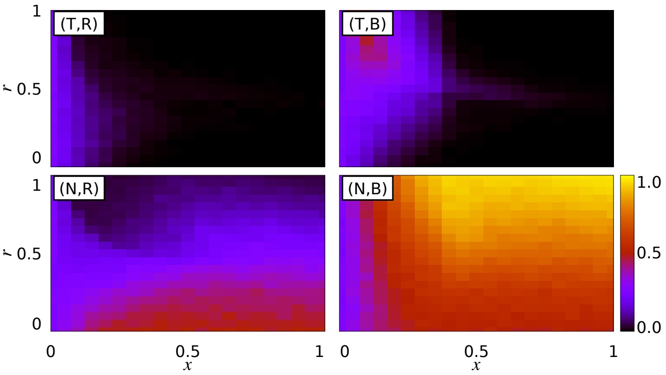

FIG. 4: The stationary density of the four strategies plotted on a 20×20 grid of (x, r) values with both x and r ranging from 0 to 1. Scale free network (unnormalized replicator dynamics).

very high frequency, around 85%. The remaining frequency, about 15%, is taken by the remaining strategy, (N,R).

B. Lattices

To understand the role of the spatial structure on the evolution of trust and trustworthiness, we simulate the trust game on different lattices. Figure 2 reports the stationary densities of the four strategies as a function of x and r on the square lattice. It is immediately evident that the trends in the evolution of different strategies remain the same, compared to the well-mixed populations (Figure 1). We obtain very similar trends in the case of the triangular and the hexagonal lattices (figures reported in the supplementary information) with the results differing only by a small numerical value.

In order to provide further evidence that the lattice structure has very little effect on the evolution of trust and trustworthiness, we also conducted several simulations to study the time evolution of the four strategies in the three lattices (figures not reported in the paper). Consistent with the results mentioned above, we found that the time evolution in the lattice is very similar to the well-mixed case. For example, for $x = 1$ and $r = 0.5$, we found that the frequencies of (T,R) and (T,B) quickly go to zero, whereas the final density of (N,R) is slightly larger that it was in well-mixed populations, but the numerical difference is very small (around 5%); consequently, the final density of (N,B) is slightly smaller in the lattices than it was in well-mixed populations.

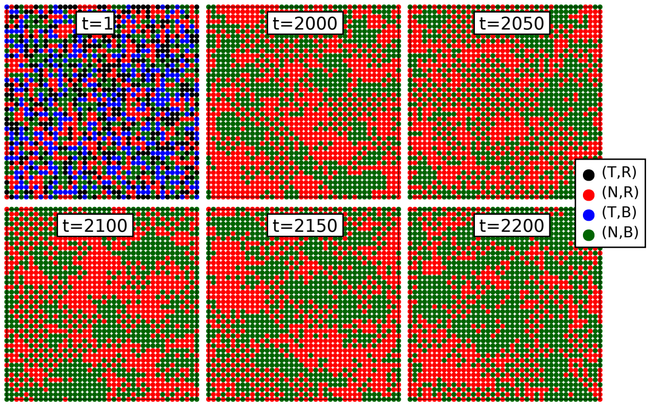

FIG. 3: Snapshots of the evolution of strategies in the trust game played on a 40×40 square lattice. In most realizations of the game, the entire system enters an absorbing state. However, in a few realizations, (N,B) and (N,R) both survive for long time. We have picked one such realization for representation purposes.

To better understand the spatial evolution of the strategies, Figure 3 presents snapshots of the game on a square lattice at late times. In most realizations of the game, the whole population adopts a single strategy and the system enters an absorbing state. However, in a few realizations, (N,B) and (N,R) both survive for long time. We have picked one such realization for representation purposes. It is clearly seen that the two surviving strategies tend to cluster together, forming metastable clusters. It is also noticed that, apart from clusters, there are patches where the two surviving strategies appear alternatively.

C. Random and scale-free networks

To investigate the role of the spatial structure on the evolution of trust and trustworthiness further, we simulate the trust game on a scale-free network generated by the Barabási-Albert algorithm and an Erdős-Rényi random network, with 500 agents each (both with an average degree close to 10). Since these networks are not regular, we consider both the normalized and the unnormalized replicator dynamics.

In the case of random networks, we obtain results very similar to the well-mixed populations and the three lattices. This holds using both the normalized and unnormalized dynamics (see supplementary information) and it provides further evidence that spatial correlations alone do not lead to the evolution of trust.

To study the effects of heterogeneity in the network of contacts, we study the trust game on scale-free networks. In this case, evolution turns out to be more nuanced and interesting, as it appears to depend on the choice of replicator dynamics. When agents imitate other agents using the normalized replicator dynamics, trust does not evolve and the results are virtually the same as in the previous cases. However, when agents imitate other agents using the unnormalized replicator dynamics, we observe rich behaviour in terms of the evolution of different strategies. Figure 4 is a heatmap of the steady state density of the four strategies, and it shows clear differences with Figure 1. The strategy (T,R), which in the other cases never evolved, this time evolves for small values of x. Similarly, the strategy (T,B), which generally vanished in the other cases, now evolves for $x < 0.5$. The difference is particularly evident for x small and r large, where there appears to be an island in which (T,B) actually evolves with frequency close to 50%.

To have a better understanding of these differences, we next explore the time evolution of the four strategies in three prototypical cases, one in which we expect virtually no differences compared to the well-mixed case ($x = 1$ and $r = 0.5$) and two in which we expect large differences ($x = 0.1, r = 0.8$ and $x = 0.1, r = 0.3$). Note indeed that Figure 4 suggests that, for $x = 1$ and $r = 0.5$, the final densities according to the unnormalized replicator dynamics should be very similar to those according to the normalized replicator dynamics, which are in turn very similar to those in well-mixed populations. By contrast, we chose the values $x = 0.1$ and $r = 0.8$ to illustrate the evolution in correspondence to the island described above where we expect trust but not trustworthiness to evolve. And we chose the values $x = 0.1$ and $r = 0.3$ to illustrate a situation in which we expect both trust and trustworthiness to evolve.

Figure 5 reports the time evolution of the four strategies for $x = 1$ and $r = 0.5$ using the unnormalized dynamics (top panel) and the normalized dynamics (bottom panel). As expected, the bottom panel is virtually identical to the well-mixed population (not reported in the figures, but discussed earlier in the text). The top panel differs from the bottom panel only in a very small detail: the strategy (T,B) evolves with a very small frequency.

Figure 6 (top panel) reports the evolution of the four strategies for $x = 0.1$ and $r = 0.8$, only for the unnormalized replicator dynamics. As expected, this time we see very large differences compared to the normalized replicator dynamics, which we do not report in the paper, being virtually identical to Figure 5. In particular, the biggest difference can be observed in the evolution of the strategy (T,B). In the previous case ($x = 1$, $r = 0.5$), this strategy almost vanished when agents imitate other agents using the unnormalized replicator dynamics, and completely vanished when they used the normalized replicator dynamics. In stark contrast, the strategy (T,B) now evolves with frequency close to 50%. Another difference can be noticed in the case of the strategy (T,R). This strategy vanished in all the earlier cases. By contrast, it now survives, although with a very small probability around 2%. Finally, Figure 6 (bottom panel) reports the evolution of the four strategies for $x = 0.1$ and $r = 0.3$, again only for the unnormalized replicator dynamics. As expected, this time the strategy (T,R) evolves with a non-negligible frequency around 15%. The strategy (N,R) evolves with an even higher frequency around 25%, compared to the 5% for $x = 0.1$ and $r = 0.8$. These increases in frequency come mainly at the expenses of the strategy (T,B), which, for $x = 0.1$ and $r = 0.8$ evolved with very high frequency (about 45%), whereas it now evolves only with frequency around 20%; and to a lesser extent at the expenses of the strategy (N,B), which, for $x = 0.1$ and $r = 0.8$ evolved with frequency around 45%, whereas it now evolves with frequency below 40%.

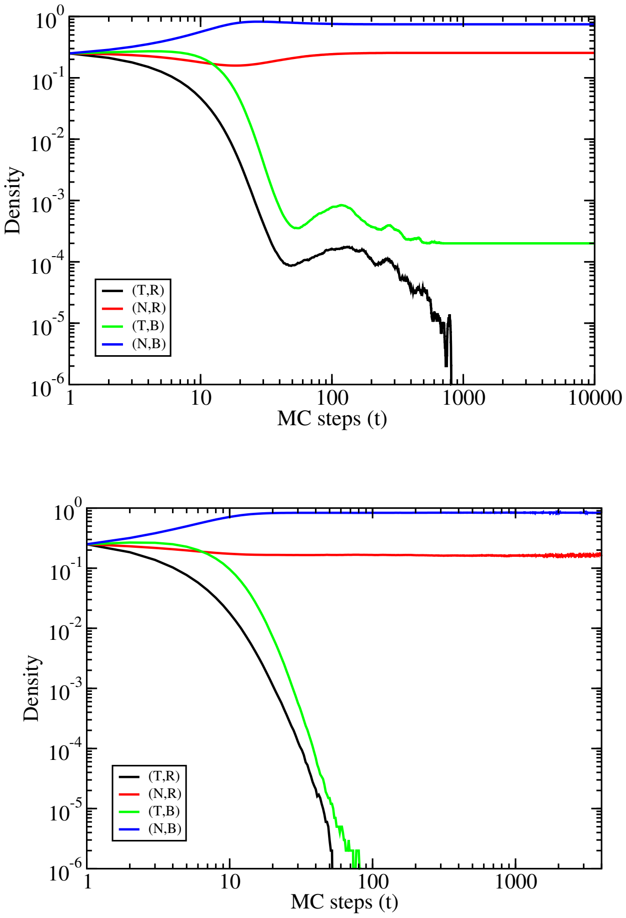

FIG. 5: Time evolution of the four strategies in a scale free network for values $x = 1$ and $r = 0.5$, for the unnormalized (top) and normalized (bottom) replicator dynamics.

IV. DISCUSSION

We have used the Monte Carlo method to study the evolution of trust and trustworthiness in well-mixed populations, three different types of lattices, random networks and scale-free networks. Since the latter two networks are not regular, in these cases we have studied the evolution of trust and trustworthiness both when agents imitate other agents by taking into account their degree (normalized replicator dynamics) and when they do not (unnormalized replicator dynamics). As

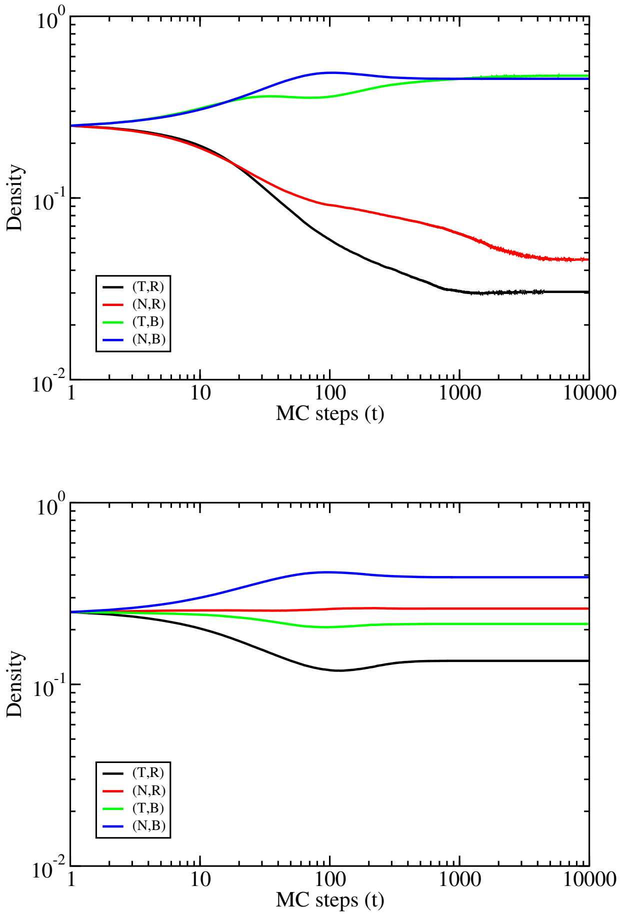

FIG. 6: Time evolution of the four strategies in a scale-free network (unnormalized replicator dynamics) for values $x = 0.1$ and $r = 0.8$ (top) and for $x = 0.1$ and $r = 0.3$ (bottom).

a measure of trust and trustworthiness, we have used a binary version of the trust game [59]. The choice made by player 1 (the trustor) was taken as a measure of trust; the choice made by player 2 (the trustee) was taken as a measure of trustworthiness. We parameterized the game through two parameters: $x \in [0, 1]$ describes the amount of money that the trustor can send to the trustee; $r \in [0, 1]$ represents the proportion of the amount received by the trustee that the he can return to the trustor.

Our exploration provided evidence of several results. First, in well-mixed populations, trust never evolves, whereas the evolution of trustworthiness depends monotonically decreasingly on r (and shows very little dependence on x). Second, to understand the effects of spatial correlations on the evolution of different strategies, we simulated the trust game on homogenous and heterogenous networks. On lattices, random networks (using both the imitation dynamics), and scale-free networks with normalized replicator dynamics, we observe that that trust does not evolve, and in these cases, the final densities of the four strategies are very similar to the corresponding final densities in well-mixed populations. This conclusively points to the fact that solely spatial structure does not lead to the evolution of trust. Third, scale-free networks with unnormalized replicator dynamics give rise to the most nuanced evolution: for small values of r and x, both trust and trustworthiness evolve, although with a relatively small frequency around 15%; for small values of x and large values of r, trust evolves with a relatively large frequency around 50%, but this time trustworthiness does not evolve. These results can readily be compared to the evolutionary prisoner's dilemma on scale-free networks [74] with normalized and unnormalized replicator dynamics where the evolution of cooperation is possible in both cases. The heterogeneous scale-free network provides a mechanism for the survival of cooperators up to larger values of temptation to defect, when compared to well-mixed populations. However, it is interesting to note that when the payoffs of an individual are normalized by their degrees, the fraction of surviving cooperators is significantly lesser. Our results hint that while heterogeneity also provides a route for the evolution of trust, it only does so when the payoff of the players is accumulated over all its interactions with its neighbours, and not averaged over them, like in the normalized dynamics.

In sum, we have operationalized trust and trustworthiness using the trust game with the trustor's investment and the trustee's return of the investment as the two key parameters and we have studied their evolution in a number of networks and our results have shown that trust and trustworthiness very rarely evolve in these networks, and even more rarely do they do it together: when trustworthiness evolves, then trust does not; when trust evolves, trustworthiness does not. Only in a relatively small region (both r and x small) and only in the case of scale-free networks and unnormalized replicator dynamics, the strategy (T,R) evolved with a non-negligible, although still relatively small (around 15%) probability.

This is the first systematic study on the evolution of trust and trustworthiness on networks. Most previous work applied the Monte Carlo method to study the evolution of cooperation in the prisoner's dilemma [27–36], the evolution of strategic fairness and altruistic punishment in the ultimatum game [37–44], and the evolution of truth-telling in the sender-receiver game [55, 56] or in other deception games [75–79]. These games are fundamentally different from the trust game used in the current analysis. The trust game is obviously different from the sender-receiver game and the ultimatum game, because they have completely different strategic structure and, consequently, sets of equilibria. But it is also different from the prisoner's dilemma: while this game is symmetric, the trust game is not. This is probably the reason that leads to the fact that, in general, the spatial structure favors the evolution of cooperation in the prisoner's dilemma, while having very little effect on the evolution of trust and trustworthiness. A handful of papers have studied the evolution of trust and trustworthiness using the trust game or some variants thereof. However, most of these works focused on well-mixed populations [80–84]. These works typically show that, with no additional mechanisms, such as choice visibility, trust and trustworthiness do not evolve in well-mixed populations. Our findings are thus in line with this preceding research. A variant of the trust game has also been studied on networks, however the analysis was mainly focused more on group effects, and on

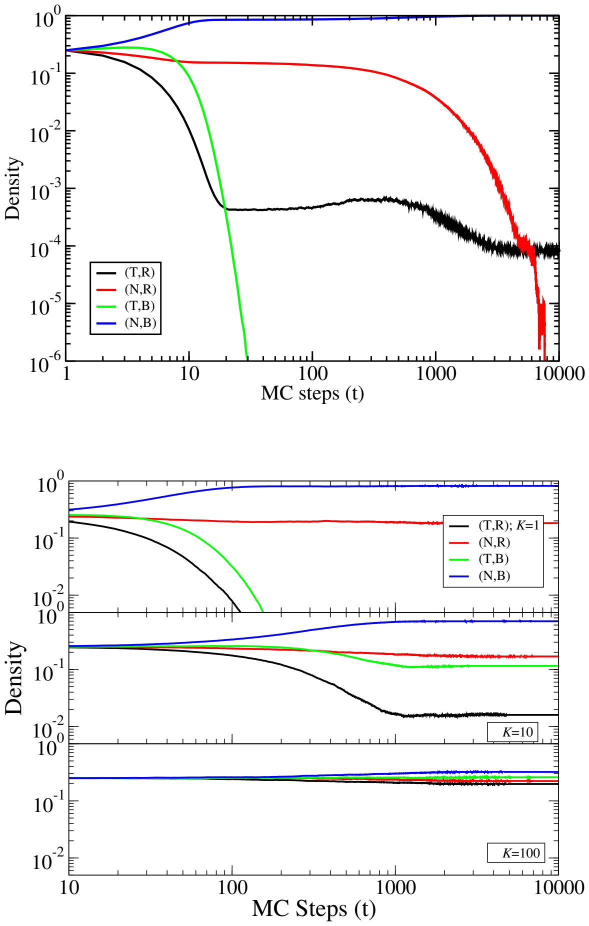

FIG. 7: (Top panel) The evolution of the four strategies as a function of time in the case of a well mixed population after adding a ‘good samaritan’ agent that always chooses the strategy (T,R). We look at the evolution of the 499 agents in the population (excluding the good samaritan) for the values $x = 1$ and $r = 0.5$. (Bottom panel) Time evolution of the four strategies (without any good samaritan) for $x = 1$, $r = 0.5$, and varying levels of noise ($K = 1, 10, 100$).

one specific network – the email network of a university in Tarragona [85].

The fact that the spatial structure, apart from one special case, does not promote the evolution of trust and trustworthiness together with the observation that, in reality, we do see a lot of trust and trustworthiness, generates the following question: What mechanisms promote the evolution of trust and trustworthiness? In Figure 7, the top panel is a plot of the evolution of the strategies of 499 agents for values $x = 1$ and $r = 0.5$ in a well-mixed population of 500 agents where the excluded individual is a ‘good samaritan’ who always chooses the strategy (T,R). We emphasize that the plot only considers the evolution of the strategies of the 499 agents which does not include the good samaritan. It can be seen that a finite, albeit small fraction of trustors, survive in the stationary state upon the inclusion of a single good samaritan agent at a value of x and r where previously trust did not evolve. It is well known that zealots can drive the evolution of cooperation in the prisoner’s dilemma game [86] and a detailed study of the effects of good samaritans on the evolution of trust and other moral behaviours provides an interesting avenue for future research. The bottom panel of Figure 7 explores the influence of noisy imitation on the dynamics. Noise can be interpreted as the lack of perfect information about the payoffs of other people, or as sub-optimal decision making. We show a comparison of the evolution of increased noise in the imitation process for $x = 1$, $r = 0.5$, and $K = 1, 10$ and 100). It is expected in the limit of $K \to \infty$ that each strategy survives with equal probability as the dynamics is random. However, even at $K = 10$, we can see that trust evolves to steady state density of around 10% and the strategy (T,R) which accounts for trusting, and trustworthy individuals also evolve to a final density of around 1%. Studying further the effects of noisy imitation as a function of the parameters of the game could lead us to novel insights.

Several other mechanisms could be responsible for the evolution of trust [87]. Possible candidates could be reward and punishment as well as apology, forgiveness, and emotions such as guilt. We know that these mechanisms promote the evolution of cooperation [88–97]. Along similar lines, it is possible that they also promote the evolution of trust and trustworthiness. In fact, in reality, we know that, for example, online transactions, which are fundamentally based on a relationship of trust, are supported by rating systems that provide a measure of the trustworthiness of the agents. Therefore, it is likely that the presence of a reputation mechanism promotes the evolution of trust. Following recent works of Fudenberg and Imhof [98] and Veller and Hayward [99], it would also be interesting to study the problem where not only do the agents evolve using imitation, but also can spontaneously mutate and adopt different strategies. Additionally, we notice that our results were obtained on particular networks and imitation rules; it is possible that other networks and/or other imitation rules lead to the evolution of trust and trustworthiness. Finally, individual differences for example in gender, age, dominance status, number of neighbours, kinship, which are well-known to affect cooperative and altruistic behaviour [36, 100–104], can also affect the evolution of trust and trustworthiness. Future work should explore these possibilities.

Author contributions

AK, VC, MP designed the study, AK conducted the simulations, AK, VC, MP discussed the results and wrote the paper.

Acknowledgments

We thank Shraddha Pathak and Sandeep Chowdhary for many useful discussions.

Data accessibility

This paper has no data.

Funding statement

Aanjaneya Kumar would like to acknowledge the Prime Minister's Research Fellowship for financial support. Matjaž Perc was supported by the Slovenian Research Agency (Grant Nos. J4-9302, J1-9112, and P1-0403).

[1] S. Pinker, *The Better Angels of our Nature: Why Wiolence has Declined* (Viking, New York, 2011).

[2] S. Pinker, *Enlightenment now: The case for reason, science, humanism, and progress* (Penguin, 2018).

[3] G. Hardin, Science **162**, 1243 (1968).

[4] R. L. Trivers, The Quarterly Review of Biology **46**, 35 (1971).

[5] R. Axelrod and W. D. Hamilton, Science **211**, 1390 (1981).

[6] M. Milinski, D. Semmann, and H.-J. Krambeck, Nature **415**, 424 (2002).

[7] R. Boyd, H. Gintis, S. Bowles, and P. J. Richerson, Proceedings of the National Academy of Sciences **100**, 3531 (2003).

[8] M. A. Nowak, Science **314**, 1560 (2006).

[9] C. Hilbe and K. Sigmund, Proceedings of the Royal Society B: Biological Sciences **277**, 2427 (2010).

[10] V. Capraro, PloS One **8** (2013).

[11] D. G. Rand and M. A. Nowak, Trends in Cognitive Sciences **17**, 413 (2013).

[12] M. Perc, J. J. Jordan, D. G. Rand, Z. Wang, S. Boccaletti, and A. Szolnoki, Phys. Rep. **687**, 1 (2017).

[13] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical Processes on Complex Networks* (Cambridge University Press, Cambridge, U.K., 2008).

[14] M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, Oxford, U.K., 2010).

[15] E. Estrada, *The Structure of Complex Networks: Theory and Applications* (Oxford University Press, Oxford, 2012).

[16] A.-L. Barabási, *Network Science* (Cambridge University Press, Cambridge, 2015).

[17] H. E. Stanley, *Introduction to Phase Transitions and Critical Phenomena* (Clarendon Press, Oxford, 1971).

[18] T. M. Liggett, *Interacting Particle Systems* (Springer, New York, 1985).

[19] D. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, Cambridge, 2000).

[20] C. Castellano, S. Fortunato, and V. Loreto, Rev. Mod. Phys. **81**, 591 (2009).

[21] D. Helbing, Rev. Mod. Phys. **73**, 1067 (2001).

[22] M. R. D'Orsogna and M. Perc, Phys. Life Rev. **12**, 1 (2015).

[23] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, Rev. Mod. Phys. **87**, 925 (2015).

[24] J. M. Pacheco, V. V. Vasconcelos, and F. C. Santos, Phys. Life Rev. **11**, 573 (2014).

[25] Z. Wang, C. T. Bauch, S. Bhattacharyya, A. d'Onofrio, P. Manfredi, M. Perc, N. Perra, M. Salathé, and D. Zhao, Phys. Rep. **664**, 1 (2016).

[26] M. Perc, Sci. Rep. **9**, in press (2019).

[27] F. C. Santos and J. M. Pacheco, Physical Review Letters **95**, 098104 (2005).

[28] J. M. Pacheco, A. Traulsen, and M. A. Nowak, Physical Review Letters **97**, 258103 (2006).

[29] J. Gómez-Gardenes, M. Campillo, L. M. Floría, and Y. Moreno, Physical Review Letters **98**, 108103 (2007).

[30] H. Ohtsuki, M. A. Nowak, and J. M. Pacheco, Physical Review Letters **98**, 108106 (2007).

[31] S. Lee, P. Holme, and Z.-X. Wu, Physical Review Letters **106**, 028702 (2011).

[32] J. Tanimoto, M. Brede, and A. Yamauchi, Phys. Rev. E **85**, 032101 (2012).

[33] Z. Wang, S. Kokubo, M. Jusup, and J. Tanimoto, Phys. Life Rev. **14**, 1 (2015).

[34] M. A. Javarone, Eur. Phys. J. B **89**, 42 (2016).

[35] M. A. Amaral and M. A. Javarone, Phys. Rev. E **97**, 042305 (2018).

[36] D. Vilone, V. Capraro, and J. J. Ramasco, Journal of Physics Communications **2**, 025019 (2018).

[37] A. Szolnoki, M. Perc, and G. Szabó, Physical Review Letters **109**, 078701 (2012).

[38] K. M. Page, M. A. Nowak, and K. Sigmund, Proceedings of the Royal Society of London. Series B: Biological Sciences **267**, 2177 (2000).

[39] M. Kuperman and S. Risau-Gusman, The European Physical Journal B **62**, 233 (2008).

[40] V. M. Eguíluz and C. J. Tessone, Advances in Complex Systems **12**, 221 (2009).

[41] R. da Silva, G. A. Kellermann, and L. C. Lamb, Journal of Theoretical Biology **258**, 208 (2009).

[42] L. Deng, W. Tang, and J. Zhang, Physica A: Statistical Mechanics and its Applications **390**, 4227 (2011).

[43] J. Gao, Z. Li, T. Wu, and L. Wang, EPL (Europhysics Letters) **93**, 48003 (2011).

[44] A. Szolnoki, M. Perc, and G. Szabó, EPL (Europhysics Letters) **100**, 28005 (2012).

[45] D. Helbing, A. Szolnoki, M. Perc, and G. Szabó, PLoS Comput. Biol. **6**, e1000758 (2010).

[46] A. Szolnoki and M. Perc, Phys. Rev. X **3**, 041021 (2013).

[47] A. Szolnoki and M. Perc, Phys. Rev. X **7**, 041027 (2017).

[48] E. Dal Bó and P. Dal Bó, Journal of Public Economics **117**, 28 (2014).

[49] L. Biziou-van Pol, J. Haenen, A. Novaro, A. Occhipinti Liberman, and V. Capraro, Judgment and Decision Making **10**, 538 (2015).

[50] E. O. Kimbrough and A. Vostroknutov, Journal of the European Economic Association **14**, 608 (2016).

[51] K. Eriksson, P. Strimling, P. A. Andersson, and T. Lindholm, Journal of Experimental Social Psychology **69**, 59 (2017).

[52] V. Capraro and D. G. Rand, Judgment and Decision Making **13**, 99 (2018).

[53] V. Capraro, G. Jagfeld, R. Klein, M. Mul, and I. van de Pol, Scientific Reports **9**, 11880 (2019).

[54] V. Capraro and M. Perc, Front. Phys. **6**, 107 (2018).

[55] V. Capraro, M. Perc, and D. Vilone, Journal of the Royal Society Interface **16**, 20190211 (2019).

[56] V. Capraro, M. Perc, and D. Vilone, Physical Review E (2020).

[57] A. Peysakhovich, M. A. Nowak, and D. G. Rand, Nature Communications **5**, 1 (2014).

[58] O. S. Curry, D. A. Mullins, and H. Whitehouse, Current Anthropology **60**, 47 (2019).

[59] J. Berg, J. Dickhaut, and K. McCabe, Games and Economic Behavior **10**, 122 (1995).

[60] E. L. Glaeser, D. I. Laibson, J. A. Scheinkman, and C. L. Soutter, The Quarterly Journal of Economics **115**, 811 (2000).

[61] I. Bohnet and R. Zeckhauser, Journal of Economic Behavior & Organization **55**, 467 (2004).

[62] D. Malhotra, Organizational Behavior and Human Decision Processes **94**, 61 (2004).

[63] M. Kosfeld, M. Heinrichs, P. J. Zak, U. Fischbacher, and E. Fehr, Nature **435**, 673 (2005).

[64] E. Fehr, Journal of the European Economic Association **7**, 235 (2009).

[65] N. D. Johnson and A. A. Mislin, Journal of Economic Psychology **32**, 865 (2011).

[66] A. Ortmann, J. Fitzgerald, and C. Boeing, Experimental Economics **3**, 81 (2000).

[67] D. Fetchenhauer and D. Dunning, Journal of Economic Psychology **30**, 263 (2009).

[68] D. Dunning, D. Fetchenhauer, and T. M. Schlösser, Journal of Economic Psychology **33**, 686 (2012).

[69] B. Corgnet, A. M. Espín, R. Hernán-González, P. Kujal, and S. Rassenti, Journal of Behavioral and Experimental Economics **64**, 20 (2016).

[70] A. Ben-Ner and F. Halldorsson, Journal of Economic Psychology **31**, 64 (2010).

[71] J. Ermisch, D. Gambetta, H. Laurie, T. Siedler, and S. Noah Uhrig, Journal of the Royal Statistical Society: Series A (Statistics in Society) **172**, 749 (2009).

[72] A. M. Espín, F. Exadaktylos, and L. Neyse, Frontiers in Psychology **7**, 728 (2016).

[73] K. Binder and D. K. Hermann, *Monte Carlo Simulations in Statistical Physics* (Springer, Heidelberg, 1988).

[74] N. Masuda, Sci. Rep. **2**, 646 (2012).

[75] A. Grafen, Journal of Theoretical Biology **144**, 517 (1990).

[76] D. Catteeuw, T. A. Han, and B. Manderick, in *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation* (2014), pp. 153–160.

[77] T. A. Han, L. M. Pereira, F. C. Santos, and T. Lenaerts, Scientific Reports **3**, 2695 (2013).

[78] T. A. Han, L. M. Pereira, and T. Lenaerts, Journal of the Royal Society Interface **12**, 20141203 (2015).

[79] L. M. Pereira, T. Lenaerts, et al., Autonomous Agents and Multi-Agent Systems **31**, 561 (2017).

[80] J. M. McNamara, P. A. Stephens, S. R. Dall, and A. I. Houston, Proceedings of the Royal Society B: Biological Sciences **276**, 605 (2009).

[81] M. L. Manapat, M. A. Nowak, and D. G. Rand, Journal of Economic Behavior & Organization **90**, S57 (2013).

[82] M. L. Manapat and D. G. Rand, Dynamic Games and Applications **2**, 401 (2012).

[83] H. Abbass, G. Greenwood, and E. Petraki, IEEE Transactions on Evolutionary Computation **20**, 470 (2015).

[84] P. Rauwolf and J. J. Bryson, Dynamic Games and Applications **8**, 891 (2018).

[85] M. Chica, R. Chiong, J. J. Ramasco, and H. Abbass, Communications in Nonlinear Science and Numerical Simulation **79**, 104870 (2019).

[86] F. C. Santos and J. M. Pacheco, J. Evol. Biol. **19**, 726 (2006).

[87] P. Andras, L. Esterle, M. Guckert, T. A. Han, P. R. Lewis, K. Milanovic, T. Payne, C. Perret, J. Pitt, S. T. Powers, et al., IEEE Technology and Society Magazine **37**, 76 (2018).

[88] D. G. Rand, A. Dreber, T. Ellingsen, D. Fudenberg, and M. A. Nowak, Science **325**, 1272 (2009).

[89] Ö. Gürerk, B. Irlenbusch, and B. Rockenbach, Science **312**, 108 (2006).

[90] A. Szolnoki and M. Perc, New J. Phys. **14**, 093016 (2012).

[91] M. Milinski, D. Semmann, H.-J. Krambeck, and J. Marotzke, Proceedings of the National Academy of Sciences **103**, 3994 (2006).

[92] L. A. Martinez-Vaquero, T. A. Han, L. M. Pereira, and T. Lenaerts, Scientific reports **5**, 10639 (2015).

[93] V. Capraro, F. Giardini, D. Vilone, and M. Paolucci, Judgment and Decision Making **11**, 589 (2016).

[94] C. D. Melo, S. Marsella, and J. Gratch, ACM Transactions on Computer-Human Interaction (TOCHI) **23**, 1 (2016).

[95] L. A. Martinez-Vaquero, T. A. Han, L. M. Pereira, and T. Lenaerts, Scientific reports **7**, 1 (2017).

[96] L. M. Pereira, T. Lenaerts, L. A. Martinez-Vaquero, and T. A. Han, in *AAMAS* (2017), pp. 1422–1430.

[97] Y. Fang, T. P. Benko, M. Perc, H. Xu, and Q. Tan, Proceedings of the Royal Society A **475**, 20190349 (2019).

[98] D. Fudenberg and L. A. Imhof, J. Econ. Theor. **131**, 251 (2006).

[99] C. Veller and L. K. Hayward, Journal of Economic Theory **162**, 93 (2016).

[100] P. D. Taylor and A. J. Irwin, Evolution **54**, 1135 (2000).

[101] M. Bao and G. Wild, Theoretical Population Biology **82**, 200 (2012).

[102] D. G. Rand, V. L. Brescoll, J. A. Everett, V. Capraro, and H. Barcelo, Journal of Experimental Psychology: General **145**, 389 (2016).

[103] D. G. Rand, Journal of experimental social psychology **73**, 164 (2017).

[104] P. Brañas-Garza, V. Capraro, and E. Rascon-Ramirez, Economics Letters **170**, 19 (2018).