*Current Zoology* 61 (2): 382–396, 2015

# Self-deception can evolve under appropriate costs

Juan Camilo RAMÍREZ*, James A. R. MARSHALL

Department of Computer Science, University of Sheffield, Sheffield S1 4DP, United Kingdom

**Abstract** Apparent biases in decision making by animals, including humans, seem to present an evolutionary puzzle, since one would expect decisions based on biased (unrealistic) information to be suboptimal. Although cognitive biases are hard to diagnose in real animals (Marshall et al., 2013b), we investigate Trivers' proposal that individuals should self-deceive first in order to better deceive others (Trivers, 2011). Although this proposal has been scrutinized extensively (Bandura et al., 2011) it has not been formally modelled. We present the first model designed to investigate Trivers' proposal. We introduce an extension to a recent model of the evolution of self-deception (Johnson and Fowler, 2011). In the extended model individuals make decisions by taking directly into account the benefits and costs of each outcome and by choosing the course of action that can be estimated as the best with the information available. It is shown that in certain circumstances self-deceiving decision-makers are the most evolutionarily successful, even when there is no deception between these. In a further extension of this model individuals additionally exhibit deception biases and Trivers' premise (that effective deception is less physiologically costly with the aid of self-deception) is incorporated. It is shown that under Trivers' hypothesis natural selection favors individuals that self-deceive as they deceive others [*Current Zoology* 61 (2): 382–396, 2015].

**Keywords** Self-deception, Deception, Overconfidence, Cognitive biases, Optimal decision-making, Optimal behavior

Deception in animals (no conscious intention being implied) refers to the signaling of false information from one individual to another and is normally beneficial to the signaller and detrimental to the receiver (Semple and McComb, 1996). For this reason some animals are observed to evolve strategies to deceive others, although natural selection is expected to also favor individuals who are able to 'see' through the deception. In addition to this, signaling may also be costly in order to be considered reliable, for example in mating situations (Zahavi, 1975). Arms races may occur between deceivers (no conscious intention being implied) and deception-uncovering species, with each group under selection to outsmart the other. Unlike deception, self-deception can be sensibly hypothesized not to be evolutionarily stable by itself, because animals who make decisions on false information seem more likely to make bad choices that could lead to negative consequences, such as injury and death. Especially in situations where conflict is likely it is sensible to expect that self-deceiving individuals tend to make suboptimal decisions, for instance risking injury through fighting a stronger opponent, and that in the long term they end up being less evolutionarily successful than others who use truthful information (Marshall et al., 2013b). Despite this, self-deception biases are claimed to occur frequently. For instance, animals and humans sometimes behave as if their subjective confidence in their skills in a given moment is above the objective measure of such attributes (Svenson, 1981; McCormick et al., 1986; Pallier et al., 2002; Alicke and Govorun, 2005). Surveys have shown most drivers rate their own skills as above average (Svenson, 1981; McCormick et al., 1986) and most students regard themselves as above-average leaders (Alicke and Govorun, 2005). It has also been documented that people who are unskilled for a task often fail to recognize their lack of competence, a phenomenon known as the Dunning-Kruger effect (Kruger and Dunning, 1999). Psychological tests have also shown that people tend to overestimate the probability of positive events (e.g., career success) and to underestimate the probability of negative events (e.g., onset of a serious illness) (Sharot, 2011b). Additional studies have shown that these optimistic expectations are not necessarily deterred by knowledge of past, realistic information. For instance, newly married couples tend to overestimate the likelihood of having long marriages despite reported divorce rates of around 50% (Sharot, 2011a). Some stroke patients, who are aware of their condition, have been observed to deceive themselves into thinking that their paralysis is due to factors other than their illness (Ramachandran, 1996). Similar studies have found

Received Oct. 21, 2014; accepted Jan. 23, 2015.

* Corresponding author. E-mail: acq11jcr@sheffield.ac.uk

that surveyed students also rate others as above average (Klar and Giladi, 1997). These results appear to show that individuals are generally unable to estimate correctly the average capability in a group (Brooks and Swann, 2011; Chambers and Windschitl, 2004) and that they have a tendency to overestimate the skills of others.

Self-deception has been defined as a misrepresentation of reality (Trivers, 2000). At its simplest, this would correspond to using a biased estimate of the probability of an event in decision-making. It generally comes in the form of a bias, which is a tendency to act prejudicedly or behave in a way that apparently does not conform to rationality. Biases can be of one of two types: cognitive biases, which are perceptual biases in the subjective experience of an individual, and behavioral biases, which are manifest in behaviors that depart from the optimal fitness-maximizing strategy (Marshall et al., 2013b). Cognitive biases are generally hard to diagnose, and attempts to classify or explain them are often controversial (Dougherty et al., 1999; Marshall et al., 2013b). The apparent overconfidence exhibited by students and drivers in the surveys mentioned above are usually diagnosed by psychologists as an example of a cognitive bias (Svenson, 1981; McCormick et al., 1986; Pallier et al., 2002; Alicke and Govorun, 2005). Even though a bias may result in seemingly unreasonable behavior it could evolve if, for instance, the bias is part or the by-product of a larger behavioral trait that overall proves to be individually advantageous. Thus cognitive biases may evolve given appropriate decision machinery, whereas we expect behavioral biases not to.

It has been proposed by Trivers (Trivers, 2011; von Hippel and Trivers, 2011b) that the most evolutionarily successful deceivers in nature are those that self-deceive first. That is to say, *unconscious deceivers* (i.e., those who unwittingly 'lie' to themselves just as they lie to others) are favored by natural selection over *conscious deceivers* (i.e., those who intentionally attempt to be deceitful while acting on truthful information). Trivers hypothesizes that this is because conscious deceivers have to pay a considerable cognitive cost in order to avoid exhibiting involuntary responses (e.g., blushing, nervousness, blinking, voice tone, etc.) that would allow others to see through the deception. On the other hand unconscious deceivers do not have to pay the same cognitive cost and tend to be better cheaters because they do not exhibit the same involuntary responses, since they believe the lie. According to Trivers' theory a tendency towards self-deception evolves as a supportive by-product of the ability to deceive others, and the cost of lying to oneself is outweighed by the benefit brought by the ability to lie convincingly to adversaries. The theory proposed by Trivers has received extensive discussion from different commentators (Bandura et al., 2011) and Trivers has addressed these criticisms (von Hippel and Trivers, 2011a). One point that has been raised is that in situations of conflict, a deceiver may succeed in discouraging competitors from fighting (e.g., by feigning a strength higher than the actual one) but it is likely that at some point the deception may be uncovered by others and that then the deceiver will face serious consequences, such as injury or death, as pointed out earlier (Marshall et al., 2013b; Frey and Voland, 2011; Funder, 2011). In such case the eventual cost of being discovered may be higher than the advantage posed by deceiving others, and self-deception should not evolve. This point has not been addressed by Trivers (Marshall et al., 2013b).

In Section 1.1 we extend a model of the evolution of overconfidence proposed by Johnson and Fowler (2011) in order to investigate the evolution of self-deception given statistically-optimal behavioral machinery. With this it is shown analytically and computationally that under certain circumstances overconfidence evolves even when decision-makers use a theoretically optimal decision rule as suggested by Marshall et al. (2013b). In this case overconfidence or underconfidence are cognitive biases assuming a particular decision machinery, since they lead to optimal behavior, rather than a suboptimal behavioral bias (Marshall et al., 2013b). The new model is extended in Section 1.2 to incorporate deception biases in order to test Trivers' theory (Trivers, 2011) by showing that deception is favored by natural selection when self-deception reduces cognitive or other costs. These self-deception biases are shown to be evolutionarily stable in a situation of conflict, one scenario not addressed by Trivers when replying to their critics (Marshall et al., 2013b; von Hippel and Trivers, 2011a). In the model presented in Section 1.1 individuals do not attempt to deceive others because the purpose is to compare the self-deception biases with (model in Section 1.2) and without (model in Section 1.1) deception between individuals. Analysis and results of the models introduced in Section 1.1 and Section 1.2 are presented in Section 2.1 and Section 2.2, respectively.

# 1 Materials and Methods

## 1.1 A simplified owner-challenger model with internal biases

In this section we present an extension to Johnson

and Fowler (J&F)'s model (Johnson and Fowler, 2011). In the model presented in this section, individuals self-deceive but do not deceive others. This extension, called the *simplified owner-challenger model*, is further extended in Section 1.2 to allow individuals to both self-deceive and deceive. The purpose of having the two models is to compare the level of self-deception that evolves in the absence of selective pressure to deceive others (in the simplified model presented in this section), and compare it with the level of self-deception that evolves when this selective pressure is present (in the generalized model presented in Section 1.2).

The definition of the simplified owner-challenger model is similar to that of J&F's (Johnson and Fowler, 2011) and can be formulated as follows. Each individual has a fighting capability, denoted by $\theta$, which is normally distributed with mean zero and standard deviation $\sqrt{1/2}$. Given any two individuals, *i* and *j*, with capabilities $\theta_i$ and $\theta_j$ respectively, the former would defeat the latter if $\theta_i > \theta_j$ should a conflict between them occur. The capability advantage *i* has over *j* is defined as $A = \theta_i - \theta_j$, which, since normal variances are additive, is a standard normal random variable (i.e., $A \sim N(0,1)$). The marginal probability that *i* defeats *j* is thus given by $p_W = P(A > 0) = 1/2$.

Every individual *i* also has an *internal bias* (i.e., a self-deception bias), denoted by $k_i$, that distorts its perception of its own capability in such a manner that *i* always acts as if its capability is $\theta_i + k_i$. In addition to this, *i*'s *perception* of *j*'s capability, denoted by $\hat{\theta}_j$, is normally-distributed with mean $\theta_j$ and standard deviation $\sigma_\varepsilon$ (i.e., $\hat{\theta}_j \sim N(\theta_j, \sigma_\varepsilon)$). The perception error size, $\sigma_\varepsilon$, is a non-negative parameter of the model. In this manner the model simulates perception errors as they occur in nature, which are due in part to environmental factors beyond the control of each individual, as well as being due to sensory noise. In this manner the advantage *i* perceives it has over an opponent *j* is given by $\hat{A} = \theta_i + k_i - \hat{\theta}_j$.

A conflict between two individuals over a resource occurs in an owner-challenger encounter as shown in Fig. 1, where *r* is the value of the contested resource and *c* is a cost both individuals pay if they fight. The encounter involves the owner of the resource, who arrives at it first, and a challenger, who arrives subsequently and decides whether to claim the resource. If the challenger claims the owner decides whether to fight for the resource or abandon it to the challenger. Both *r* and *c* are constant and positive, and each individual decides in sequence whether to fight or not.

In a more realistic scenario, *r*, *c*, and $\sigma_\varepsilon$ would likely vary from individual to individual and from encounter to encounter. For instance, an individual who has collected many resources will value a newly encountered resource less than an individual who has collected none. Similarly, the cost of a fight will probably be higher for an individual who has been injured badly from pre-

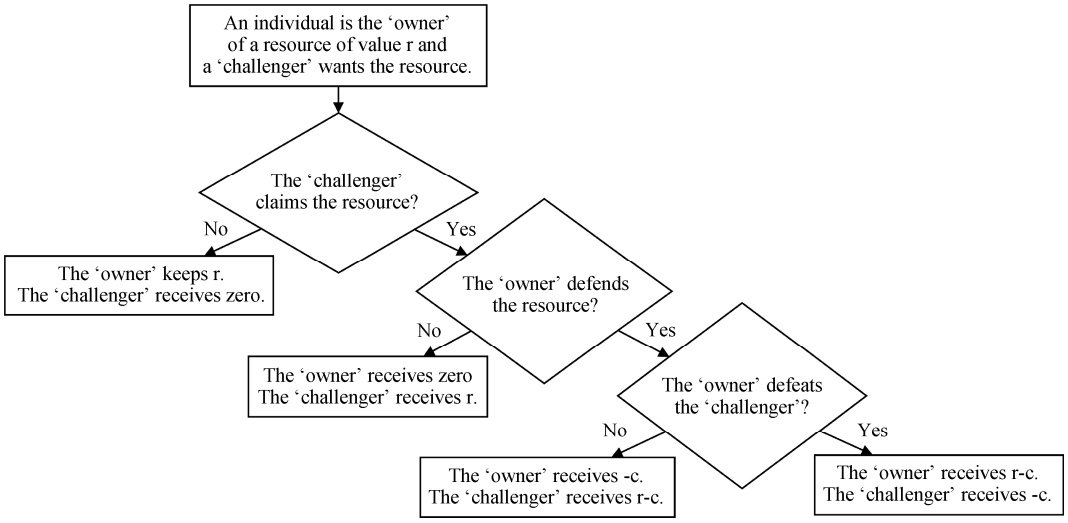


**Fig. 1 An owner-challenger encounter occurs when one individual is the owner of a resource and then a challenger arrives with the intention of claiming the resource**

Both parties decide asynchronously whether to fight over the resource or surrender it to the other individual. If both fight then both pay a cost $-c$ but the winner additionally receives the resource value r. The strongest individual wins the fight. If both have the same capability then the winner is decided randomly with each individual having equal probability of winning.

viously lost fights than for an individual who has lost none. In addition to this, in a natural scenario perceptual capabilities as well as conditions of the local environment (e.g., low visibility that affects the ability to visualize the opponent) would be likely to vary from encounter to encounter, resulting in different perception errors between individuals. However, $r$, $c$, and $\sigma_\varepsilon$ have been kept constant among all individuals for simplicity because in this manner the formal analysis of the model (Section 2.1) and a further extension (Section 1.2) of it are much more manageable and by doing so we identify minimal conditions that are sufficient for the evolution of the biases that we are interested in.

There are two differences between the owner-challenger model and the one proposed by J&F (Johnson and Fowler, 2011). The first is that in J&F's model the two competing individuals decide synchronously whether to fight or not whereas in the owner-challenger model the two decisions are made asynchronously and in sequence. The second is that in J&F's model individuals make their decisions based only on their perceived advantage over their respective opponents while ignoring the benefits ($r$) and costs ($c$) of each decision, whereas in the owner-challenger model these variables are taken into account by each individual. In this manner, the owner-challenger model addresses two main criticisms of J&F's (Marshall et al., 2013a,b; Johnson and Fowler, 2013). The first one is that since in J&F's model contests over resources are synchronous they can lead to valuable resources remaining unclaimed if no individual chooses to contest them, while the second is that in J&F's model individuals use an arbitrary and unrealistic decision rule in deciding whether to contest (Marshall et al., 2013a,b). The first criticism is addressed by allowing individuals to use whether they arrived at a resource first or second to determine their strategy, thereby creating a uncorrelated asymmetry (Maynard Smith, 1982) and allowing low value resources to be claimed by one individual. The second criticism is addressed by enabling individuals to use the estimated payoffs associated with different outcomes, and an estimate of the probability of winning, to determine whether to contest a resource.

A realistic scenario in the owner-challenger model is that where both individuals use all the relevant information when making their respective decisions. However the mathematical analysis of the model becomes difficult when both decision-makers behave in this manner. For this reason we analyze first a simplified version of the model where the decision of the challenger $j$ is always to claim the resource and fight whereas the owner $i$ makes its decision (after having been challenged by $j$) by using the following reasoning; first $i$ estimates its own probability of winning as $\hat{p}_W = P(\hat{\theta}_j < \theta_i + k_i)$. Then $i$ estimates its expected payoff from the hypothetical fight as $\hat{F} = \hat{p}_W (r - c) + (1 - \hat{p}_W)(-c) = \hat{p}_W \; r - c$. This individual then decides to defend the resource if and only if this estimated payoff is higher than zero. This in turn occurs if and only if $\hat{p}_W > c / r$ (Marshall et al., 2013b). This decision rule is rational from the perspective of an owner because it uses all the relevant information available to estimate the expected payoff from a fight and the final decision is made if and only if the evidence suggests that this estimate is positive. In the long run this rule should yield a positive payoff to an owner on average after repeated encounters with random challengers.

The estimate $\hat{F}$ does not include the weighted payoff received by an individual when the opponent withdraws from conflict, therefore every owner works under the assumption that the opponent is always intent to fight. This assumption is clearly correct from the perspective of the owner because its decision-making takes place only after having been challenged. However a rational challenger should not always claim, since this ignores the probability that the owner will defend the resource rather than abandon it uncontested. Therefore a challenger that always claims should be expected to perform worse in the long term (i.e., after repeated encounters against random owners) than an owner. The simplified model is proposed in this manner, with always-aggressive challengers, in order to determine analytically what values of $r$, $c$, and $\sigma_\varepsilon$ make internal biases necessary for owners to receive the highest long-term payoffs, even when these individuals use the rational decision rule stated above and when no deception between individuals is present. The analysis of this model and the results are presented in Section 2.1.

### 1.2 The generalized owner-challenger model with role-dependent internal and external biases

In this section we introduce a generalized version of the simplified model presented in Section 1.1 in order to simulate the scenario where every decision-maker additionally has an *external bias* (i.e., a deception bias) that alters the capability this individual signals to any opponent. The larger the external bias the greater the baseline capability signalled to competitors. Given any two individuals, $x$ and $y$, what $x$ perceives is $y$'s projected capa-

bility, distorted first by *y*'s external bias and then by *x*'s own perception error. The actual attribute remains unchanged but *y* may be able to deceive *x* into thinking that *y*'s capability is greater (or lower) than it actually is, thus making *x* less (or more) willing to fight. The model aims to test the theory proposed by Trivers (2011). By incorporating the premises of the theory (namely costs paid for conscious deception of others) computational simulations are run to determine in what circumstances, if any, self-deception evolves in order to facilitate the deception of opponents. An individual with a non-zero external bias exerts a form of deception, or dishonest signaling. The use of a positive external bias is similar to *deimatic behavior*, in which an animal, feeling in danger, makes a physical display, possibly involving changes in shape, position, and/or color, in order to appear threatening (probably more than the animal actually is) and to dissuade an opponent from attacking. Examples of deimatic individuals include some species of frog, who, in the presence of a threat, inflate themselves with air and raise their hind legs in order to appear larger (Martins, 1989). The dishonest signal sent by an individual with a positive external bias could also be compared to *Batesian mimicry*, where a harmless individual imitates the signals of a harmful one, in order to discourage attacks from predators. Examples of Batesian species include *Lampropeltis elapsoides*, a nonvenomous snake who exhibits the color pattern of the venomous *Micrurus fulvius* (Kikuchi and Pfennig, 2010).

In the generalized model each decision-maker has two types of bias. An internal bias, denoted by *k*, that influences the perception the individual has of itself (as in the simplified model of the previous section), and an external bias, denoted by *s*, that distorts the capability it displays to opponents. Both biases comprise together a *deception pair* denoted by [*k*, *s*]. Any individual *x* with internal bias $k_x$ and external bias $s_x$ believes that its own capability is $\theta_x + k_x$ and attempts to deceive any potential opponent *y* into believing that *x*'s capability is $\theta_x + s_x$. Then what *y* perceives is a normal deviate of the projected capability with standard deviation $\sigma_\varepsilon$, the perception error size.

Each decision-maker in the model holds two deception pairs, [$k_o$, $s_o$] and [$k_c$, $s_c$]. The first one of these is expressed when the individual is playing the role of an owner and the second one when the individual is a challenger. The first pair can be referred to as the individual's *owner biases* and the latter as the *challenger biases*. Alternatively the first pair can be referred to as the *owner strategy* of the individual while the second pair is the *challenger strategy*. A strategy is *symmetrically biased* if its internal and external biases are equal otherwise it is *asymmetrically biased*. Asymmetrically-biased individuals represent organisms in nature that exercise 'conscious' deception because they attempt to project an image of themselves that differs from their true self-perception. On the other hand, symmetrically-biased individuals represent organisms that do not deceive or deceive 'unconsciously', because if they spread false information it is only because they 'believe' it as well.

In each encounter every individual expresses only the pair of biases that match the role (owner or challenger) the individual is playing at that moment. Given an owner *x* and a challenger *y*, *x* estimates its probability of winning as $\hat{p}_{W,x} = P\left(\hat{\theta}_y < \theta_x + k_{o,x}\right)$, where $\hat{\theta}_y \sim N\left(\theta_y + s_{c,y}, \sigma_\varepsilon\right)$, given that in the encounter this individual estimates its capability as $\theta_x + k_{o,x}$ (with *x*'s internal owner bias) and that of *y* as a normal deviate of $\theta_y + s_{c,y}$ (with *y*'s external challenger bias) with perception error size $\sigma_\varepsilon$. On the other hand, *y* estimates its probability of winning as $\hat{p}_{W,y} = P\left(\hat{\theta}_x < \theta_y + k_{c,y}\right)$, where $\hat{\theta}_x \sim N\left(\theta_x + s_{o,x}, \sigma_\varepsilon\right)$, given that in the encounter this individual estimates its capability as $\theta_y + k_{c,y}$ (with *y*'s internal challenger bias) and that of *x* as a normal deviate of $\theta_x + s_{o,x}$ (with *x*'s external owner bias) with perception error size $\sigma_\varepsilon$. The owner *x* estimates its expected payoff as $\hat{F}_x = \hat{p}_{W,x}(r-c) + \left(1-\hat{p}_{W,x}\right)(-c) = \hat{p}_{W,x} r - c$ and fights if and only if this estimate is positive. Similarly, the challenger *y* estimates its expected payoff as $\hat{F}_y = \hat{p}_{W,y} r - c$ and decides to fight if and only if $\hat{F}_y > 0$. As explained in Section 1.1, this decision rule (i.e., fighting if and only if the estimated payoff is positive) is rational from the perspective of an owner but not necessarily so from the perspective of a challenger. This is because for the challenger a rationally estimated payoff would necessarily include an estimate of the probability of the owner contesting the resource. However the model becomes difficult to analyze if the challenger is set to estimate this probability. For this reason we consider the simplified scenario where both challengers and owners use the same rule because the former can be realistically assumed to be conservative when forced to work with imperfect information. In this manner the challenger only challenges when it estimates

that it can win the resource even if the owner fights back. Despite not being rigorously rational this challenger behavior is sensible and realistic.

Trivers' premise that deception is more costly in the absence of self-deception (Trivers, 2011) is incorporated into the model by having each individual pay a *conscious deception cost* that penalizes asymmetrical strategies, regardless of whether a fight actually takes place or not. The cost paid by an individual is proportional to the discrepancy between the internal and the external biases in the strategy exercised by this individual in the encounter, with a proportionality constant $\lambda \in [0,1]$. That is to say, the cost paid by the owner $x$ increases with the difference between this individual's owner biases and is given by $\lambda|k_{o,x} - s_{o,x}|$, whereas the cost paid by the challenger $y$ increases with the difference between this individual's challenger biases and is given by $\lambda|k_{c,y} - s_{c,y}|$.

Dishonest signaling may serve as a way to avoid the cost derived from a physical conflict by discouraging an opponent from fighting (an individual with a high external bias may dissuade an opponent from fighting); however, in nature such signaling is also costly, even though the cost paid in exchange for the ability to cheat opponents (e.g., through having to invest in ornamentation) may be less than the one paid for taking part in a fight (e.g., through sustaining an injury) (Backwell et al., 2000; Zahavi, 1975). This premise is incorporated into the model by having each individual pay a *dishonest signaling cost* proportional to the square of the external bias in the strategy played by the individual in an encounter against an opponent. The proportionality constant is denoted $\omega \in [0,1]$ and the cost paid by an owner $x$ is thus given by $\omega s_{o,x}^2$ whereas the cost paid a challenger $y$ is given by $\omega s_{c,y}^2$. The conscious deception cost and the dishonest signaling cost paid by an individual are subtracted from the payoff received by this decision-maker from the encounter. For instance, if an owner $x$ wins a fight against a challenger $y$ then $x$'s final payoff is $F_x = r - c - \lambda\left|k_{o,x} - s_{o,x}\right| - \omega s_{o,x}^2$, whereas $y$'s is $F_y = -c - \lambda\left|k_{c,y} - s_{c,y}\right| - \omega s_{c,y}^2$. Clearly positive factors $\lambda$ and $\omega$ together impose a selective pressure on decision-makers driving them towards becoming less deceptive and more symmetrically biased. We hypothesize that without the former parameter individuals should evolve to be asymmetrically biased whereas without the latter individuals should evolve to be maximally deceptive. Given any two individuals, $x$ and $y$, $y$'s internal bias can evolve so that $y$ disregards the uninformative signal originated from $x$'s external bias. In the absence of the dishonest signaling cost this would escalate, therefore this cost prevents signallers' external biases and receivers' internal biases from increasing indefinitely in an evolutionary arms race.

A set of evolutionary simulations were run with the role-dependent owner-challenger model under Triver's premise (i.e., with large enough values of $\lambda$ and $\omega$) as follows. Firstly a population of decision-makers is initialized randomly with standard normal biases, then each generation every individual $x$ is paired at random with exactly one adversary $y$ in the population. The fitness of $x$ is calculated as the average of its payoff when playing owner and challenger against $y$, and it increases with the resources (each one of these with value $r$) $x$ manages to protect (as an owner) and/or usurp from $y$ (as a challenger) and decreases with the number of fights $x$ involves itself in (because each fight comes with a cost $c$). The fitness of $x$ depends on the decisions this individual makes and how advantageously it influences the decisions of $y$, who is also trying to maximize its own gain. Fitness proportional selection (Baker, 1987) is used to determine which individuals reproduce, with normally-distributed mutations. Evolution runs until no considerable changes are observed and the population is assumed to be in equilibrium. Full details of the model are presented in the Supplementary Information. Results from the model are presented in Section 2.2.

# 2 Results

### 2.1 Results with the simplified owner-challenger model with internal biases

The expected payoff $F(k)$ of an owner with internal bias $k$ depends on the value of the resources contested (i.e., $r$), the cost of each fight (i.e., $c$), and the error made when estimating the capability of an opponent (i.e., $\sigma_\varepsilon$). We approximate this with a numerical method described in Appendix A. The expected payoffs for different values of $r/c$ and $k$ when $\sigma_\varepsilon = 1$ are plotted in Figure 2A. The plot shows that the highest payoff is obtained by owners with biases near zero when $r/c = 2$. But as this ratio increases it is owners with negative biases who receive the highest expected payoffs. Therefore owners that underestimate their own strength are the ones that in the long term perform the best against always-aggressive challengers when $\sigma_\varepsilon = 1$ and the value of the contested resource ($r$) outweighs the cost of a

confrontation ($c$). In Appendix B it is formally shown that when $r/c \in (0,1]$ owners never retaliate against always-aggressive challengers and end up receiving the same payoff (zero), regardless of owners' bias; this is because the value of the resource is offset by the cost of the inevitable fight. Therefore it can be concluded that as long as $r/c \leq 1$ no owner should perform better than the other and no bias can be considered optimal. On the other hand, the biases that maximize $F(k)$ when $r/c \in (1, +\infty)$ were found numerically and plotted in Figure 2B. In Appendix B it is shown that any owner $i$'s fighting probability is given by $p_F = P(A > z - k_i)$, where $z$ is an advantage threshold for conflict that depends on $r/c$ and $\sigma_\varepsilon$. It is also shown that if $i$ is an optimal-decision maker then $p_F = 1/2$. Therefore after a large enough number of repeated encounters with random challengers an optimal owner should fight back in half of these encounters, because $P(A>0) = 1/2$. This means that given any $r/c$ and $\sigma_\varepsilon$, only owners with biases equal to $z$ may be optimal decision-makers because only these have fighting probabilities equal to 1/2. This is confirmed by the numerical results displayed in Figure 2B. If $\sigma_\varepsilon = 0$ then $z = 0$ (details in Appendix B) and the highest payoffs are received by owners with approximately zero bias. Unbiased individuals also get the best payoffs when $\sigma_\varepsilon > 0$ and $r/c = 2$ because then $z = 0$. If $\sigma_\varepsilon > 0$ then $z$ decreases below zero as $r/c$ increases above 2 and therefore negative owner biases yield the highest payoffs. Additionally if $\sigma_\varepsilon > 0$ then $z$ increases above zero as $r/c$ decreases below 2 and owners achieve the maximum gain with positive biases. It can be concluded that owners require biases to optimize their payoffs if and only if the perception error is present (i.e., if $\sigma_\varepsilon > 0$). That is to say, internal biases serve as a means to compensate for information noise, given the assumed decision rule.

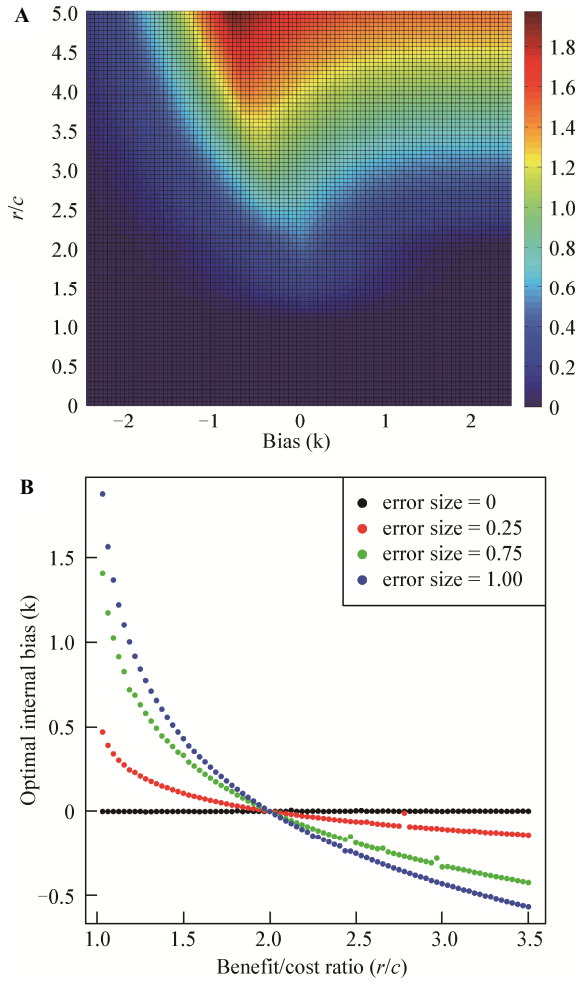


**Fig. 2 A. Owner's expected payoff ($F$) in the simplified model against random always-aggressive challengers as a function of confidence biases ($k$) and benefit cost ratios ($r/c$) when $\sigma_\varepsilon = 1$. B. The self-deception biases ($k$) that yield the highest payoffs to an owner in the long term when facing random always-aggressive challengers along different benefit/cost ratios ($r/c$) and perception error sizes ($\sigma_\varepsilon$).**

### 2.2 Results with the generalized owner-challenger model with role-dependent biases

We measured the difference between deception and self-deception biases when populations were in evolutionary equilibrium in different simulations with different values of $r/c$ and $\sigma_\varepsilon$. If these biases tend to evolve to have the same value when it is costly to have them differ, then this would support the theory proposed by Trivers (2011). Fig. 3 shows the average owner bias differences ($|k_o - s_o|$) and the average challenger bias differences ($|k_c - s_c|$) when the population is in evolutionary equilibrium in simulations run with parameters r/c $\in [1, 5]$, $\sigma_\varepsilon \in [0,4]$, $\lambda \in \{0,0.5,1\}$ and $\omega = 1$. That is to say, Figure 3 shows the level of symmetry of owner and challenger strategies in equilibrium with ($\lambda = 0.5$ and $\lambda = 1$) and without ($\lambda = 0$) Trivers' premise that deception is more costly to the deceiver when it is unconscious. The plots show that when $\lambda > 0$ the symmetry in owner and challenger strategies generally increases as $r/c$ and $\sigma_\varepsilon$ increase together. With high enough values of these two parameters the internal and external biases evolve to be approximately equal, which is consistent with Trivers' theory because as the parameters increase, natural selection increasingly favors individuals that self-deceive just as much as they deceive others because they avoid the cognitive cost of conscious deception, and the effects of acting based on false information are more than offset by the effects of manipulating others' perceptions. Figure 3 also shows that asymmetry can be

stable as the perception error decreases and $r/c$ increases. That is to say, as higher costs are paid for the ability of being consciously deceitful, it pays off more to be an unconscious deceiver, unless the perception errors are low (allowing the owner to make decisions on more certain information) and the value of the contested resource greatly outweighs the costs of a fight.

Fig. 3A shows that if $\lambda = 0$ then owner bias differences are generally lowest when $r/c \in [2,2.5]$ and they increase as $r/c$ increases and decreases away from this interval. That is to say, if $r/c \in [2,2.5]$ then evolutionary equilibrium generally occurs when the population exercise owner strategies that are symmetrically biased (i.e., when owners are unconscious deceivers), otherwise equilibrium generally occurs when the population exercise owner strategies that are symmetrically unbiased (i.e., when owners are conscious deceivers). The figure also shows that owner bias differences increase and decrease with $\sigma_\varepsilon$. This means that as the information available becomes noisier then it pays off more to be an asymmetrically-biased owner (i.e., a consciously-deceiving owner). A similar pattern occurs in challenger strategies, as shown in Fig. 3B, although the bias differences observed in these strategies when the population is in evolutionary equilibrium are generally higher. That is to say, differences in challenger strategies in-

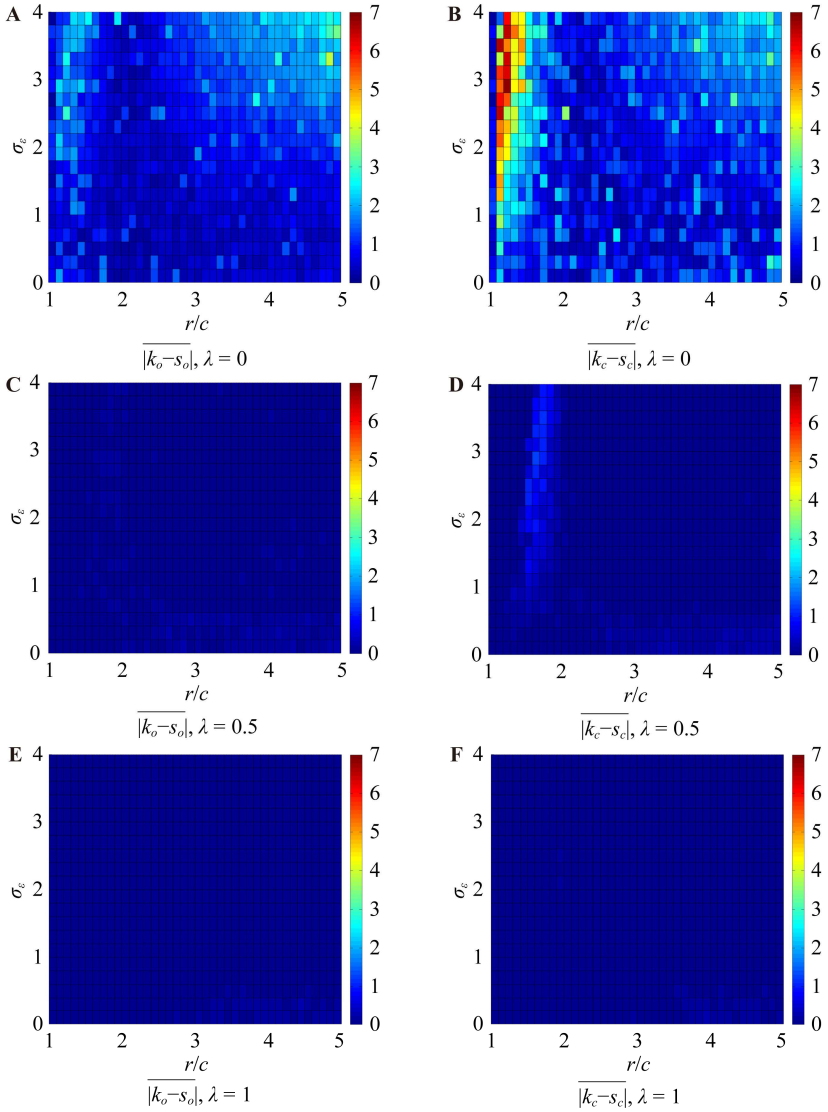


**Fig. 3 Differences between internal (self-deception) and external (deception) biases evolved with different benefit/cost ratios ($r/c$), perception errors ($\sigma_\varepsilon$), with fixed conscious deception costs ($\lambda$) and a fixed dishonest signaling cost ($\omega = 1$).**
With each combination of these parameters, evolutionary simulations are run with populations composed of individuals with role-dependent biases. When equilibrium is reached the average difference of owner biases ($|k_o - s_o|$) and the average difference of challenger biases ($|k_c - s_c|$) in the population are calculated and plotted. These plots show values of r/c and $\sigma_\varepsilon$ with which symmetrical (darker blue) and asymmetrical (darker red) strategies are stable. As conscious deception costs increase differences between internal and external biases tend to decrease towards zero.

crease with $\sigma_\varepsilon$ and as $r/c$ increases and decreases away from [2,3.5]. All the above is similar to what is observed in Figure 2B where the magnitude of optimal internal bias exhibit a similar relationship with $r/c$ and $\sigma_\varepsilon$, i.e., higher magnitudes as $r/c$ increases and decreases from 2 and as $\sigma_\varepsilon$ increases. It is reasonable to assume that this similarity is due to the same causes (i.e., information noise) although a formal demonstration of this (such as the one provided for the simplified model in Section 2.1) is difficult in the generalized model with role-dependent biases.

Additional evolutionary simulations were run with the same parameters with external biases absent. The purpose of this was to compare the evolved internal biases in owner and challenger strategies in the absence and presence of external biases. The difference in mag nitude of internal biases in equilibrium when these evolve in the presence and absence of external biases was measured by running separate evolutionary simulations with ($s_o$, $s_c \neq 0$) and then without ($s_o$, $s_c = 0$) deception biases. Individuals pay a conscious deception cost ($\lambda = 1$) and a dishonest signaling cost ($\omega = 1$) only in simulations where external biases are present. The average internal bias in owner strategies when the population is in evolutionary equilibrium in simulations with external biases is denoted by $\overline{k_o^d}|_{\lambda=1}$ and the average internal bias in challenger strategies is denoted by $\overline{k_c^d}|_{\lambda=1}$. The average internal bias in owner strategies when individuals are in evolutionary equilibrium in simulations without external biases is denoted by $\overline{k_o^{nd}}|_{\lambda=0}$ and the average internal bias in challenger strategies is denoted by $\overline{k_c^{nd}}|_{\lambda=0}$. The difference between evolved internal biases with and without external biases ($\overline{k_o^d}|_{\lambda=1} - \overline{k_o^{nd}}|_{\lambda=0}$ and $\overline{k_c^d}|_{\lambda=1} - \overline{k_c^{nd}}|_{\lambda=0}$) for each choice of $r/c$ and $\sigma_\varepsilon$ are plotted in Fig. 4. The two plots show that there are values of $r/c$ and $\sigma_\varepsilon$ for which these differences are generally positive and that these differences tend to increase as $r/c$ and $\sigma_\varepsilon$ increase. This implies that with high enough values of $r/c$ and $\sigma_\varepsilon$ the magnitude of the evolved internal biases increases in the presence of external biases, which means that the ability to deceive others requires an increase in self-deception in order to be evolutionarily stable. It can be hypothesized that the internal biases that evolve without external biases do so for a reason similar to the one explained in Section 2.1 for the simplified model with always-aggressive challengers (i.e., noise in the information available), although a formal demonstration of this is harder in the generalized model with role-dependent biases.

In evolutionary simulations with both internal and external biases where individuals are forced to pay a cost for the ability of conscious deception (i.e., for exercising asymmetrical strategies) there is a difference in the magnitude of the internal biases that evolve compared to the internal biases evolved in the absence of this cost, as shown in Fig. 5. The average internal bias

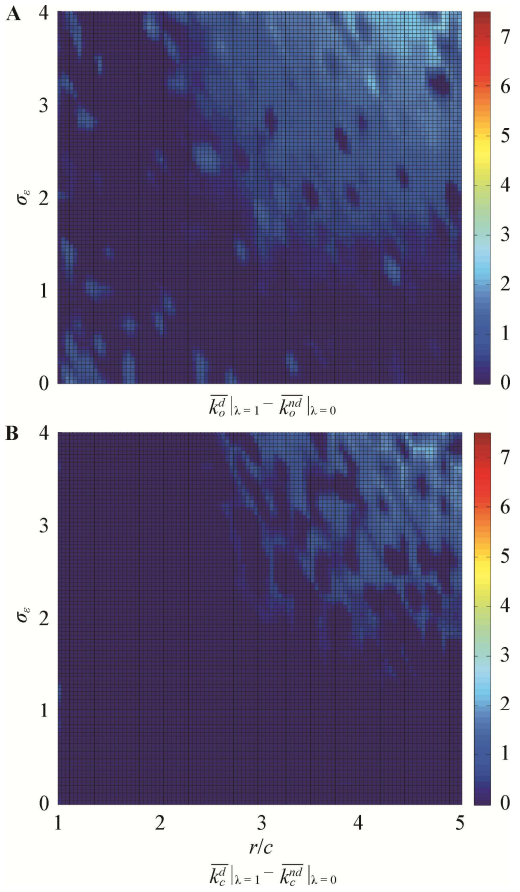


**Fig. 4 Difference between internal (self-deception) biases evolved in the presence and absence of external biases**

When external biases are present, individuals pay a conscious deception cost ($\lambda = 1$) and a dishonest signaling cost ($\omega = 1$). The notations $\overline{k_o^{nd}}|_{\lambda=0}$ and $\overline{k_c^{nd}}|_{\lambda=0}$ are used to refer to the average internal bias in owner and challenger strategies, respectively, in populations in evolutionary equilibrium when individuals evolve with no external biases (i.e., when $s_{o,x}$, $s_{c,x} = 0$ for every individual $x$). The notations $\overline{k_o^d}|_{\lambda=1}$ and $\overline{k_c^d}|_{\lambda=1}$ are used to refer to the average internal bias in owner and challenger strategies, respectively, in populations in evolutionary equilibrium when individuals evolve with external biases. Figure 4A shows the owner difference $\overline{k_o^d}|_{\lambda=1} - \overline{k_o^{nd}}|_{\lambda=0}$ for different benefit/cost ratios ($r/c$) and perception errors ($\sigma_\varepsilon$), whereas Figure 4B shows the challenger difference $\overline{k_c^d}|_{\lambda=1} - \overline{k_c^{nd}}|_{\lambda=0}$. The plots show that these differences tend to increase with $r/c$ and $\sigma_\varepsilon$. This implies that as these parameters increase, an increase in the ability to deceive others (from zero external bias to nonzero external bias) requires an increase in the magnitude of self-deception in order to be stable.

in the equilibrium population is measured when individuals evolve with ($\lambda = 1$) and without ($\lambda = 0$) paying a cost for conscious deception and then the differences between these averages is calculated for each choice of $r/c$ and $\sigma_\varepsilon$. The two plots in Fig. 5 show that these differences tend to increase with $r/c$ and $\sigma_\varepsilon$ and that they are generally above zero with high enough values of these two parameters. This implies that the magnitude of the self-deception that evolves under the influence of the conscious-deception cost tends to become larger than the self-deception that evolves without this cost as $r/c$ and $\sigma_\varepsilon$ increase. That is to say, with high enough values of $r/c$ and $\sigma_\varepsilon$, self-deception is effectively higher under Trivers' premise that conscious deception is costly (Trivers, 2011). In addition to this, self-deception evolved under Trivers' premise increases as the information decision-makers use becomes noisier (i.e., as $\sigma_\varepsilon$ increases) and as the value of the contested resource grows relative to the cost of a fight (i.e., as $r/c$ increases).

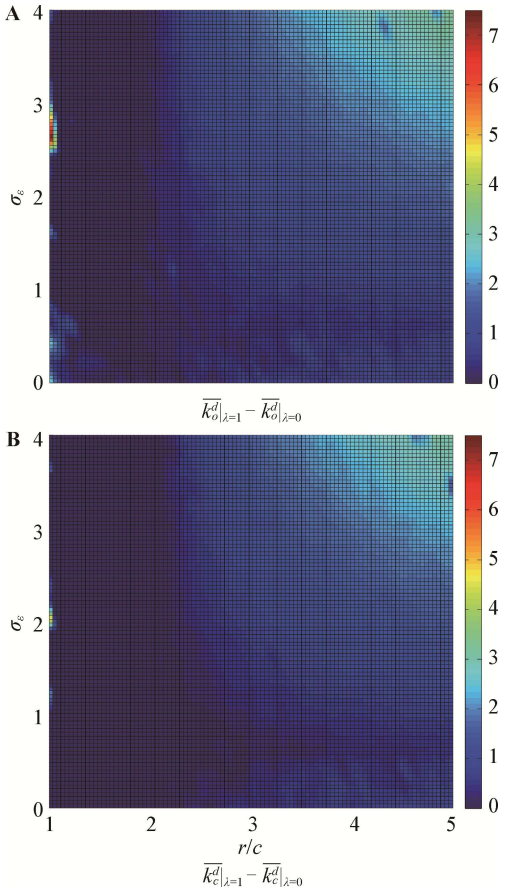


**Fig. 5 Differences in the magnitude of internal biases evolved with ($\lambda = 1$) and without ($\lambda = 0$) conscious deception costs**

In both plots the dishonest signaling cost is $\omega = 1$. The average owner internal bias evolved in the presence of external biases when $\lambda = 1$ is denoted by $\overline{k_o^d}|_{\lambda=1}$ and the same average when $\lambda = 0$ is denoted by $\overline{k_o^d}|_{\lambda=0}$. The average challenger internal bias evolved in the presence of external biases when $\lambda = 1$ is denoted by $\overline{k_c^d}|_{\lambda=1}$ and the same average when $\lambda = 0$ is denoted by $\overline{k_c^d}|_{\lambda=0}$. The plots show that the differences $\overline{k_o^d}|_{\lambda=1} - \overline{k_o^d}|_{\lambda=0}$ and $\overline{k_c^d}|_{\lambda=0} - \overline{k_c^d}|_{\lambda=1}$ are generally non-zero and that they increase with $r/c$ and $\sigma_\varepsilon$. This implies that internal biases evolved with conscious deception costs (i.e., with $\lambda = 1$) tend to become larger than internal biases evolved without this assumption (i.e., when $\lambda = 0$) as $r/c$ and $\sigma_\varepsilon$ increase.

## 3 Discussion

The owner-challenger model extends the one proposed by Johnson and Fowler (Johnson and Fowler, 2011) and offers two improvements over the original model. The first is that resources are never left unclaimed and the second is that individuals use a rational decision rule by taking into account the costs and benefits of each decision. Two versions of the model are considered: (1) the model with internal biases, introduced in Section 1.1, and (2) the model with role-dependent internal and external biases, introduced in Section 1.2. The model with internal biases aims to determine the evolvability of self-deceptive cognitive biases, given that decision-makers use a rational decision rule. The model with role-dependent internal and external biases introduces dishonest signaling and aims to test the theory proposed by Trivers, which states that self-deception should evolve if individuals face a selective pressure to deceive each other and that self-deceiving deceivers have an evolutionary advantage over other deceitful individuals because the former do not have to pay the cognitive costs of concealing a lie consciously. The baseline results with the owner-challenger model with internal biases introduced in Section 1.1 show that, given an assumed optimal decision rule taking proper account of probabilities costs and benefits of outcomes, suggested by Marshall et al. (2013b), biases provide a way for owners in the model to compensate for perception errors when their opponents are certain to fight. This is illustrated in Figure 2B, where it is shown that if there are perception errors ($\sigma_\varepsilon > 0$) then optimal behavior requires non-zero biases, the sole exception being when $r/c$ =2. If errors are not present ($\sigma_\varepsilon = 0$) then owners do not require any biases to behave optimally. But if errors are present then self-deception biases are required to gain the best payoffs in the long term. These results provided a baseline optimal level of self-deception to compare the extended model against.

The extended owner-challenger model with role-dependent internal and external biases was introduced in Section 1.2 with the purpose of examining the evolution

of self-deception as a supporting mechanism of deception. Symmetrically-biased individuals are those who self-deceive just as much as they attempt to deceive others, and can be considered to be 'unconscious' deceivers. On the other hand asymmetrically-biased individuals are those who project an image of themselves that differs from their self-perception, and are analogous to 'conscious' deceivers. The premise of Trivers' theory was incorporated into the model by having each decision-maker pay a dishonest signaling cost (for having the ability to deceive others through external biases) and a conscious deception cost (for exercising asymmetrically-biased strategies, i.e., for being consciously deceitful), emulating the physiological costs that deceivers in nature have to pay, according to Trivers' proposal (Trivers, 2011). Evolutionary simulations with the model show that when these costs are present then symmetrically-biased, self-deceiving individuals are more evolutionarily successful than those who attempt to deceive others while attempting to act on truthful information, when the benefit/cost ratio and the perception error are high enough (Figure 3). In other words, self-deceiving deceivers are more likely to evolve as the benefit/cost ratio and the perception error increase, when the conscious deception and dishonest signaling costs are present. The internal biases evolved when individuals attempt to deceive others generally exceed those that are evolved when individuals cannot deceive others (Figure 4), as information becomes noisier (i.e., as $\sigma_\varepsilon$ increases) and the benefit/cost ratio becomes larger. That is to say, in order to be evolutionarily stable, an increase in deceitful behavior requires an increase in self-deceiving behavior when there are physiological costs for exercising deception and for doing so consciously. Further simulations show that internal biases are also generally higher when there is a conscious deception cost (i.e., when $\lambda = 1$) than when this cost is absent (i.e., when $\lambda = 0$), and also that the difference between the internal biases evolved with and without this cost generally increases with $r/c$ and $\sigma_\varepsilon$ (Figure 5). From these numerical experiments it can be concluded that Trivers' theory generally holds true in situations of conflict if two conditions are met. First, the value of the contested resource must sufficiently exceed the cost of the fight required to claim the resource. Second, there must be a high enough degree of uncertainty in the information the decision-maker uses to assess its chances of winning the fight. As the value of the resource and the uncertainty increase, from the perspective of an individual it tends to payoff more in the long term to self-deceive as much as to attempt to deceive others, when conscious deception and dishonest signaling are physiologically costly. Then it should be expected that when these conditions are met, self-deceiving fighters evolve, as predicted by Trivers' theory (von Hippel and Trivers, 2011a; Marshall et al., 2013b).

The theory by Trivers has received has received considerable discussion, and it is possible that it will continue to be debated whether this theory correctly explains the apparent self-deception biases observed in humans, such as the ones presented in the introduction. The model presented in this article aims to test this theory in the particular case of a situation of conflict. The motivation for proposing this model is the point raised by commentators that the risk of injury faced by a self-deceiving, deceitful fighter is likely to be higher than the benefit received from discouraging an opponent from fighting by means of a deceitful exhibition of strength, and that therefore self-deceiving deceivers should not evolve. The model presented in this article simulates a situation of conflict where it is shown that, under the premises of the theory, self-deceiving, deceitful fighters do evolve. Given this, the model we have proposed serves as a first attempt to formally address the evolution of self-deception in situations of conflict, and the results obtained complement Trivers' proposal (von Hippel and Trivers, 2011a; Marshall et al., 2013b). In the future it would be of interest to examine the impact of more biologically-realistic assumptions on this result, such as population variation in perception errors, costs of conflict and so on.

## References

Alicke MD, Govorun O, 2005. The better-than-average effect. In: Alicke MD, Dunning DA, Krueger J ed. The Self in Social Judgment: Studies in Self and Identity. New York: Psychology Press, 85–106.

Backwell PR, Christy JH, Telford SR, Jennions MD, Passmore NI, 2000. Dishonest signaling in a fiddler crab. Proceedings of the Royal Society, Biological Sciences 267: 719–24.

Baker JE, 1987. Reducing bias and inefficiency in the selection algorithm. In: Grefenstette JJ ed. Proceedings of the Second International Conference on Genetic Algorithms. Hillsdale, NJ: L. Erlbaum Associates, 14–21.

Bandura A, Brooks ML, SwannWB, Jr., Buss DM, Dunning D et al., 2011. Open Peer Commentary on The evolution and psychology of self-deception. Behavioral and Brain Sciences, 34: 16–41.

Brooks ML, Swann WB, 2011. Is social interaction based on guile or honesty? Behavioral and Brain Sciences 34: 17–18.

Chambers JR, Windschitl PD, 2004. Biases in social comparative judgments: The role of nonmotivated factors in above-average

and comparative-optimism effects. Psychological Bulletin 130: 813–838.
Dougherty MRP, Gettys CF, Ogden EE, 1999. MINERVA-DM: A memory processes model for judgments of likelihood. Psychological Review 106: 180–209.
Frey U, Voland E, 2011. The evolutionary route to self-deception: Why offensive versus defensive strategy might be a false alternative. Behavioral and Brain Sciences 34: 21–22.
Funder DC, 2011. Directions and beliefs of self-presentational bias. Behavioral and Brain Sciences 34: 23.
Johnson DDP, Fowler JH, 2011. The evolution of overconfidence. Nature 477: 317–20.
Johnson DDP, Fowler JH, 2013. Complexity and simplicity in the evolution of decision-making biases. Trends in Ecology & Evolution 28: 446–447.
Kikuchi DW, Pfennig DW, 2010. Predator cognition permits imperfect coral snake mimicry. The American Naturalist 176: 830–834.
Klar Y, Giladi EE, 1997. No one in my group can be below the group's average: A robust positivity bias in favor of anonymous peers. Journal of Personality and Social Psychology 73: 885–901.
Kruger J, Dunning D, 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology 77: 1121–1134.
Marshall JAR, Trimmer PC, Houston AI, 2013a. Unbiased individuals use valuable information when making decisions: A reply to Johnson and Fowler. Trends in Ecology & Evolution, 28: 444–445.
Marshall JAR, Trimmer PC, Houston AI, McNamara JM, 2013b. On evolutionary explanations of cognitive biases. Trends in Ecology & Evolution, 469–473.
Martins M, 1989. Deimatic behavior in Pleurodemabrachyops. Journal of Herpetology 23: 305–307.
Maynard Smith J, 1982. Evolution and the Theory of Games. Cambridge: Cambridge University Press.
McCormick IA, Walkey FH, Green DE, 1986. Comparative perceptions of driver ability: A confirmation and expansion. Accident Analysis & Prevention 18: 205–208.
Pallier G, Wilkinson R, Danthiir V, Kleitman S, Knezevic G et al., 2002. The role of individual differences in the accuracy of confidence judgments.The Journal of General Psychology 129: 257–99.
Ramachandran SV, 1996. The evolutionary biology of self-deception, laughter, dreaming and depression: Some clues from anosognosia. Medical Hypotheses 47: 347–362.
Semple S, McComb K, 1996. Behavioural deception. Trends in Ecology & Evolution 11: 434–437.
Sharot T, 2011a. The optimism bias. Current Biology 21: R941–R945.
Sharot T, 2011b. The Optimism Bias: A Tour of the Irrationally Positive Brain. New York: Pantheon Books.
Svenson O, 1981. Are we all less risky and more skillful than our fellow drivers? Acta Psychologica 7: 143–148.
Trivers R, 2000. The elements of a scientific theory of self-deception. Annals of the New York Academy of Sciences 907: 114–131.
Trivers R, 2011. The Folly of Fools: The Logic of Deceit and Self-deception in Human Life. New York: Basic Books.
von Hippel W, Trivers R, 2011a. Reflections on self-deception. Behavioral and Brain Sciences 34: 41–56.
von Hippel W, Trivers R. 2011b. The evolution and psychology of self-deception. Behavioral and Brain Sciences 34: 1–56.
Zahavi A, 1975. Mate selection: A selection for a handicap. Journal of Theoretical Biology 53: 205–214.

# Supplementary Information

## 1 Details of the Evolutionary Model Presented in the Main Document

The evolutionary model used in Section 1.2 of the main document (*Self-deception can evolve under appropriate costs*) is as follows. A population of 500 individuals with role-dependent biases is initialized with random standard normal biases. Every generation the population is assorted in such a way that every individual $x$ is paired at random with exactly one adversary $y$ in the population. The capability difference between $x$ and $y$ (denoted by $A$) is a randomly-chosen standard normal value. Two encounters between $x$ and $y$ are simulated. In the first encounter $x$ plays owner and $y$ plays challenger. In the second encounter the roles are inverted. In each encounter each individual expresses only the biases corresponding to the role played by the individual. That is to say, when $x$ plays the role of an owner it estimates its own capability as $\theta_x + k_{o,x}$ and attempts to deceive $y$ into believing that $x$'s capability is $\theta_x + s_{o,x}$. On the other hand, when $x$ plays the role of a challenger it estimates its own capability as $\theta_x + k_{c,x}$ and attempts to project onto $y$ a capability equal to $\theta_x + s_{c,x}$. The biases of $y$ are expressed analogously. The fitness of $x$ is calculated as the average of the payoff received by this individual in the two encounters. The fitness of $y$ is calculated in the same manner.

One half of the population are selected through stochastic universal sampling (Baker, 1987). A new population is formed consisting of the selected individuals. Those that fail to be selected are replaced by randomly-chosen copies of the selected ones to complete the new population. Nearly 1% of the new population members have their genetically-encoded biases mutated with Gaussian noise. Evolution runs until the 5,000-th generation, when no considerable changes are observed and the population is assumed to be in equilibrium.

## 2 Implementation of the Model Presented in the Main Document

The source code that implements the model presented in Section 1.1 and the generalized model presented in Section 1.2 of the main document (*Self-deception can evolve under appropriate costs*) can be found in the compressed archive *Sources.zip* downloadable from http://goo.gl/FIqLzP. These sources were written in the *R* programming language (version 3.0.2) and *Matlab* (version 8.1.0.604). The operating system used was *Scientific Linux* release 6.5 (Carbon).

The simplified model with only internal biases and always-aggressive challengers described in Section 1.1 of the main document and the numerical analysis described in Section 2.1 are implemented in source file *Model_1.R* whereas the generalized model with role-dependent internal and external biases described in Section 1.2 and the simulations and numerical analyses described in Section 2.2 are implemented in source file *Model_2.R*. These sources are adaptations from the *R* code by Johnson and Fowler (2011).

The *Matlab* source files *SurfacePlot.m*, *SurfacePlot2.m* and *SurfacePlot3.m* are used for producing the plots in the main document (Fig. 2, 3, 4 and 5 in *Self-deception can evolve under appropriate costs*) from data generated by the *R* source code. Section 3 of this Online Supplementary Information presents the instructions for producing these plots using the sources in the compressed archive *Sources.zip* downloadable from http://goo.gl/FIqLzP.

## 3 Instructions for Producing the Plots Presented in the Main Document

### 3.1 Figure 2

Fig. 2A in the main document is produced following the steps below.

(1) Executing function *Model_1_1()* in the *R* source file *Model_1.R*, which outputs the data file *Expected_payoffs.csv*.

(2) Executing function *SurfacePlot()* in *Matlab* source file *SurfacePlot.m* in the same directory as *Expected_payoffs.csv*. The plot is output in the same directory as a file named *Expected_payoffs.pdf*.

Fig. 2B is produced by running function *Model_1_2()* in the *R* source file *Model_1.R*. The plot is output in the same directory as a file named optimal *owner_biases.pdf*.

### 3.2 Figure 3

Fig. 3A and Fig. 3B in the main document are produced following the steps below.

(1) Running *Model_2_1()* in the *R* source file *Model_2.R*, which outputs a folder named *Model_2_1* containing two .csv files named *1_owners.csv* and *2_challengers.csv*. In order to generate these two files, the program needs to output several auxiliary files into this folder first. While the program is running the size of the folder can reach over 300 MB, but these auxiliary files are deleted by the program upon completion.

(2) Running the *Matlab* source file *SurfacePlot2.m* in the same directory as *1_owners.csv* and *2_challengers.csv* (these two files can copied from folder *Model_2_1*). This program reads the two .csv files and produces the plots. The plots are output in the same directory as two files named *owners_sf_plot.pdf* and *challengers_sf_plot.pdf*.

Fig. 3C and Fig. 3D in the main document are produced following the steps below.

(1) Running *Model_2_2()* in the *R* source file *Model_2.R*, which outputs a folder named *Model_2_2* containing two .csv files named *1_owners.csv* and *2_challengers.csv*. In order to generate these two files, the program needs to output several auxiliary files into this folder first. While the program is running the size of the folder can reach over 100 MB, but these auxiliary files are deleted by the program upon completion.

(2) Running the *Matlab* source file *SurfacePlot2.m* in the same directory as *1_owners.csv* and *2_challengers.csv* (these two files can copied from folder *Model_2_2*). This program reads the two .csv files and produces the plots. The plots are output in the same directory as two files named *owners_sf_plot.pdf* and *challengers_sf_plot.pdf*.

Fig. 3E and Fig. 3F in the main document are produced following the steps below.

(1) Running *Model_2_3()* in the *R* source file *Model_2.R*, which outputs a folder named *Model_2_3* containing two .csv files named *1_owners.csv* and *2_challengers.csv*. In order to generate these two files, the program needs to output several auxiliary files into this folder first. While the program is running the size of the folder can reach over 100 MB, but these auxiliary files are deleted by the program upon completion.

(2) Running the *Matlab* source file *SurfacePlot2.m* in the same directory as *1_owners.csv* and *2_challengers.csv* (these two files can copied from folder *Model_2_3*). This program reads the two .csv files and produces the plots. The plots are output in the same directory as two files named *owners_sf_plot.pdf* and *challengers_sf_plot.pdf*.

### 3.3 Figure 4

Fig. 4A and Fig. 4B in the main document are produced following the steps below.

(1) Running *Model_2_4()* in the *R* source file *Model_2.R*, which outputs a folder named *Model_2_4* containing two .csv files named *1_owners.csv* and *2_challengers.csv*.

(2) Running the *Matlab* source file *SurfacePlot3.m* in the same directory as *1_owners.csv* and *2_challengers.csv* (these two files can copied from folder *Model_2_3*). This program reads the two .csv files and produces the plots. The plots are output in the same directory as two files named *owners_sf_plot.pdf* and *challengers_sf_plot.pdf*.

### 3.4 Figure 5

Fig. 5A and Fig. 5B in the main document are produced following the steps below.

(1) Running *Model_2_5()* in the *R* source file *Model_2.R*, which outputs a folder named *Model_2_5* containing two .csv files named *1_owners.csv* and *2_challengers.csv*.

(2) Running the *Matlab* source file *SurfacePlot3.m* in the same directory as *1_owners.csv* and *2_challengers.csv* (these two files can copied from folder *Model_2_3*). This program reads the two .csv files and produces the plots. The plots are output in the same directory as two files named *owners_sf_plot.pdf* and *challengers_sf_plot.pdf*.

## Appendix A: Expected payoff of an owner in the simplified model (Section 1.1)

The expected payoff of any owner with internal bias $k$ is calculated as a function of $k$, as

$$F(k) = \sum_{A \in S_A} P(A) I\left[\hat{p}_W(A,k) > c/r\right]\left[I(A>0)r - c\right]$$

This function approximates the owner's payoff as a summation of the weighted partial payoffs the owner receives in simulated encounters with challengers in a set $S_A$ of uniformly sampled values of $A$. The weighting factor is the probability of each $A$, denoted by $P(A)$. Each partial payoff in the summation is expressed in terms of an auxiliary identity function of the form $I(C_x)$, which returns unity if the condition represented by $C_x$ holds true and zero otherwise. The expression $\hat{p}_W(A,k)$ denotes the individual's estimated probability of winning, given $A$ and $k$[①]. Partial payoffs are added in the calculation of expected payoff if and only if $\hat{p}_W(A,k) > r/c$, i.e., if the owner decides to fight back. The partial payoff is $r-c$ if the owner is stronger than its opponent in the simulated owner-challenger encounter and $-c$ otherwise.

## Appendix B: Optimal owner behavior in the simplified model (Section 1.1)

In the model introduced in Section 1.1 it holds that if an owner $i$ accepts an opponent $j$'s challenge then the *marginal probability* that $i$ defeats $j$ is given by $p_W = P(\theta_i > \theta_j) = P(A > 0) = 1/2$ and therefore an owner's expected payoff from a single fight is $1/2r-c$. An unsophisticated owner that ignores relevant information and fights randomly each time with *fighting probability* $p_F$ is then expected to receive a mean payoff $F_{min} = p_F(1/2r-c)$ after repeated encounters with random challengers. Owners in this model, however, choose to fight only when their estimated expected payoff is positive, therefore their payoffs in the long term should be higher than $F_{min}$. Theorem 1 shows that an owner may fight only if $r/c$ is in the interval $(1, +\infty)$. In addition to this, Theorem 2 shows that for each $r/c$ there is a *capability superiority threshold for conflict*, denoted by $z$, and that every owner $i$ decides to repel a challenger if and only if $A > z-k_i$. Thus the fighting probability of an owner with bias $k_i$ is given by $p_F = P(A > z-k_i)$ and it increases as $r/c$ and $k_i$ increase and as $z$ and $c/r$ decrease.

**Theorem 1:** When $r/c \in (0,1]$ an owner never fights back.

**Proof:** If $r/c \in (0,1]$ then $c/r \in [1, +\infty)$ and $\hat{p}_W$ can never be above $c/r$. Therefore the owner's decision rule (i.e., "fight back if and only if $\hat{p}_W > c/r$") is never satisfied.

**Theorem 2:** Every owner $i$ fights $j$ if and only if $\theta_i + k_i > \theta_j + z$ where $z$, the capability superiority threshold for conflict, is the solution to the equation $\int_{-\infty}^{z} \Phi'(x)dx = c/r$ and $\Phi'(x)$ is the density function of a normal distribution with mean zero and standard deviation $\sigma_\varepsilon$.

**Proof:** From the fact that $\hat{p}_W = P\left(\hat{\theta}_j < \theta_i + k_i\right)$ and $\hat{\theta}_j \sim N(\theta_j, \sigma_\varepsilon)$ it follows that $\hat{p}_W = \int_{-\infty}^{\theta_i + k_i} \Phi(x)dx$, where $\Phi(x)$ is the density function of the normally-distributed $\hat{\theta}_j$. Let $v$ be the solution to equation $\int_{-\infty}^{v} \Phi(x)dx = c/r$ and let $z = v - \theta_j$. Therefore $\hat{p}_W > c/r$ (i.e., the owner fights back) if and only if $\theta_i + k_i > v = \theta_j + z$, where $z$ satisfies the equation $\frac{c}{r} = \int_{-\infty}^{v} \Phi(x)dx = \int_{-\infty}^{\theta_j + z} \Phi(x)dx = \int_{-\infty}^{z} \Phi'(x)dx$

[①] The estimate $\hat{p}_W(A,k)$ is given by $P\left(\hat{\theta}_j < \theta_i + k\right) = P\left(\hat{\theta}_j - \theta_i - k < 0\right)$. Since $\hat{\theta}_j$ is normally distributed with mean $\theta_j$ and standard deviation $\sigma_\varepsilon$ then the sum $\hat{\theta}_j - \theta_i - k$ is normally distributed with mean $\theta_j - \theta_i - k$ and standard deviation $\sigma_\varepsilon$. Therefore $P\left(\hat{\theta}_j - \theta_i - k < 0\right)$ can be restated as $P(\Theta < 0)$, where $\Theta \sim N(-A-k, \sigma_\varepsilon)$.

and $\Phi'(x)$ is the density function of a normal distribution with mean zero and standard deviation $\sigma_\varepsilon$.

Marshall et al. (2013b) show that an optimally-behaving individual in J&F's model (Johnson and Fowler, 2011) whose opponent is known to fight necessarily retaliates if its marginal probability of winning satisfies the inequality $p_W > c/r$. An optimal owner in the owner-challenger model presented in Section 1.1 should exhibit exactly the same behavior because its decision is made only in the knowledge that its opponent is determined to fight. If every owner had access to perfect information it would be able to compute accurately its probability of winning and follow the decision rule "fight if and only if $p_W > c/r$" (Marshall et al., 2013b). However, each owner only has access to its own estimated probability of winning ( $\hat{p}_W$ ), which is likely to deviate from the actual value ($p_W$) due to the individual's perception error ($\sigma_\varepsilon$) and internal bias ($k$). The individual uses this information to make a rational decision but due to the uncertainty present it is possible that at some point the individual makes the wrong choice.

Given any randomly selected owner $i$ and challenger $j$, then by definition $i$'s marginal probability of winning is $p_W = P(A>0) = 1/2$ and this means $i$ is objectively expected to be stronger than half the opponents it encounters. Therefore an optimal owner $i$ that hypothetically takes part in repeated encounters with random challengers should decide to fight in approximately half of those encounters, otherwise $i$ would be missing opportunities to defeat and increase its fitness; or it would risk itself in fights that are bound to be lost, which would in turn have a detrimental effect on its long term fitness. Then it can be predicted that the owners with the highest long-term payoffs must have internal biases that make $p_F$ equal to 1/2. That is to say, optimal behavior is a sufficient (but not necessary) condition for $p_F = 1/2$. Similarly, a fighting probability $p_F = 1/2$ is a necessary (but not sufficient) condition of optimality.