INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cite this article: Capraro V, Perc M, Vilone D. 2019 The evolution of lying in well-mixed populations. *J. R. Soc. Interface* **16**: 20190211. http://dx.doi.org/10.1098/rsif.2019.0211

Received: 25 March 2019
Accepted: 24 June 2019

Subject Category:
Life Sciences–Physics interface

Subject Areas:
evolution

Keywords:
honesty, deception, lying, evolution

Author for correspondence:
Valerio Capraro
e-mail: v.capraro@mdx.ac.uk

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.4575797.

THE ROYAL SOCIETY PUBLISHING

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, UK
[2]Faculty of Natural Sciences and Mathematics, University of Maribor, Koroška cesta 160, SI-2000 Maribor, Slovenia
[3]Complexity Science Hub Vienna, Josefstädterstraße 39, 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

VC, 0000-0002-0579-0166; DV, 0000-0002-3485-9249

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 favour the receiver at a cost to the sender, black lies that favour the sender at a cost to the receiver, spiteful lies that harm both the sender and the receiver, and Pareto white lies that favour both the sender and the receiver. We find that spiteful lies give rise to trivial behaviour, where senders quickly learn that their best strategy is to send a truthful message, while 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 behaviour in human experiments.

1. Introduction

There are arguments and data in favour 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 favours 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 twenty-first century [7], and scholars from disciplines as diverse as economics, psychology, sociology, biology and anthropology have explored what factors favour people's cooperative behaviour [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 behaviours—moral behaviours [37]. This suggests that the same methods could be applied effectively to study the evolution of other types of moral behaviours 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 society as a whole. For example, tax evasion costs approximately US$100 billion a year to the US government alone [39], whereas, still in the USA, insurance fraud costs about US$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 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 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 behaviour in general, as well as provides pointers for devising innovative human experiments to test the theory.

2. The sender–receiver game

Behavioural 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 us to study only black lies. In particular, we focus on the variant of the sender–receiver game introduced by Erat & Gneezy [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 with 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 (⅘) come from the fact that, when the sender lies and the receiver does not believe the message sent by the sender, the receiver does not guess the true outcome of the dice with probability (⅘).

Following the taxonomy introduced by Erat & 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 ⅙ and plays L with probability ⅚; 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.

3. 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 under-performing strategies are imitated as well; for example, because of 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 (MCSs), 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. The code used to conduct the simulations is reported in the electronic supplementary material.

4. Results

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

4.1. 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 (a) and believers (b), as functions of the game parameters (r, s). For each couple (r, s), the corresponding densities are obtained by averaging over 2000 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 $r, s < 0$. Figure 1a highlights that, in this case, all senders are honest, whereas figure 1b 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 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 $((1/6)T + (5/6)L, (1/6)B + (5/6)N)$. 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 figure 2a), 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 favourable for senders to lie. This has the effect that lying tends to spread. However, since, in the domain of black lies, the receiver's best response to lying (L) is to not believe the sender's message (N), while

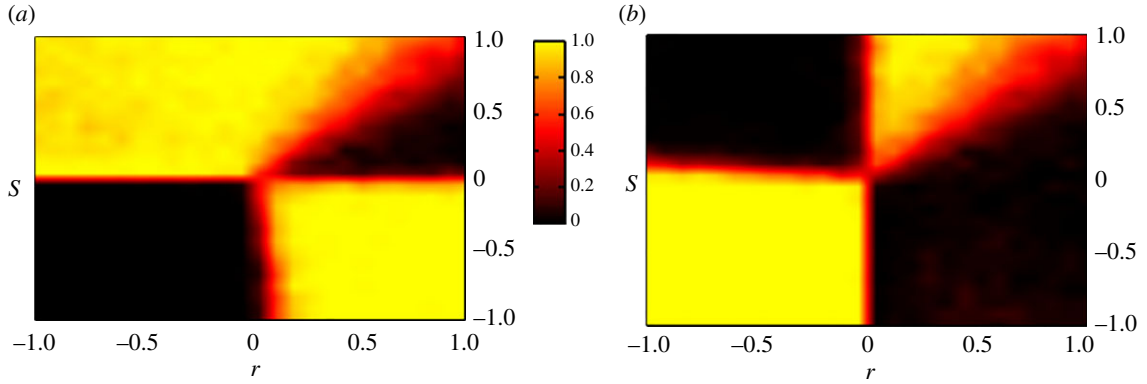

Figure 1. Density of liars (*a*) and believers (*b*) in the steady state. In the domain of spiteful lies, all senders are honest and all receivers believe the sender's message. In the domain of altruistic white lies and black lies, most senders lie and most senders do not believe the sender's message. 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.

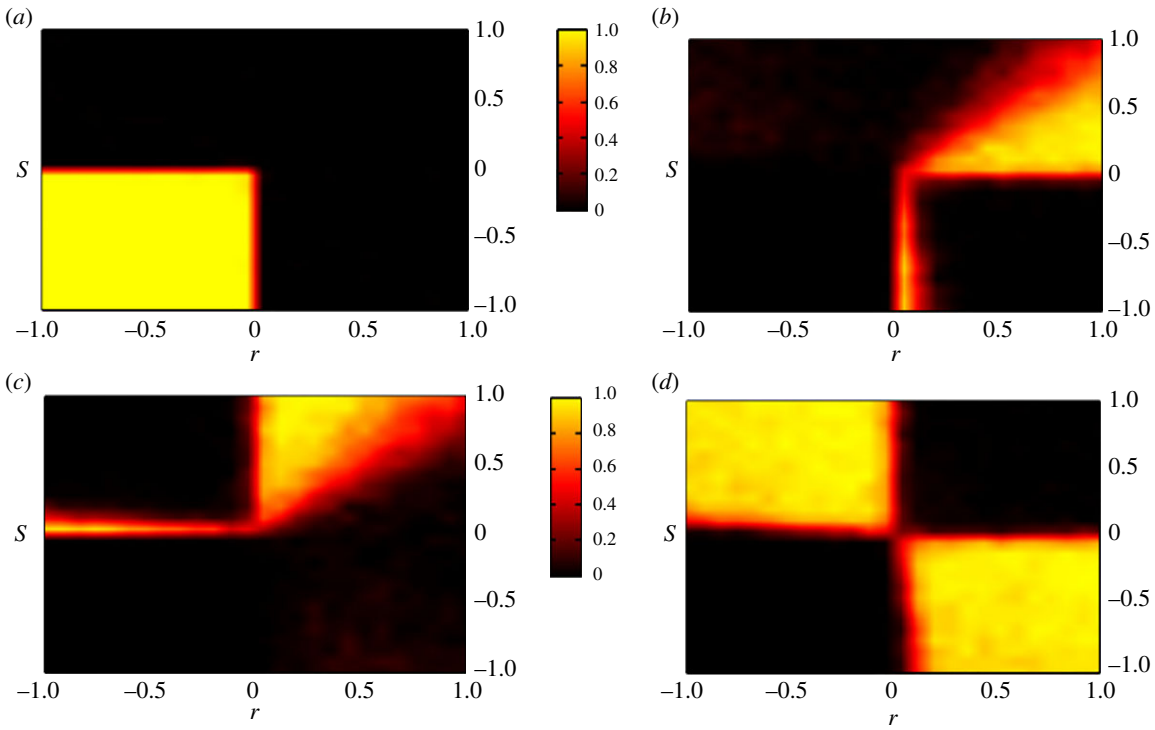

Figure 2. (*a*) Final densities of the pure strategy profile (*T*, *B*), which turns out to evolve only in the domain of spiteful lies. (*b*) Final densities of the pure strategy profile (*T*, *N*), which turns out to evolve in three cases; namely, for altruistic white lies, for black lies and for Pareto white lies, although with different frequencies depending on the exact parameter values. (*c*) 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. (*d*) 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.

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, the sender's best response to *N* is telling the truth (*T*). This opposite dynamics results 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 favourable 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, the sender's best response to *N* is *L*, the emergence of *N* is contrasted by the emergence of *L*. This opposite dynamics results in a mixed state, which, again, depends non-trivially on 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 receiver. The resulting dynamics is quite complex and the steady state highly depends on the parameters (*r*, *s*), and both *L* and *N* can span all possible frequency values from 0 to 1,

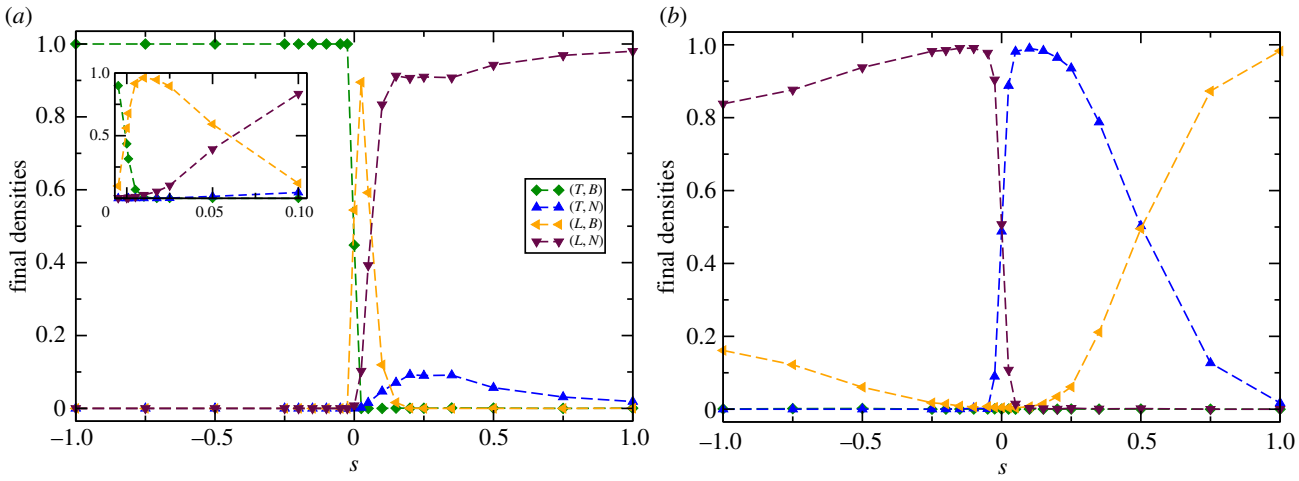

Figure 3. (*a*) Final densities of different strategies as a function of the parameter s, for fixed $r = -0.50$. When $s < 0$, only the strategy (T, B) survives. For $s > 0$, (L, B) quickly increases to around 0.9 and then it quickly decreases to 0; (L, N) quickly increases up to around 0.9, and then 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 $s \in [-0.005, 0.1]$. (*b*) 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.

in a monotonic way: keeping r constant, the final frequencies of L and B both increase with s.

4.2. 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).

Figure 2*a* 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. Figure 2*b* 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). Figure 2*c* 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, figure 2*d* 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.

4.3. 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, $((1/6)T + (5/6)L, (1/6)B + (5/6)N)$. 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*a*, 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*a*, 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

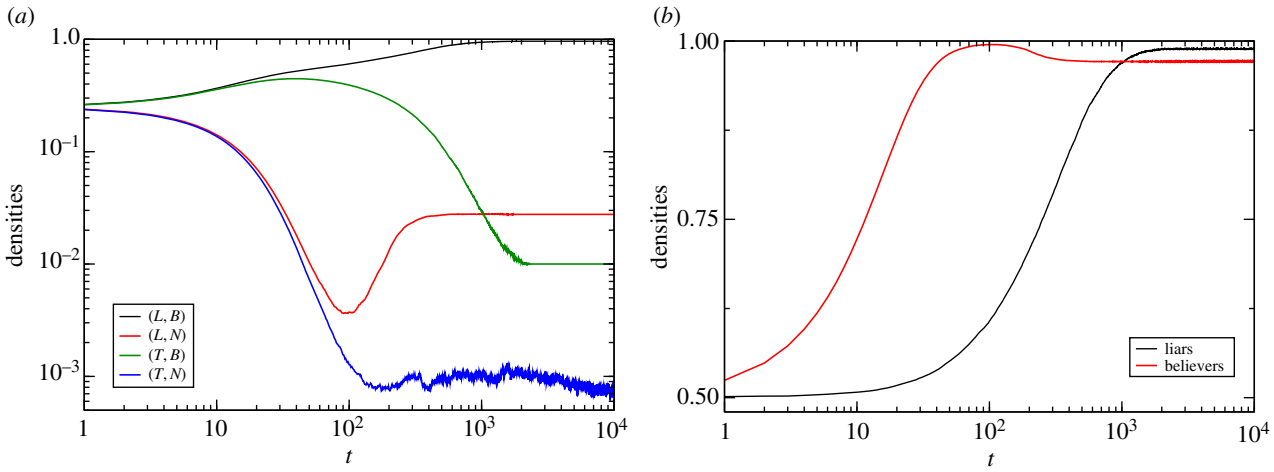

Figure 4. (*a*) Time series of the frequencies of four basic strategies for $r = -1$ and $s = 0.01$, that is, around the (L, B) maximum (figure 3*a*). The frequency of (L, B) increases monotonically up to near 1, while all other strategies tend to appear with very small frequencies, although their evolution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

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 behaviour, in figure 4, we report the time series of each strategy in the interval of (L, B) dominance. Specifically, figure 4*a* 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. Figure 4*b* 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*b*. 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 in the frequency of appearance of (T, N), as s increases, appears to be perfectly mirrored by an increase in the frequency of (L, B).

4.4. 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*a*,*b* highlights 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 behaviour of senders and the behaviour of receivers evolve in a non-monotone way. Similarly, for $r = 1$ and $s = 1$ (blue line, Pareto white lie), the behaviours 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 (figure 6*a*), we note that the non-monotonic 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 by 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 figure 5*b*; 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 figure 6*b*, 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 increases with the population size:

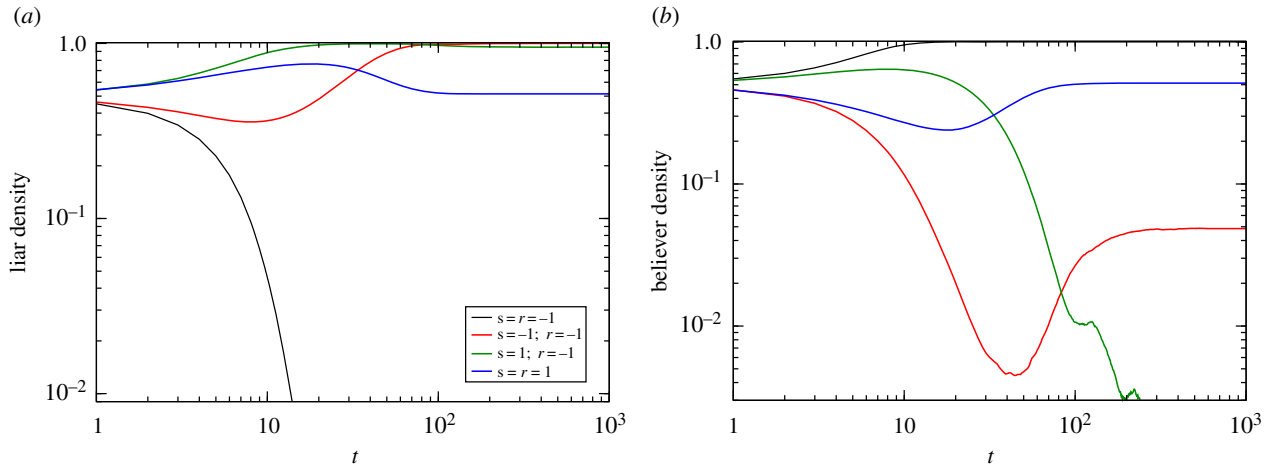

Figure 5. (*a*) Time evolution of liars at the corners of the domain of the parameters (*r*, *s*). The evolution is monotone only in the case of $r = -1$ and $s = -1$ (spiteful lies). (*b*) Time evolution of believers at the corners of the domain of the parameters (*r*, *s*). Time evolution of liars 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.

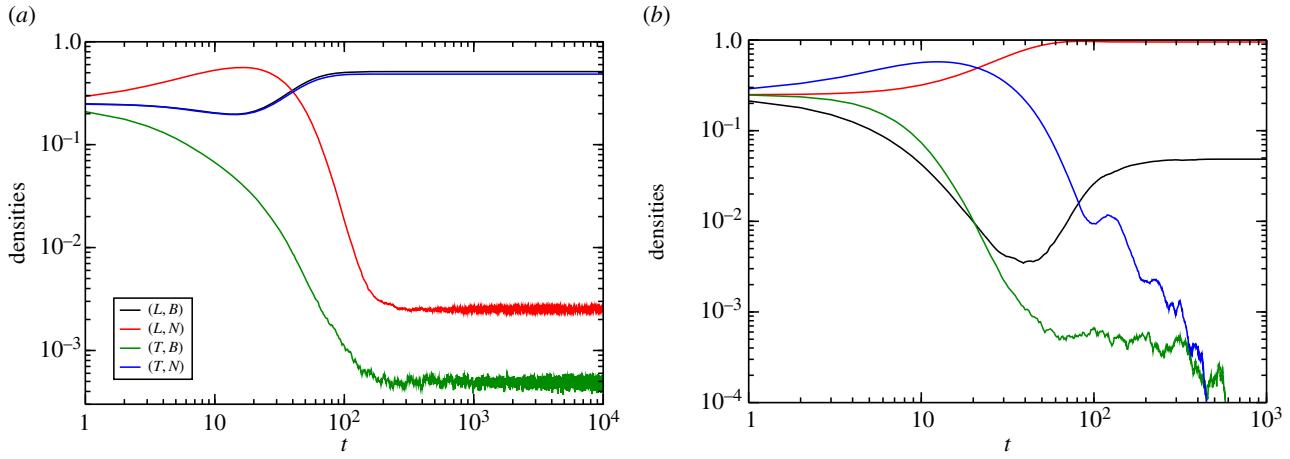

Figure 6. (*a*) Time evolution of the four pure strategy profiles for $r = 1$ and $s = 1$ (Pareto white lies). The non-monotonic evolution of liars is primarily driven by a non-monotonic evolution of the strategy (*L*, *N*). The non-monotonic evolution of believers is driven by a combination of (*T*, *B*) and (*L*, *B*). (*b*) 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.

indeed, for very large systems ($N \gtrsim 10^4$), in some cases we observe oscillations before the densities reach the final state.

4.5. 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 depends fully on the payoffs, and both lying and non-believing can evolve with all probabilities between 0 and 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 receiver's resource. This creates a conflict of interests

among the sender and the receiver which corresponds to what we called (following Erat & 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, 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 us 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 pre-commitments 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 dilemma, both in pairwise [63] and group interactions [64,65], and found that cooperation can evolve under a number of different circumstances, such as when the cost of commitment is sufficiently small compared with 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 behaviours, moral behaviours [37]. Therefore, it is time now to move beyond the borders of cooperation and start applying similar methods to the evolution of other moral behaviours, 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 favour 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 behaviour, 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 behaviour. A recent work by Curry *et al*. [66] shows that seven moral rules are universal across societies: love your family, help your group, return favours, be brave, defer to authority, be fair and respect others' property. Clearly, not all these behaviours can be studied using simple games, but some are. For instance, 'returning favours' could be studied using a sequential Prisoner's Dilemma or the trust game; 'help your group' could be studied using games with labelled 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 using games with special frames, as, for example, the dictator game in the take frame, for which taking turns out is 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.

Data accessibility. This article has no additional data.

Competing interests. We declare we have no competing interests.

Funding. M.P. was supported by the Slovenian Research Agency (grant nos. J4-9302, J1-9112 and P1-0403). D.V. was supported by the European Union's Horizon 2020Project PROTON (grant no. 699824).

Reference

1. Pinker S. 2011 *The better angels of our nature: why violence has declined*, vol. 75. New York, NY: Viking.
2. Pinker S. 2019 *Enlightenment now: the case for reason, science, humanism, and progress*. Baltimore, MD: Penguin Books.
3. Arthus-Bertrand Y. 2014 *Human (movie)*. Neuilly-sur-Seine, France: Bettencourt Schueller Foundation.
4. Garrett RK. 2009 Echo chambers online? Politically motivated selective exposure among internet news users. *J. Comput. Mediated Commun.* **14**, 265–285. (doi:10.1111/j.1083-6101.2009.01440.x)
5. Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, Stanley HE, Quattrociocchi W. 2016 The spreading of misinformation online. *Proc. Natl Acad. Sci. USA* **113**, 554–559. (doi:10.1073/pnas.1517441113)
6. Nowak MA, Highfield R. 2011 *SuperCooperators: altruism, evolution, and why we need each other to succeed*. New York, NY: Free Press.
7. Kennedy D, Norman C. 2005 What don't we know? *Science* **309**, 75. (doi:10.1126/science.309.5731.75)
8. Trivers RL. 1971 The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57. (doi:10.1086/406755)
9. Axelrod R. 1984 *The evolution of cooperation*. New York, NY: Basic Books.
10. Ostrom E. 2000 Collective action and the evolution of social norms. *J. Econ. Perspect.* **14**, 137–158. (doi:10.1257/jep.14.3.137)
11. Henrich J, Boyd R, Bowles S, Camerer C, Fehr E, Gintis H, McElreath R. 2001 In search of homo economicus: behavioral experiments in 15 small-scale societies. *Am. Econ. Rev.* **91**, 73–78. (doi:10.1257/aer.91.2.73)
12. Milinski M, Semmann D, Krambeck H-J. 2002 Reputation helps solve the 'tragedy of the commons'. *Nature* **415**, 424–426. (doi:10.1038/415424a)
13. Fehr E, Gächter S. 2002 Altruistic punishment in humans. *Nature* **415**, 137–140. (doi:10.1038/415137a)

14. Gintis H, Bowles S, Boyd R, Fehr E. 2003 Explaining altruistic behavior in humans. *Evol. Human Behav.* **24**, 153–172. (doi:10.1016/S1090-5138(02)00157-5)
15. Tomasello M, Carpenter M, Call J, Behne T, Moll H. 2005 Understanding and sharing intentions: the origins of cultural cognition. *Behav. Brain Sci.* **28**, 675–691. (doi:10.1017/S0140525X05000129)
16. Nowak MA. 2006 Five rules for the evolution of cooperation. *Science* **314**, 1560–1563. (doi:10.1126/science.1133755)
17. Bowles S, Gintis H. 2011 *A cooperative species: human reciprocity and its evolution*. Princeton, NJ: Princeton University Press.
18. Rand DG, Nowak MA. 2013 Human cooperation. *Trends Cogn. Sci.* **17**, 413–425. (doi:10.1016/j.tics.2013.06.003)
19. Capraro V. 2013 A model of human cooperation in social dilemmas. *PLoS ONE* **8**, e72427. (doi:10.1371/journal.pone.0072427)
20. Stanley HE. 1971 *Introduction to phase transitions and critical phenomena*. Oxford, UK: Clarendon Press.
21. Binder K, Hermann DK. 1988 *Monte Carlo simulations in statistical physics*. Heidelberg, Germany: Springer.
22. Estrada E. 2012 *The structure of complex networks: theory and applications*. Oxford, UK: Oxford University Press.
23. Boccaletti S, Bianconi G, Criado R, del Genio C, Gómez-Gardenes J, Romance M, Sendiña-Nadal I, Wang Z, Zanin M. 2014 The structure and dynamics of multilayer networks. *Phys. Rep.* **544**, 1–122. (doi:10.1016/j.physrep.2014.07.001)
24. Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA. 2014 Multilayer networks. *J. Complex Netw.* **2**, 203–271. (doi:10.1093/comnet/cnu016)
25. Barabási A-L. 2015 *Network science*. Cambridge, UK: Cambridge University Press.
26. Castellano C, Fortunato S, Loreto V. 2009 Statistical physics of social dynamics. *Rev. Mod. Phys.* **81**, 591–646. (doi:10.1103/RevModPhys.81.591)
27. Szabó G, Fáth G. 2007 Evolutionary games on graphs. *Phys. Rep.* **446**, 97–216. (doi:10.1016/j.physrep.2007.04.004)
28. Perc M, Szolnoki A. 2010 Coevolutionary games—a mini review. *BioSystems* **99**, 109–125. (doi:10.1016/j.biosystems.2009.10.003)
29. Perc M, Gómez-Gardeñes J, Szolnoki A, Floría LM, Moreno Y. 2013 Evolutionary dynamics of group interactions on structured populations: a review. *J. R. Soc. Interface* **10**, 20120997. (doi:10.1098/rsif.2012.0997)
30. Wang Z, Wang L, Szolnoki A, Perc M. 2015 Evolutionary games on multilayer networks: a colloquium. *Eur. Phys. J. B* **88**, 124. (doi:10.1140/epjb/e2015-60270-7)
31. D'Orsogna MR, Perc M. 2015 Statistical physics of crime: a review. *Phys. Life Rev.* **12**, 1–21. (doi:10.1016/j.plrev.2014.11.001)
32. Giardini F, Vilone D. 2016 Evolution of gossip-based indirect reciprocity on a bipartite network. *Sci. Rep.* **6**, 37931. (doi:10.1038/srep37931)
33. Pastor-Satorras R, Castellano C, Van Mieghem P, Vespignani A. 2015 Epidemic processes in complex networks. *Rev. Mod. Phys.* **87**, 925–979. (doi:10.1103/RevModPhys.87.925)
34. Wang Z, Bauch CT, Bhattacharyya S, d'Onofrio A, Manfredi P, Perc M, Perra N, Salathé M, Zhao D. 2016 Statistical physics of vaccination. *Phys. Rep.* **664**, 1–113. (doi:10.1016/j.physrep.2016.10.006)
35. Perc M. 2016 Phase transitions in models of human cooperation. *Phys. Lett. A* **380**, 2803–2808. (doi:10.1016/j.physleta.2016.06.017)
36. Perc M, Jordan JJ, Rand DG, Wang Z, Boccaletti S, Szolnoki A. 2017 Statistical physics of human cooperation. *Phys. Rep.* **687**, 1–51. (doi:10.1016/j.physrep.2017.05.004)
37. Capraro V, Rand DG. 2018 Do the right thing: experimental evidence that preferences for moral behavior, rather than equity or efficiency per se, drive human prosociality. *Judgm. Decis. Mak.* **13**, 99–111. (doi:10.2139/ssrn.2965067)
38. Capraro V, Perc M. 2018 Grand challenges in social physics: in pursuit of moral behavior. *Front. Phys.* **6**, 107. (doi:10.3389/fphy.2018.00107)
39. Gravelle JG. 2015 *Tax havens: international tax avoidance and evasion*. Washington, DC: Congressional Research Service.
40. FBI, Insurance fraud. https://www.fbi.gov/stats-services/publications/insurance-fraud/.
41. Pennycook G, Cannon TD, Rand DG. 2018 Prior exposure increases perceived accuracy of fake news. *J. Exp. Psychol.: Gen.* **147**, 1865–1880. (doi:10.1037/xge0000465)
42. Mazar N, Amir O, Ariely D. 2008 The dishonesty of honest people: a theory of self-concept maintenance. *J. Mark. Res.* **45**, 633–644. (doi:10.1509/jmkr.45.6.633)
43. Ariely D, Jones S. 2012 *The (honest) truth about dishonesty: how we lie to everyone—especially ourselves*, vol. 336. New York, NY: Harper-Collins.
44. Gino F, Ayal S, Ariely D. 2009 Contagion and differentiation in unethical behavior. *Psychol. Sci.* **20**, 393–398. (doi:10.1111/j.1467-9280.2009.02306.x)
45. Gino F, Schweitzer ME, Mead NL, Ariely D. 2011 Unable to resist temptation: how self-control depletion promotes unethical behavior. *Organ. Behav. Human Decis. Process.* **115**, 191–203. (doi:10.1016/j.obhdp.2011.03.001)
46. Shalvi S, Dana J, Handgraaf MJ, De Dreu CK. 2011 Justified ethicality: observing desired counterfactuals modifies ethical perceptions and behavior. *Organ. Behav. Human Decis. Process.* **115**, 181–190. (doi:10.1016/j.obhdp.2011.02.001)
47. Shalvi S, Eldar O, Bereby-Meyer Y. 2012 Honesty requires time (and lack of justifications). *Psychol. Sci.* **23**, 1264–0. (doi:10.1177/0956797612443835)
48. Shalvi S, Gino F, Barkan R, Ayal S. 2015 Self-serving justifications. *Curr. Dir. Psychol. Sci.* **24**, 125–130. (doi:10.1177/0963721414553264)
49. Verschuere B, Spruyt A, Meijer EH, Otgaar H. 2011 The ease of lying. *Conscious Cogn.* **20**, 908–911. (doi:10.1016/j.concog.2010.10.023)
50. Biziou-van Pol L, Haenen J, Novaro A, Occhipinti-Liberman A, Capraro V. 2015 Does telling white lies signal pro-social preferences? *Judgm. Decis. Mak* **10**, 538–548. (doi:10.2139/ssrn.2617668)
51. Capraro V. 2017 Does the truth come naturally? Time pressure increases honesty in one-shot deception games. *Econ. Lett.* **158**, 54–57. (doi:10.1016/j.econlet.2017.06.015)
52. Capraro V. 2018 Gender differences in lying in sender-receiver games: a meta-analysis. *Judgm. Decis. Mak.* **13**, 345–355. (doi:10.31234/osf.io/jaewt)
53. Erat S, Gneezy U. 2012 White lies. *Manage. Sci.* **58**, 723–733. (doi:10.1287/mnsc.1110.1449)
54. Gneezy U, Kajackaite A, Sobel J. 2018 Lying aversion and the size of the lie. *Am. Econ. Rev.* **108**, 419–453. (doi:10.1257/aer.20161553)
55. Capraro V, Schulz J, Rand DG. 2019 Time pressure and honesty in a deception game. *J. Behav. Exp. Econ.* **79**, 93–99. (doi:10.1016/j.socec.2019.01.007)
56. Fischbacher U, Föllmi-Heusi F. 2013 Lies in disguise—an experimental study on cheating. *J. Eur. Econ. Assoc.* **11**, 525–547. (doi:10.1111/jeea.12014)
57. Gneezy U. 2005 Deception: the role of consequences. *Am. Econ. Rev.* **95**, 384–394. (doi:10.1257/0002828053828662)
58. Smith JM. 1991 Honest signalling: the Philip Sidney game. *Anim. Behav.* **42**, 1034–1035. (doi:10.1016/S0003-3472(05)80161-7)
59. Sutter M. 2009 Deception through telling the truth?! Experimental evidence from individuals and teams. *Econ. J.* **119**, 47–60. (doi:10.1111/j.1468-0297.2008.02205.x)
60. Grafen A. 1990 Biological signals as handicaps. *J. Theor. Biol.* **144**, 517–546. (doi:10.1016/S0022-5193(05)80088-8)
61. Számadó S. 2011 The cost of honesty and the fallacy of the handicap principle. *Anim. Behav.* **81**, 3–10. (doi:10.1016/j.anbehav.2010.08.022)
62. Catteeuw D, Han TA, Manderick B. 2014 Evolution of honest communication through social punishment. In *Proc. of the 2014 Genetic and Evolutionary Computation Conference (GECCO 2014), Vancouver, 12–16 July 2014*, pp. 153–160. New York, NY: ACM.
63. Han TA, Pereira LM, Santos FC, Lenaerts T. 2013 Good agreements make good friends. *Sci. Rep.* **3**, 2695. (doi:10.1038/srep02695)
64. Han TA, Pereira LM, Lenaerts T. 2015 Avoiding or restricting defectors in public goods games? *J. R. Soc. Interface* **12**, 20141203. (doi:10.1098/rsif.2014.1203)
65. Han TA, Pereira LM, Lenaerts T. 2017 Evolution of commitment and level of participation in public goods games. *Auton. Agents Multi-Agent Syst.* **31**, 561–583. (doi:10.1007/s10458-016-9338-4)
66. Curry OS, Mullins DA, Whitehouse H. 2019 Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Curr. Anthropol.* **60**, 47–69.
67. Szolnoki A, Perc M, Szabó G. 2012 Defense mechanisms of empathetic players in the spatial ultimatum game. *Phys. Rev. Lett.* **109**, 078701. (doi:10.1103/PhysRevLett.109.078701)
68. Page KM, Nowak MA, Sigmund K. 2000 The spatial ultimatum game. *Proc. R. Soc. London Ser. B* **267**, 2177–2182. (doi:10.1098/rspb.2000.1266)

69. Kuperman M, Risau-Gusman S. 2008 The effect of the topology on the spatial ultimatum game. *Eur. Phys. J. B* **62**, 233–238. (doi:10.1140/epjb/e2008-00133-x)
70. Eguíluz VM. 2009 Critical behavior in an evolutionary ultimatum game with social structure. *Adv. Complex Syst.* **12**, 221–232. (doi:10.1142/S0219525909002179)
71. da Silva R, Kellermann GA, Lamb LC. 2009 Statistical fluctuations in population bargaining in the ultimatum game: static and evolutionary aspects. *J. Theor. Biol.* **258**, 208–218. (doi:10.1016/j.jtbi.2009.01.017)
72. Deng L, Tang W, Zhang J. 2011 The coevolutionary ultimatum game on different network topologies. *Physica A* **390**, 4227–4235. (doi:10.1016/j.physa.2011.06.076)
73. Gao J, Li Z, Wu T, Wang L. 2011 The coevolutionary ultimatum game. *Europhys. Lett.* **93**, 48003. (doi:10.1209/0295-5075/93/48003)
74. Szolnoki A, Perc M, Szabó G. 2012 Accuracy in strategy imitations promotes the evolution of fairness in the spatial ultimatum game. *Europhys. Lett.* **100**, 28005. (doi:10.1209/0295-5075/100/28005)
75. Deng L, Wang C, Tang W, Zhou G, Cai J. 2012 A network growth model based on the evolutionary ultimatum game. *J. Stat. Mech: Theory Exp.* **2012**, P11013. (doi:10.1088/1742-5468/2012/11/P11013)
76. Iranzo J, Floria LM, Moreno Y, Sanchez A. 2012 Empathy emerges spontaneously in the ultimatum game: small groups and networks. *PLoS ONE* **7**, e43781. (doi:10.1371/journal.pone.0043781)
77. Miyaji K, Wang Z, Tanimoto J, Hagishima A, Kokubo S. 2013 The evolution of fairness in the coevolutionary ultimatum games. *Chaos Solitons Fractals* **56**, 13–18. (doi:10.1016/j.chaos.2013.05.007)
78. Krupka EL, Weber RA. 2013 Identifying social norms using coordination games: why does dictator game sharing vary? *J. Eur. Econ. Assoc.* **11**, 495–524. (doi:10.1111/jeea.12006)
79. Capraro V, Vanzo A. 2019 The power of moral words: loaded language generates framing effects in the extreme dictator game. *Judgm. Decis. Mak.* **14**, 309–317.