The evolution of lying in well-mixed populations

Valerio Capraro,[1,*] Matjaž Perc,[2,3] and Daniele Vilone[4,5]

[1]*Department of Economics, Middlesex University, The Burroughs, London NW4 4BT, U.K.*
[2]*Faculty of Natural Sciences and Mathematics, University of Maribor, Koroška cesta 160, SI-2000 Maribor, Slovenia*
[3]*Complexity Science Hub Vienna, Josefstädterstraße 39, A-1080 Vienna, Austria*
[4]*LABSS (Laboratory of Agent Based Social Simulation), Institute of Cognitive Science and Technology, National Research Council (CNR), Via Palestro 32, 00185 Rome, Italy*
[5]*Grupo Interdisciplinar de Sistemas Complejos (GISC), Departamento de Matemáticas, Universidad Carlos III de Madrid, 28911 Leganés, Spain*

Lies can have profoundly negative consequences for individuals, groups, and even for societies. Understanding how lying evolves and when it proliferates is therefore of significant importance for our personal and societal well-being. To that effect, we here study the sender-receiver game in well-mixed populations with methods of statistical physics. We use the Monte Carlo method to determine the stationary frequencies of liars and believers for four different lie types. We consider altruistic white lies that favor the receiver at a cost to the sender, black lies that favor the sender at a cost to the receiver, spiteful lies that harm both the sender and the receiver, and Pareto white lies that favor both the sender and the receiver. We find that spiteful lies give rise to trivial behavior, where senders quickly learn that their best strategy is to send a truthful message, whilst receivers likewise quickly learn that their best strategy is to believe the sender's message. For altruistic white lies and black lies, we find that most senders lie while most receivers do not believe the sender's message, but the exact frequencies of liars and non-believers depend significantly on the payoffs, and they also evolve non-monotonically before reaching the stationary state. Lastly, for Pareto white lies we observe the most complex dynamics, with the possibility of both lying and believing evolving with all frequencies between 0 and 1 in dependence on the payoffs. We discuss the implications of these results for moral behavior in human experiments.

Introduction

There are arguments and data in favor of the statement that we live safer, richer, and healthier than ever before [1, 2]. But the gap between rich and poor is currently growing out of all reasonable proportions. And it is difficult to look away from the armed conflicts, hunger, and poverty without thinking that we ought to be able to do better. While we try our best to be compassionate, civilized, and social, and while there is an abundance of technological breakthroughs and innovations that make our lives better, many human societies are still seriously failing to meet the most basic needs of millions around the world [3]. We are also dangerously depleting natural resources, our industries and ways of life are changing the climate, and we have fallen victim to echo chambers and misinformation, to the point that it is often impossible to discern truth from lies [4, 5].

Although the above-outlined issues are diverse and multifaceted, they do share one common property. Their solutions require cooperation. And we do cooperate – in fact, we are champions of cooperation, to the point that we exercise "SuperCooperation" [6]. But since natural selection in all of biology favors the fittest and the most successful individuals, there is still an innate selfishness in us that greatly challenges our cooperative drive. Cooperation is costly, and exercising it weighs down on individual well-being and prosperity. We therefore often succumb to the Darwin within, and we forget about less privileged others, and about future generations, and the health of our climate, and about many related issues that would require large-scale cooperation to be improved. Not surprisingly, understanding and promoting cooperation in human societies has once been declared one of the grandest challenges of the 21st century [7], and scholars from disciplines as diverse as economics, psychology, sociology, biology, and anthropology have explored what factors favor people's cooperative behavior [8–19].

Methods of physics, in particular the Monte Carlo method and related approaches in statistical physics and network science [20–25], have also emerged as being very useful for studying many social phenomena. Statistical physics of social dynamics [26], of evolutionary games in structured populations [27–30], of crime [31], of gossip [32], and of epidemic processes and vaccination [33, 34], are all examples of this exciting development, with human cooperation being no exception [35, 36]. However, empirical work has shown that cooperation is only one kind of a more general class of behaviors – moral behaviors [37]. This suggests that the same methods could be applied effectively to study the evolution of other types of moral behaviors as well [38].

Using this as motivation, here we use methods of statistical physics to study the evolution of lying, or deception. Why deception? Deception has significant negative impacts on government, companies, and the society as a whole. For example, tax evasion costs approximately 100 billion a year to the US government alone [39], whereas, still in the USA, insurance fraud costs about 40 billion a year to insurance companies [40]. More recently, research has also focused on the spreading of fake news and misinformation [5], which, by favouring the emergence of inaccurate beliefs about the real state of the society, may represent a serious threat to democracy [41]. Thus not surprisingly, studying dishonesty has a long history of interest among social scientists [42–56], with

[*] Electronic address: v.capraro@mdx.ac.uk

the sender-receiver game being a popular theoretical paradigm to measure (dis)honesty [57].

In what follows, we re-introduce the sender-receiver game in a way that is appropriate to use with the Monte Carlo method, and we determine the stationary frequencies of liars and believers for four different lie types in well-mixed populations. In particular, we consider altruistic white lies, black lies, spiteful lies, and Pareto white lies, and we study in detail the dynamics that emerges as a result. As we will show, with spiteful lies in play senders and receivers both quickly learn that their best strategy is to send a truthful message and believe it, respectively. But for other types of lying, the dynamics becomes more nuanced. For example, for altruistic white lies and black lies, we will show that most senders lie while most receivers do not believe the sender's message, while for Pareto white lies, we will show that both lying and believing can evolve with any frequencies between 0 and 1. Our research thus adds a theoretically rigorous quantitative component to studying dishonesty, which has important implications for better understanding moral behavior in general, as well as provides pointers for devising innovative human experiments to test the theory.

The sender-receiver game

Behavioral scientists have invented several tasks to measure people's (dis)honesty. The more popular ones are the die-rolling-paradigm [56], the matrix task [42], the Philip Sidney game [58], and the sender-receiver game [57]. In this work, we focus on the sender-receiver game, which is particularly suitable for the application of the Monte Carlo method, being a game with two players and (practically) two strategies, whereas the die-rolling-paradigm and the matrix task are both decision problems, with no strategic component, in which one person has to decide whether to lie for their benefit, or not. Moreover, the sender-receiver game allows us to study four different types of lies (black lies, spiteful lies, altruistic white lies, and Pareto white lies), whereas the Philip Sydney game, although strategically similar to the sender-receiver game, permits to study only black lies. In particular, we focus on the variant of the sender-receiver game introduced by Erat and Gneezy in [53].

The game is as follows. There are two potential allocations of money between the sender and the receiver, Option A and Option B. The sender rolls a six-face dice and is the only one who sees the outcome. After looking at the outcome, the sender chooses a message to send to the receiver among six possible messages: "The outcome was i", with $i \in \{1,2,3,4,5,6\}$. After receiving the message, the receiver has to guess the true outcome of the dice roll. If the receiver guesses the true outcome, then Option A is implemented as a payment; if the receiver fails to guess the true outcome, then Option B is implemented.

Although, in principle, this game has six strategies for each player, it can be reduced to a game with two strategies for each player in an obvious way. The sender has indeed essentially two strategies: he either tells the truth to the receiver about the outcome of the dice, or he lies. Similarly, also the receiver has essentially two strategies: she either believes the message sent by the sender, or not: if the receiver believes the sender, she reports the same number as the one sent by the sender; otherwise, if the receiver does not believe the sender, she draws randomly a number from the remaining five numbers of the dice.

Therefore, we can write the payoff matrix of the sender-receiver game as follows. Let $A = (a_R, a_S)$ and $B = (b_R, b_S)$ be the payoffs associated to Option A and Option B, respectively, where S stands for the sender and R stands for the receiver. If the number chosen by the receiver is equal to the actual outcome of the dice, the sender gets the payoff a_S, and the receiver gets the payoff a_R. Conversely, if the number chosen by the receiver is not equal to the actual outcome of the dice, the sender gets the payoff b_S, and the receiver gets the payoff b_R.

Without loss of generality, we can reduce the number of parameters from four to two by setting $a_S = a_R = 0$. Finally, by setting $s = b_S$ and $r = b_R$, we can rewrite the game in a 2×2 matrix form, as follows

	B	N
T	$0, 0$	s, r
L	s, r	$\frac{4}{5}s$,$\frac{4}{5}r$

where T stands for "Truth", L stands for "Lie", B stands for "Believe", and N stands for "Not Believing". The ratios $\frac{4}{5}$ come from the fact that, when the sender lies and the receiver does not believe the message sent by the sender, then the receiver does not guess the true outcome of the dice with probability $\frac{4}{5}$.

Following the taxonomy introduced by Erat and Gneezy [53], we distinguish four types of lies, depending on the consequences in payoffs:

- Pareto white lies are those that benefit both the sender and the receiver: $r, s > 0$.
- Altruistic white lies are those that benefit the receiver at a cost to the sender: $r > 0$, $s < 0$.
- Black lies are those that benefit the sender at a cost to the receiver: $r < 0$, $s > 0$.
- Spiteful lies are those that harm both the sender and the receiver: $r, s < 0$.

We conclude this section by reporting the equilibrium analysis. If $r, s < 0$, there are two equilibria in pure strategies, (T, B) and (L, N), and one equilibrium in mixed strategies $(T/6 + 5L/6, B/6 + 5N/6)$ – that is, the sender plays T with probability $1/6$ and plays L with probability $5/6$; analogous for the receiver. If $sr < 0$ (i.e., if $r > 0$ and $s < 0$ or $s > 0$ and $r < 0$), then there are no equilibria in pure strategies and there is one equilibrium in mixed strategies, that is, again, $(T/6 + 5L/6, B/6 + 5N/6)$. Finally, if $s, r > 0$, there are two equilibria in pure strategies, (T, N), (L, B), and one equilibrium in mixed strategies, again, $(T/6 + 5L/6, B/6 + 5N/6)$. The cases $r = 0$ and/or $s = 0$ are trivial, because the corresponding player/s is/are indifferent between the strategies.

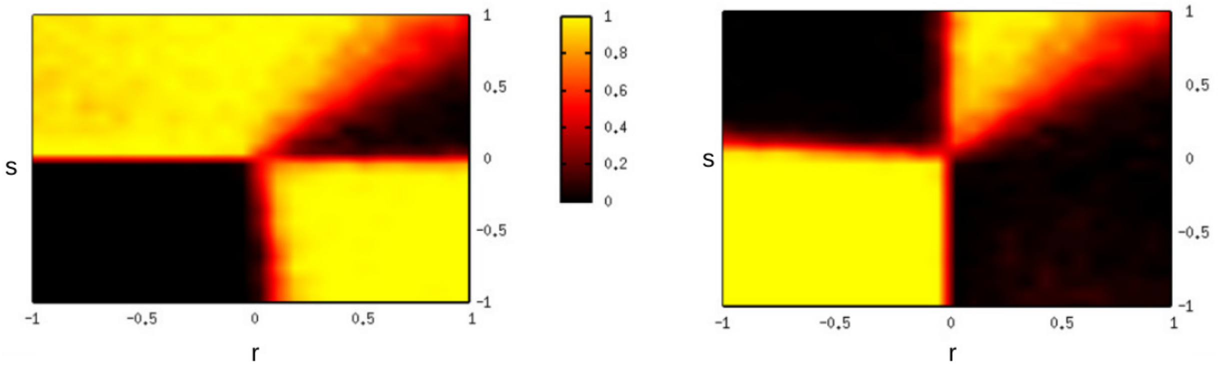

FIG. 1: Density of liars (left panel) and believers (right panel) in the steady state. In the domain of spiteful lies, all senders are honest and all receivers believe the sender's message. In the domain of altruistic white lies and black lies, most senders lie and most senders do not believe the sender's message. However, the exact final frequencies depend on the specific parameters. In the domain of Pareto white lies, the steady state depends significantly on the parameter values. System size used is $N = 500$, and the results are averaged over 2000 independent realizations.

The Monte Carlo method

We consider the sender-receiver game among N players, who interact pairwise in a well-mixed population. At each round of the game, one player acts as a sender, and the other player acts as a receiver. Each player can assume either role, which is decided by a coin toss at the start of each encounter. When acting as a sender, a player can either tell the truth (T) or lie (L). When acting as a receiver, on the other hand, a player can either believe (B) the message received from the sender, or not (N). This gives rise to four different strategies, namely (T, B), (T, N), (L, B), and (L, N). Initially, each player is randomly assigned as either T or L (when she acts as a sender), and as either B or N (when she acts as a receiver).

We simulate the game using the Monte Carlo method. For a well-mixed population with N players, the following elementary steps apply. First, a player x is randomly drawn from the population. Player x then plays the sender-receiver game with four randomly chosen other players from the population in a pairwise manner as described above, thereby obtaining the payoff π_x. Secondly, another player y is also randomly drawn from the population, and he also plays the sender-receiver game with four randomly chosen other players from the population, thereby obtaining the payoff π_y. Lastly, player y imitates the strategy of player x in accordance with the probability $w = \{1 + \exp[(\pi_y - \pi_x)/K]\}^{-1}$, where K quantifies the uncertainty during the strategy adoption process. In the $K \to \infty$ limit, payoffs cease to matter and strategies change at random; conversely, in the $K \to 0$ limit, player y imitates x only if $\pi_x > \pi_y$; between these two limits, the strategies of better performing players tend to be imitated, although underperforming strategies are imitated as well, for example due to errors in the decision making, imperfect information, and external influences that may adversely affect the evaluation of the payoff of the other player. Without loss of generality, here we set $K = 0.1$, in agreement with previous research that showed this to be a representative value [36].

The time is measured in Monte Carlo steps (MCS), whereby one MCS corresponds to executing all three elementary steps N times. During one MCS, each player changes strategy, on average, only once. For a systematic numerical analysis, we have determined the fraction of strategies in the final stationary state when varying the values of s and r. For an adequate accuracy, we have used sufficiently large system sizes, varied from $N = 500$ to 1000, as well as long enough thermalization and sampling times, varied from 10^4 to 10^6 MCS. To further remove statistical fluctuations, we have also averaged the final outcome over up to 2000 independent realizations.

Results

We considered a well-mixed population and investigated the final configuration reached by the system once the dynamics has reached its steady state.

Final densities of liars and believers as a function of lie type

As a first step of our analysis, we look at the final densities of liars and believers, as functions of the type of lie.

Figure 1 shows the final densities of liars (left panel) and believers (right panel), as functions of the game parameters (r, s). For each couple (r, s), the corresponding densities are obtained by averaging over 2,000 independent realizations on a system of size $N = 500$. The simulations were conducted with r, s increasing from -1 to 1, with steps of length 0.08. We verified that the dynamics has actually reached the final state.

We start from the case $r, s < 0$. The left panel highlights that, in this case, all senders are honest, whereas the right panel puts in evidence that all receivers believe the message sent by the sender. This result is not a priori obvious. The case $r, s < 0$ corresponds to spiteful lies, in which both the sender and the receiver are harmed by a lie that is believed. As we have seen before, in this domain, the sender-receiver game has three equilibria (T, B), (L, N), and $(\frac{1}{6}T + \frac{5}{6}L, \frac{1}{6}B + \frac{5}{6}N)$.

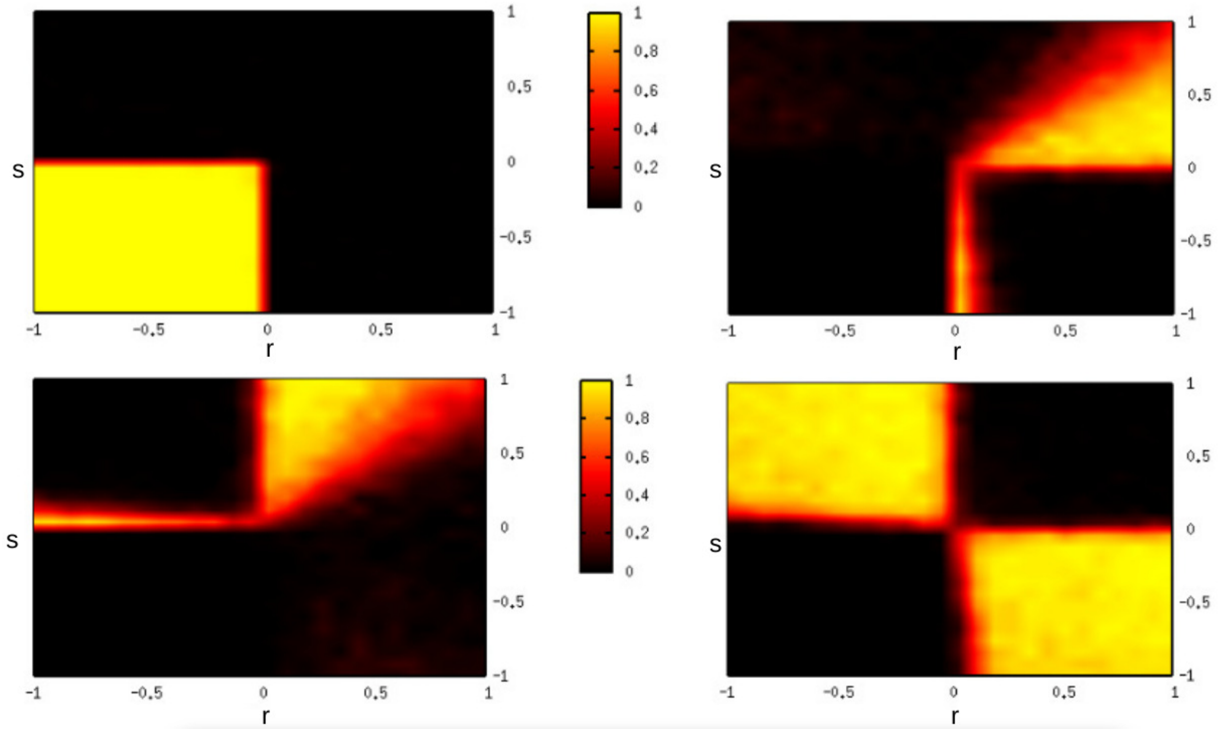

FIG. 2: Upper-left panel: Final densities of the pure strategy profile (T, B), which turns out to evolve only in the domain of spiteful lies. Upper-right panel: Final densities of the pure strategy profile (T, N), which turns out to evolve in three cases, namely, for altruistic white lies, for black lies, and for Pareto white lies, although with different frequencies depending on the exact parameter values. Lower-left panel: Final densities of the pure strategy profile (L, B), which also turns out to evolve in the domains of altruistic white lies, black lies, and Pareto white lies, but with different frequencies depending on the exact parameter values. Lower-right panel: Final densities of the pure strategy profile (L, N), which turns out to evolve only in the domains of altruistic white lies and black lies, and, in both cases, with very high frequencies.

The simulations show that two of these equilibria are discarded and all agents tend to coordinate on (T, B). A theoretical reason for why this happens is that this equilibrium is the only one that is Pareto optimal in that it maximizes the payoff for both players. Therefore, (T, B) is the strategy that has the most chances to be imitated. Also note that, as shown in this figure (see also the upper-left panel of Figure 2), the finding that only the (T, B) equilibrium survives in the evolution is robust to changing the payoff parameters, r and s, as long as they remain in the domain of spiteful lies. In other words, in the domain of spiteful lies, senders quickly learn that their best strategy is to report the truth, while receivers quickly learn that their best strategy is to believe the sender's message.

Now, keeping $r < 0$ constant, we note that, when s increases and overcomes zero, there is a state transition, which corresponds to the fact that the parameters (r, s) enter the domain of black lies, where, assuming that receivers believe the senders' messages, it is favorable for senders to lie. This has the effect that lying tends to spread. However, since, in the domain of black lies, receiver's best response to lying (L) is to not believe the sender's message (N), while L emerges, also N emerges. The emergence of N in turn contrasts the emergence of L among senders, because, in the domain of black lies, senders' best response to N is telling the truth (T). This opposite dynamics result in a mixed steady state in which most, but not all, senders lie, and most, but not all, receivers, do not believe the sender's message. One might at this point wonder whether this stationary state is equal to the unique mixed strategies equilibrium, and, in particular, whether it is independent of the parameters (r, s), or not. The answers are negative. We will show in the next sections that, in fact, the steady state depends on the parameters (r, s) non-trivially.

A similar logic applies when we keep $s < 0$ and let r increase from -1 to 1. As soon as r becomes positive, there is a state transition corresponding to the fact that the parameters (r, s) enter the domain of altruistic white lies. In this domain, assuming that receivers believe that senders tell the truth, then it is favorable for receivers to not believe the sender's message. This has the effect that strategy N tends to emerge. However, since in the domain of altruistic white lies, sender's best response to N is L, the emergence of N is contrasted by the emergence of L. This opposite dynamics result in a mixed state state, which, again, depends non-trivially from the exact parameters (r, s) as we will show in the next sections.

The quadrant in which both r and s are positive is the more variegate one. These parameters correspond to Pareto white lies, lies that benefit both the sender and the receiver. The resulting dynamics is quite complex and the steady state highly depend on the parameters (r, s), and both L and N can span all possible frequency values from 0 to 1, in a monotonic way: keeping r constant, the final frequencies of L and B both increase with s.

Density of the pure strategies

In the previous section, we have reported the final densities of liars and believers as a function of the type of lie. However, liars can come in two forms: liars who, when playing in the role of the receiver, believe the sender's message and liars who, when playing in the role of the receiver, do not believe the sender's message. Similarly, believers can come in two forms: believers who, when playing in the role of the sender, send a truthful message and believers who, when playing in the role of the sender, send a deceptive message. To gain insights about which strategies are more likely to evolve, in this section we report and discuss the final densities of the four pure strategy profiles (T, B), (T, N), (L, B) and (L, N).

The upper-left panel of Figure 2 highlights that the strategy profile (T, B), according to which a player reports the truth when acting as a sender and believes the sender's message when acting as a receiver, appears in the steady state only for $r, s < 0$ (spiteful lies). In all other types of lie, the pure strategy profile (T, B) never evolves.

Particularly interesting is the strategy profile (T, N), according to which a player tells the truth when acting as a sender, but does not believe the sender's message, when acting as a receiver. This situation is similar to what Sutter [59] termed "sophisticated deception", telling the truth while expecting to not be believed. The upper-right panel of Figure 2 highlights that this strategy profile appears in a number of non-trivial cases. When s is negative and r is positive and close to zero (T, N) appears with high probability, close to 1. This case corresponds to altruistic white lies that have a very small cost for the sender. Instead, when r is negative and s is positive (black lies), (T, N) emerges, but it does so with very small probability. In the domain of Pareto white lies ($r, s > 0$), (T, N) almost always emerges (especially for $r \geq s$). In particular, when r gets close to 1 and s is between 0 and 0.5, (T, N) emerges with very high probability, close to 1.

The case (L, B) is symmetric to the case (T, N). The lower-left panel of Figure 2 shows that this strategy profile does not emerge at all in the domain of spiteful lies ($r < 0, s < 0$) and it emerges with small probability in the domain of altruistic white lies ($r > 0, s < 0$). In the domain of black lies ($r < 0, s > 0$), we note a fast emergence of the strategy (L, B) for small values of s, close to 0, in which this strategy profile evolves even with probability close to 1. However, for larger values of s it quickly vanishes. Again, the domain of Pareto white lies is the more variegate one. Indeed, in this case, the strategy profile (L, B) emerges with high probability when $s \geq r$, whereas for $s < r$, its probability is very small.

Finally, the lower-right panel of Figure 2 shows that the strategy profile (L, N) does not emerge in the domains of spiteful lies and Pareto white lies, but it does emerge in the domains of altruistic white lies and black lies, with very high, although not equal to 1, probabilities.

Sections

We have said earlier that, in the domains of black lies ($r < 0$, $s > 0$) and altruistic white lies ($r > 0$, $s < 0$), the steady state depends on the specific values of r and s in a non-trivial way, and that, in particular, it is not equal to the unique Nash equilibrium of the game, $(\frac{1}{6}T + \frac{5}{6}L, \frac{1}{6}B + \frac{5}{6}N)$. Here we show this interesting fact by reporting the dynamics along the two sections $r = \pm 0.50$, as functions of the sole parameter s.

We start by setting $r = -0.50$. When $s < 0$, we have already seen in the previous section that the only strategy profile that survives is (T, B). This is indeed reflected in Figure 3 (left), which puts in evidence that, in this region, the frequency of (T, B) (green line) is equal to 1, whereas all other frequencies are equal to 0. Then, when s becomes positive, there is a sudden change of state. Interestingly, liars quickly emerge, but in a non-symmetric way: the frequency of (L, B) quickly increases up to almost 1 for $s \simeq 0.01$, as shown in the inset of Figure 3 (left), then it quickly decreases again to 0. On the other hand, the frequency of (L, N) rapidly increases up to around 0.9, and then slowly keeps increasing up to reaching a value near 1. The maximum of the frequency of (L, B) is rather surprising for its narrowness: the final density of (L, B) is 0 for $s < 0$; then it quickly increases for positive but very small values of s; then it quickly decreases again to 0. To better understand this peculiar behavior, in Figure 4 we report the time series of each strategy in the interval of (L, B) dominance. Specifically, the left panel of Figure 4 highlights that the frequency of (L, B) increases monotonically up to near 1, while all other strategies tend to appear with very small frequencies, although their evolution is rather different. In particular, (L, N) evolves non-monotonically, while (T,N) is even oscillatory. The right panel of Figure 4 reports the evolution of liars and believers in the same interval of (L, B) dominance. (More details about the time evolution of the various densities will be given in the next section.) Regarding truth-telling, the strategy (T, B), which was the only surviving strategy for $s < 0$, in the domain $s > 0$ completely vanishes. On the other hand, the strategy (T, N) emerges in a non-monotonic way: as $s > 0$ increases, the frequency of (T, N) first increases up to a value around 0.1, and then slowly decreases to values near 0. Therefore, for $r = -0.5$ and $s > 0$, receivers never believe the sender's message, while senders lie with high frequency, but not equal to 1.

The case $r = 0.50$ is somewhat more articulated, as shown in Figure 3 (right). When $s < 0$, liars emerge with frequency 1, however, this does not appear to be due to the emergence of a single profile of strategies. Indeed, for $s < 0$ we see a coexistence of the strategy profiles (L, B) and (L, N), although the latter one appears to emerge with higher frequency, especially when s increases and approaches 0, in which (L, N) reaches frequencies very close to 1. Then, as soon as s reaches 0, there is a change of state: the strategy profile (T, N) appears with frequency very close to 1; however, as s increases towards 1, then (T, N) appears with lower and lower frequencies. This decrease of the frequency of appearance of (T, N), as s increases, appears to be perfectly mirrored by an increase of the frequency of (L, B).

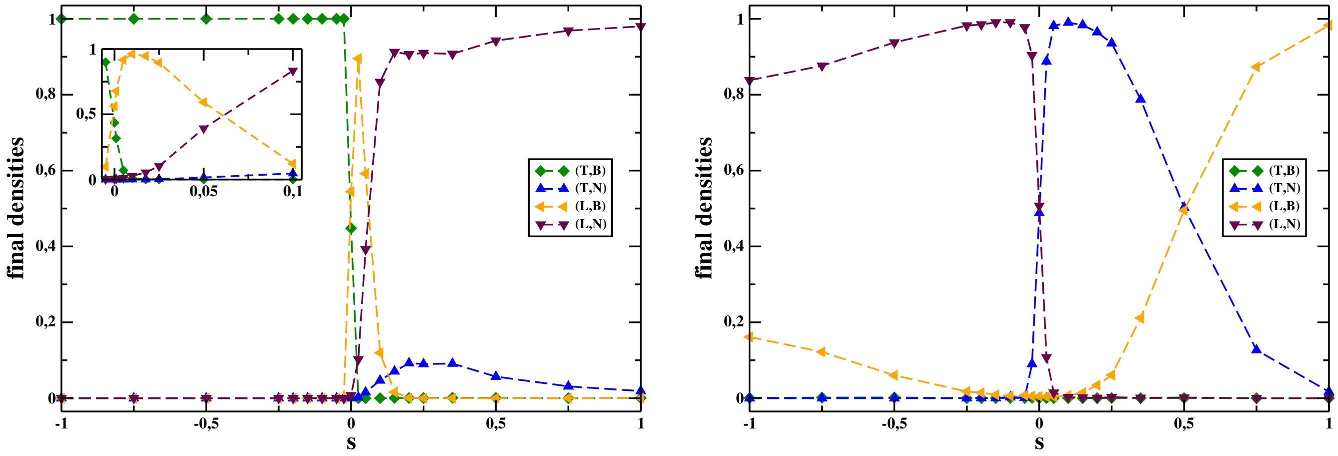

FIG. 3: Left panel: Final densities of different strategies as a function of the parameter s, for fixed $r = -0.50$. When $s < 0$, only the strategy (T, B) survives. For $s > 0$, (L, B) quickly increases to around 0.9 and then it quickly decreases to 0; (L, N) quickly increases up to around 0.9, and then slowly keeps increasing up to reaching values close to 1; (T, B) completely vanishes; (T, N) first emerges for small values of s, then vanishes; inset: zoom of the interval $s \in [-0.005, 0.1]$. Right panel: Final densities of the different strategies as a function of the parameter s, for fixed $r = 0.50$. For $s < 0$, only (L, B) and (L, N) emerge, although the latter with much higher probability. For $s > 0$, (T, N) quickly emerges, but then it slowly disappears, contrasted by the emergence of (L, B). In all cases the system size used is $N = 1000$ with random initial conditions.

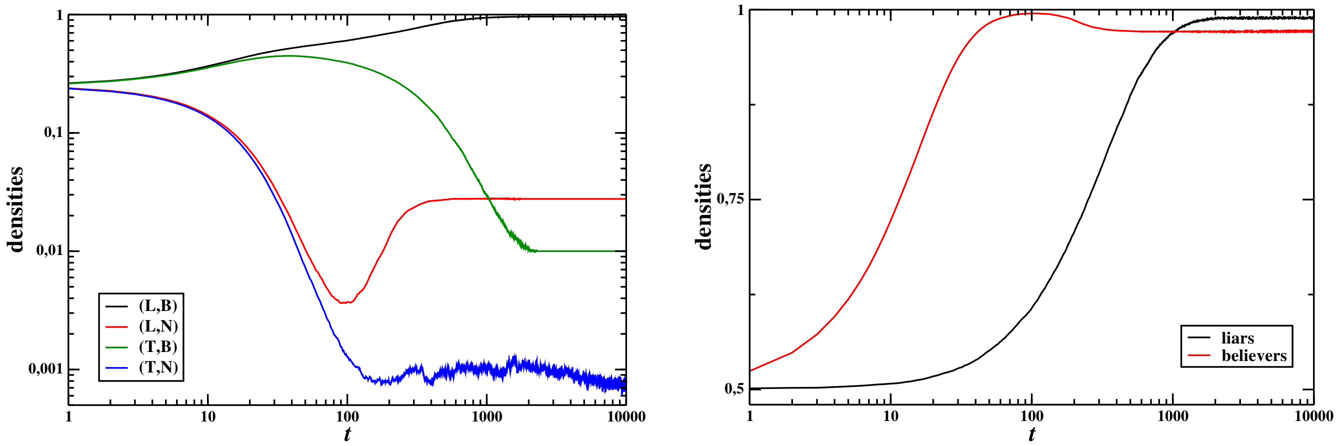

FIG. 4: Left panel: Time series of the frequencies four basic strategies for $r = -1$ and $s = 0.01$, that is, around the (L, B) maximum, see Figure 3 (left). The frequency of (L, B) increases monotonically up to near 1, while all other strategies tend to appear with very small frequencies, although their evolution is rather different. In particular, (L, N) evolves non-monotonically, while (T, N) is even oscillatory. Right panel: Time series of the frequencies of liars and believers for the same parameter values utilized in the left panel. System of size $N = 1000$ with random initial conditions.

Time evolution

We conclude by reporting the time evolution of liars and believers at the corner of the domain of the parameters (r, s). We verified the time evolution also for other values of (r, s), and we found qualitatively similar patterns (as long as $r, s \neq 0$, clearly).

Figure 5, left and right, highlight that, before reaching the steady state, the evolution is interesting, being sometimes monotone and sometimes not. For $r = 1$ and $s = -1$ (red line, altruistic white lie), we note that both the behavior of senders and the behavior of receivers evolve in a non-monotone way. Similarly, for $r = 1$ and $s = 1$ (blue line, Pareto white lie), the behavior of both senders and receivers evolve non-monotonically. A non-monotone evolution, although less remarked, appears also in the case of black lies ($r = -1$, $s = 1$, green line). Conversely, in the case of spiteful lies, we see a very quick convergence to the strategy (T, B), in line with the discussion above that, in this case, senders quickly learn that their best strategy is to tell the truth and receivers quickly learn that their best strategy is to believe the sender's message.

Figure 6 reports in more detail the time evolution of the four basic strategies for $r = 1, s = \pm 1$, that is, when the densities of liars and believers evolve non-monotonically. In the case of Pareto white lies (left panel), we note that the non-monotonic

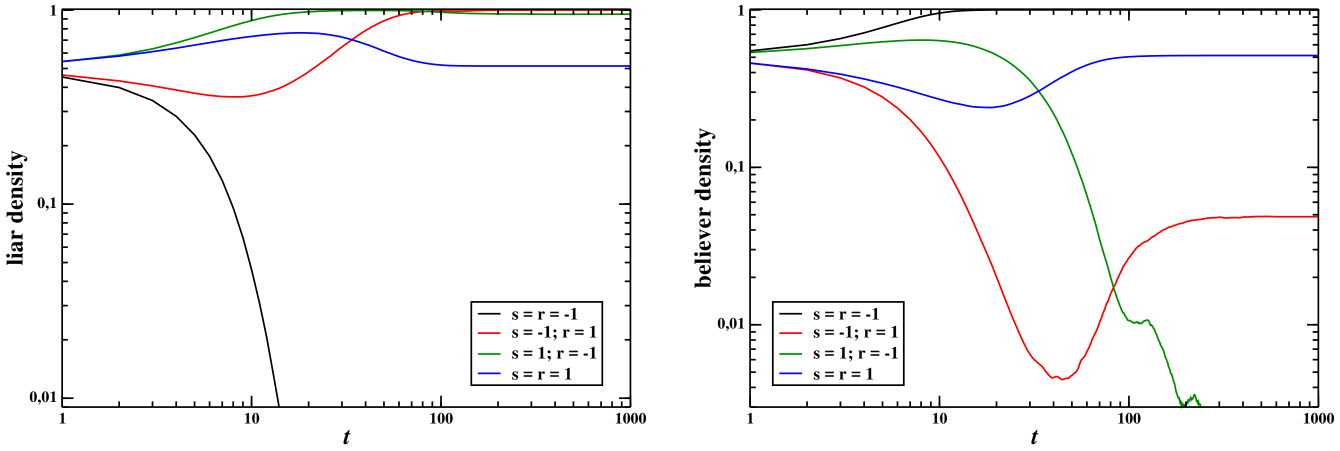

FIG. 5: Left panel: Time evolution of liars at the corners of the domain of the parameters (r, s). The evolution is monotone only in the case of $r = -1$ and $s = -1$ (spiteful lies). Right panel: Time evolution of believers at the corners of the domain of the parameters (r, s). Time evolution of liars at the corners of the domain of the parameters (r, s). The evolution is monotone only in the case of $r = -1$ and $s = -1$ (spiteful lies). In all cases the system size used is $N = 500$ with random initial conditions.

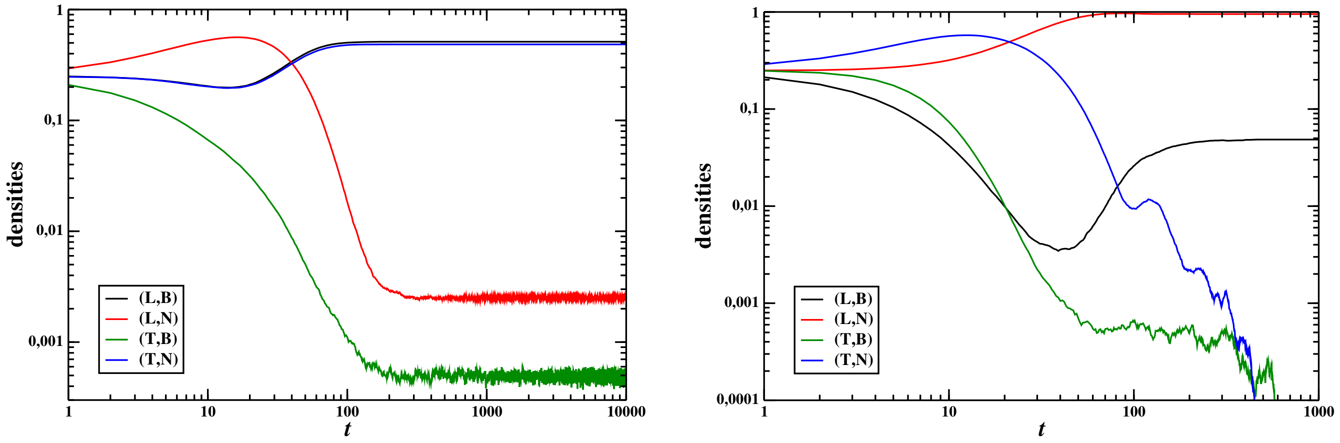

FIG. 6: Left panel: Time evolution of the four pure strategy profiles for $r = 1$ and $s = 1$ (Pareto white lies). The non-monotonic evolution of liars is primarily driven by a non-monotonic evolution of the strategy (L, N). The non-monotonic evolution of believers is driven by a combination of (T, B) and (L, B). Right panel: Time evolution of the four pure strategy profiles for $r = 1$ and $s = -1$ (altruistic white lies). The non-monotonic evolution of liars and believers is mainly driven by a non-monotonic evolution of the strategy (L, B). Systems of size $N = 500$ with random initial conditions.

evolution of liars is primarily driven by a non-monotonic evolution of the strategy (L, N), whose frequency first increases up to about 0.8 and then suddenly decreases of two orders of magnitudes, to values below 0.01, and then keeps oscillating. Similarly, still in the domain of Pareto white lies, the non-monotonic evolution of believers is driven by a combination of (T, B) and (L, B): at the beginning of the dynamics, the frequency of (L, B) is approximately constant, while the frequency of (T, B) decreases, giving rise to the initial decrease of believers observed in the right panel of Figure 5; then, between $t \simeq 20$ and $t \simeq 100$, the frequency of (T, B) doubles from about 0.4 to about 0.8, where it stabilizes, while the frequency of (T, B) keeps decreasing. After $t \simeq 100$, the frequency of (T, B) starts alternating. This change in the dynamics contributes to the overall non-monotonicity observed in the evolution of the frequency of believers. A similar line of reasoning holds in the case of altruistic white lies. As shown in the right panel of Figure 6, the non-monotonic evolution of liars and believers is mainly driven by a non-monotonic evolution of the strategy (L, B).

Finally, it is worth noticing that the non-monotonic behaviour in time increase with the population size: indeed, for very large systems ($N \gtrsim 10^4$), in some cases we observe oscillations before the densities reach the final state.

Discussion

We have used the Monte Carlo method to explore the evolution of lying in well-mixed populations, where individuals are playing the sender-receiver game [53, 57]. We have shown that the evolution follows non-trivial trajectories. In particular, honesty and dishonesty may appear or disappear with very high probability depending on the particular payoffs of the game. Similarly, also believing and non-believing can emerge or vanish with very high probabilities. More specifically, following Erat and Gneezy's taxonomy of lies [53], we distinguished four basic types of lies: black lies, spiteful lies, altruistic white lies, and Pareto white lies. In the domain of spiteful lies, senders quickly learn that their best strategy is to send a truthful message, and receivers quickly learn that their best strategy is to believe the sender's message. The cases of altruistic white lies and black lies are instead characterized by the fact that, at the steady state, most senders lie while most receivers do not believe the sender message. However, the exact proportions of senders and non-believers depend significantly on the particular payoffs, and they also evolve in a non-monotonic way, before eventually reaching the steady state. The case of Pareto white lies is an even more variegate one. Here, the steady state depend fully on the payoffs, and both lying and non-believing can evolve with all probabilities between 0 to 1.

Previous research has explored the evolution of honesty using the Philip Sidney game [58]. In this game, the Sender is initially in either of two states, healthy or needy, with probability p and $1-p$, respectively. The Sender can either pay a cost c to signal his state or stay quiet. The Receiver does not know the state of the Sender, but can observe the signal. After observing the signal (if sent), the Receiver decides whether to donate his resource to the Sender. The Sender and the Receiver are assumed to be related, by a relatedness coefficient r. Each player's payoff is the sum of his survival probability and a fraction r of the other player's survival probability. Survival probabilities are defined as follows: the Receiver is sure to survive only if he does not donate his resource; the Sender is sure to survive only if he receives the Receivers resource. This creates a conflict of interests among the Sender and the Receiver which corresponds to what we called (following Erat and Gneezy [53]) the "black lie" condition. A classic work on the Philip Sidney game found that, if the cost of the signal is sufficiently high, then honest signalling can evolve [60]. See [61] for a review of this "Handicap Principle" and its variants. More recent research revealed that punishment can promote the evolution of honesty in cases in which the conflict of interests among the Sender and the Receiver is moderate and signalling is cheap or even cost-free [62]. Our work departs from this line of research along two main dimensions. First, in the Sender-Receiver game, signalling is cost-free and there is no punishment. Even in this case, our results indicate that honesty can evolve in some circumstances (especially in the case of spiteful lies and Pareto white lies, but also, to some extent, in the case of black lies). Second, the Sender-Receiver game allows to study the evolution of honesty not only in the domain of black lies, but also in the domains of spiteful lies, Pareto white lies, and altruistic white lies.

Related to our work is also the recent literature on precommitments in social dilemmas. In this context, a social dilemma is preceded by a pre-play stage in which players can send messages (commitment proposals) and other players can accept or refuse the proposal. Proposers can lie about the commitment. For example, after promising that they would cooperate, proposers can dishonour their promise and defect. On the other hand, responders can refuse a commitment proposal because they do not believe the proposer. Han and colleagues explored analytically and numerically the evolution of cooperation in this type of social dilemmas, both in pairwise [63] and group interactions [64, 65], and found that cooperation can evolve under a number of different circumstances, such as for example when the cost of commitment is sufficiently small compared to the cost of cooperation. Our work differs from this line of research in that we focus specifically on honesty and believing, with no consequences on cooperative behaviour. This allows us to clearly identify the four classes of lies (black, spiteful, altruistic, Pareto), and to study the evolution of lying as a function of lie type.

Statistical physics, and, in particular, the Monte Carlo method, has proven valuable for the study of the evolution of cooperation in social dilemmas [36]. Yet, cooperation in social dilemmas is only one particular instance of a more general class of behaviors, moral behaviors [37]. Therefore, it is time now to move beyond the borders of cooperation and start applying similar methods to the evolution of other moral behaviors, such as, indeed, honesty [38]. To the best of our knowledge, this is the first study using techniques from statistical physics to study the evolution of lying in the six-dice sender-receiver game. Of course, some questions remain to be addressed in future research, such as: What happens for general n-dice sender-receiver games? What happens on networks? What interventions can be done to favor the evolution of honesty? What if imitation is replaced with other forms of strategy change? Just to name a few. These are important questions, whose answers can greatly contribute to the improvement of the society we live in, and they can provide a nuanced quantitative view of honest behavior, as well as inform the design of future human experiments with testable theoretical predictions.

Extending the domain of application of the Monte Carlo method from cooperation to honesty, our work also suggests that similar techniques could be applied to study the evolution of other forms of moral behavior. A recent work by Curry *et al.* [66] shows that seven moral rules are universal across societies: love your family, help your group, return favors, be brave, defer to authority, be fair, and respect others' property. Clearly, not all these behaviors can be studied using simple games, but some are. For instance, "returning favors" could be studied using a sequential prisoner's dilemma or the trust game; "help your group" could be studied using games with labeled players, in which individuals come with a label describing the group they belong to; "fairness" could be studied through the ultimatum game, as indeed has already been done [67–77]; respect others' property can be studied utilizing games with special frames, as, for example, the dictator game in the take frame, for which taking turns out to be considered more morally wrong than giving [78, 79].

In sum, we believe that illuminating if, when, and how techniques of statistical physics can be applied to study the evolution of morality among humans, should be considered as a primary direction for future research.

Acknowledgments

Matjaž Perc was supported by the Slovenian Research Agency (Grants J4-9302, J1-9112 and P1-0403). Daniele Vilone was supported by the European Union's Horizon 2020 Project PROTON (Grant 699824).

[1] S. Pinker, *The better angels of our nature: Why violence has declined*, vol. 75 (Viking New York, 2011).

[2] S. Pinker, *Enlightenment now: The case for reason, science, humanism, and progress* (Penguin Books, 2019).

[3] Y. Arthus-Bertrand, *Human (movie)* (Bettencourt Schueller Foundation, Neuilly-sur-Seine, France, 2014).

[4] R. K. Garrett, Journal of Computer-Mediated Communication **14**, 265 (2009).

[5] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, Proceedings of the National Academy of Sciences **113**, 554 (2016).

[6] M. A. Nowak and R. Highfield, *SuperCooperators: Altruism, Evolution, and Why We Need Each Other to Succeed* (Free Press, New York, 2011).

[7] D. Kennedy and C. Norman, Science **309**, 75 (2005).

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

[9] R. Axelrod, *The Evolution of Cooperation* (Basic Books, New York, 1984).

[10] E. Ostrom, Journal of Economic Perspectives **14**, 137 (2000).

[11] J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, and R. McElreath, Am. Econ. Rev. **91**, 73 (2001).

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

[13] E. Fehr and S. Gächter, Nature **415**, 137 (2002).

[14] H. Gintis, S. Bowles, R. Boyd, and E. Fehr, Evolution and Human Behavior **24**, 153 (2003).

[15] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll, Behavioral and Brain Sciences **28**, 675 (2005).

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

[17] S. Bowles and H. Gintis, *A Cooperative Species: Human Reciprocity and Its Evolution* (Princeton University Press, Princeton, NJ, 2011).

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

[19] V. Capraro, PLoS ONE **8**, e72427 (2013).

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

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

[22] E. Estrada, *The structure of complex networks: theory and applications* (Oxford University Press, Oxford, 2012).

[23] S. Boccaletti, G. Bianconi, R. Criado, C. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin, Phys. Rep. **544**, 1 (2014).

[24] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, J. Complex Netw. **2**, 203 (2014).

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

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

[27] G. Szabó and G. Fáth, Phys. Rep. **446**, 97 (2007).

[28] M. Perc and A. Szolnoki, BioSystems **99**, 109 (2010).

[29] M. Perc, J. Gómez-Gardeñes, A. Szolnoki, and L. M. Floría and Y. Moreno, J. R. Soc. Interface **10**, 20120997 (2013).

[30] Z. Wang, L. Wang, A. Szolnoki, and M. Perc, Eur. Phys. J. B **88**, 124 (2015).

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

[32] F. Giardini and D. Vilone, Scientific Reports **6**, 37931 (2016).

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

[34] 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).

[35] M. Perc, Phys. Lett. A **380**, 2803 (2016).

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

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

[38] V. Capraro and M. Perc, Frontiers in Physics **6**, 107 (2018).

[39] J. Gravelle, *Tax havens: International tax avoidance and evasion* (DIANE Publishing, 2010).

[40] FBI, *Insurance fraud*, URL `https://www.fbi.gov/stats-services/publications/insurance-fraud/`.

[41] G. Pennycook, T. D. Cannon, and D. G. Rand, Journal of Experimental Psychology: General **147**, 1865 (2018).

[42] N. Mazar, O. Amir, and D. Ariely, Journal of Marketing Research **45**, 633 (2008).

[43] D. Ariely and S. Jones, *The (honest) truth about dishonesty: how we lie to everyone – especially ourselves*, vol. 336 (HarperCollins New York, NY, 2012).

[44] F. Gino, S. Ayal, and D. Ariely, Psychological Science **20**, 393 (2009).

[45] F. Gino, M. E. Schweitzer, N. L. Mead, and D. Ariely, Organizational Behavior and Human Decision Processes **115**, 191 (2011).

[46] S. Shalvi, J. Dana, M. J. Handgraaf, and C. K. De Dreu, Organizational Behavior and Human Decision Processes **115**, 181 (2011).

[47] S. Shalvi, O. Eldar, and Y. Bereby-Meyer, Psychological Science **23**, 1264 (2012).

[48] S. Shalvi, F. Gino, R. Barkan, and S. Ayal, Current Directions in Psychological Science **24**, 125 (2015).

[49] B. Verschuere, A. Spruyt, E. H. Meijer, and H. Otgaar, Consciousness and Cognition **20**, 908 (2011).

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

[51] V. Capraro, Economics Letters **158**, 54 (2017).

[52] V. Capraro, Judgment and Decision Making **13**, 345 (2018).

[53] S. Erat and U. Gneezy, Management Science **58**, 723 (2012).

[54] U. Gneezy, A. Kajackaite, and J. Sobel, American Economic Review **108**, 419 (2018).

[55] V. Capraro, J. Schulz, and D. G. Rand, Journal of Behavioral and Experimental Economics **79**, 93 (2019).

[56] U. Fischbacher and F. Föllmi-Heusi, Journal of the European Economic Association **11**, 525 (2013).

[57] U. Gneezy, American Economic Review **95**, 384 (2005).

[58] J. M. Smith, Animal Behaviour (1991).

[59] M. Sutter, The Economic Journal **119**, 47 (2009).

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

[61] S. Számadó, Animal Behaviour **81**, 3 (2011).

[62] D. Catteeuw, B. Manderick, et al., in *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation* (ACM, 2014), pp. 153–160.

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

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

[65] T. A. Han, L. M. Pereira, and T. Lenaerts, Autonomous Agents and Multi-Agent Systems **31**, 561 (2017).

[66] O. S. Curry, D. A. Mullins, and H. Whitehouse, Current Anthropology **1** (in press).

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

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

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

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

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

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

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

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

[75] L. Deng, C. Wang, W. Tang, G. Zhou, and J. Cai, Journal of Statistical Mechanics: Theory and Experiment **2012**, P11013 (2012).

[76] J. Iranzo, L. M. Floria, Y. Moreno, and A. Sanchez, PloS ONE **7**, e43781 (2012).

[77] K. Miyaji, Z. Wang, J. Tanimoto, A. Hagishima, and S. Kokubo, Chaos, Solitons & Fractals **56**, 13 (2013).

[78] E. L. Krupka and R. A. Weber, Journal of the European Economic Association **11**, 495 (2013).

[79] V. Capraro and A. Vanzo, Judgment and Decision Making (in press).