Opinion

Cell PRESS

Focus Issue: Overconfidence and deception in behaviour

On evolutionary explanations of cognitive biases

James A.R. Marshall[1], Pete C. Trimmer[2], Alasdair I. Houston[2], and John M. McNamara[3]

[1] Department of Computer Science and Kroto Research Institute, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, UK
[2] School of Biological Sciences, University of Bristol, Woodland Road, Bristol BS8 1UG, UK
[3] School of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, UK

Apparently irrational biases such as overconfidence, optimism, and pessimism are increasingly studied by biologists, psychologists, and neuroscientists. Functional explanations of such phenomena are essential; we argue that recent proposals, focused on benefits from overestimating the probability of success in conflicts or practising self-deception to better deceive others, are still lacking in crucial regards. Attention must be paid to the difference between cognitive and outcome biases; outcome biases are suboptimal, yet cognitive biases can be optimal. However, given that cognitive biases are subjectively experienced by affected individuals, developing theory and collecting evidence on them poses challenges. An evolutionary theory of cognitive bias might require closer integration of function and mechanism, analysing the evolution of constraints imposed by the mechanisms that determine behaviour.

The problem of overconfidence

In human psychology, overconfidence is typically taken to be the overestimation of one's own capabilities (see Glossary). This, and other apparent cognitive biases such as optimism, are well-documented phenomena [1] whose underlying neural mechanisms are becoming known [2,3]. However, a convincing evolutionary explanation of such phenomena is lacking.

Two recent high-profile publications have advanced proposals for evolutionary explanations of overconfidence and, given the general interest in the topic, garnered some attention. The first proposal, a model by Johnson and Fowler (J&F) [4], is adapted from the classic Hawk–Dove model of evolutionary game theory [5,6]. J&F consider a scenario in which individuals compare their estimated fighting ability against that of potential opponents when deciding whether to contest a resource, doing so only if they perceive themselves as more capable. By identifying conditions under which individuals should overestimate their fighting ability, J&F claim to show that overconfidence should evolve. The second is Trivers' theory of self-deception [7–9], which, although not the only candidate explanation [10], is perhaps the most influential currently. Among Trivers' primary arguments for the evolution of cognitive bias are that selective pressure exists for animals to deceive each other and that deception is more effective, and less cognitively costly, when the deceiver believes the deception; in the context of animal conflict, the explanation of overconfidence would be that acting as if one's abilities are greater than they really are can more effectively dissuade others from competition.

In this opinion article, we consider the logic of these and other approaches to explaining the evolution of cognitive biases. We emphasise that, to understand overconfidence in evolutionary terms, it is important to distinguish between the psychological definition of overconfidence and an operational definition in terms of rational behaviour. In many circumstances, rational behaviour can be taken to be the behaviour that maximizes the expected (mean) value of some reward [11]. In this framework, overconfidence is a

Glossary

Cognitive bias: an inaccurate view of the world. This is a psychological definition. A cognitive bias might produce rational behaviour or might result in an outcome bias.
Likelihood ratio: the most powerful statistic for determining, given data, which of two hypotheses is true. Given by computing the probability that each of the two hypotheses could have generated the observations, then taking the ratio of these. The likelihood ratio features in optimal, unbiased decision making between two options (Box 2).
Outcome bias: a departure from rational behaviour. This is an operational definition.
Overconfidence and optimism: bias relating to the probability of positive events ('optimism bias') or ability ('superiority bias') [2]. There are two interpretations of such bias: the operational and the psychological. The operational definition is that behaviour is not rational, and the resulting outcome bias is such that individuals behave as if the probability of a positive outcome is greater than it actually is. More generally, we define overconfidence and optimism as behaving as if good things or success are more likely to occur than is the case. The psychological definition is that individuals believe that the probability of good things or success is greater than it actually is; in other words they experience a cognitive bias. For the distinction between operational (behavioural outcome) and psychological (cognitive) definitions of bias see [25,27,28].
Prior probability: the probability that, in the absence of any relevant evidence, a hypothesis is true. For example, the probability, without having assessed an opponent, that the opponent is stronger. Can be set through learning within the lifetime of an individual or genetically over evolutionary time in a sufficiently predictable environment.
Rational behaviour: this term has various meanings in the literature [11]. We use it the sense in which it is used in behavioural ecology and evolutionary biology: the behaviour that maximises fitness (what [11] calls B-rationality). This is an operational definition.

Corresponding author: Marshall, J.A.R. (james.marshall@sheffield.ac.uk).

0169-5347/$ – see front matter
 http://dx.doi.org/10.1016/j.tree.2013.05.013

departure from rational behaviour in which an individual behaves as if a reward is more likely to occur than is actually the case. Many examples of psychological overconfidence might not constitute irrational behaviour, in that even if beliefs are biased, behaviour might still be rational. It is important to understand the conditions under which natural selection produces cognitive biases as a means of achieving behaviour that maximizes fitness. It is also important to be clear about what needs an explanation. Some biases do not produce rational behaviour. Explaining these biases is a challenge for evolutionary biology.

Throughout we refer primarily to overconfidence, but an evolutionary explanation of underconfidence, pessimism, and other negative biases is similarly important. Negative biases are simply the mathematical opposite of positive biases and therefore the same logical framework should underlie the explanation of both.

Overconfidence as evolved bias?

We start with J&F's overconfidence result [4]. Their model is based on contests over a resource. Let V be the expected fitness benefit from gaining the resource and C be the expected fitness cost of injury sustained in a fight. J&F simulate the evolution of rules, keeping V and C fixed. The end result is that the rule that evolves in a given environment depends on V and C. Crucially, making a decision about whether to contest a resource is based solely on the estimated chance of winning a fight, which is arrived at from an estimate of the opponent's capability and a (potentially biased) assessment of the focal individual's capability. Thus, the individual is constrained in that it cannot explicitly take into account V and C in making its decision (Figure 1), although the rules that evolve do reflect the values of V and C.

In Box 1, we expose the logic of J&F's model to show how optimal decision makers are constrained to almost always use a biased estimate of their individual capability. It is particularly important to note that this estimate is a parameter in a model. For J&F's result to correspond to psychological overconfidence, the animal must interpret the parameter as determining the probability of winning.

Box 1. Constrained rules and bias

Here we illustrate the consequences of constraining the form of decision rules, as in [4], for the evolution of bias. Suppose that an individual knows that its opponent will always fight. If the focal individual fights, there are two possible outcomes: it either wins and gains the resource (of value V) or loses and sustains injury (with cost C). Thus, if the animal's probability of winning the fight is p, its expected net benefit from fighting is:

$$pV - (1 - p)C. \quad [I]$$

Because, unlike the classic Hawk–Dove game [5,6], the benefit of not fighting is always zero, optimal behaviour requires that the individual fights if $pV - (1 - p)C > 0$; that is, its chance of winning a fight satisfies $p > p_c$ where

$$p_c = C/(V + C). \quad [II]$$

This critical value of p is below 1/2 if $V > C$ and above 1/2 if $V < C$. For example, if $V = 9$ and $C = 1$, $p_c = 0.1$, and thus if an animal had probability $p = 0.3$ of winning a fight it would still be optimal for it to fight. The decision rule used in [4] constrains an animal to fight if and only if its estimate of p is above 1/2. Given this constraint, the probability estimate must be greater than its true value to achieve optimal behaviour, and this estimate can be increased by adding a positive bias to the animal's estimate of its own capability. Similarly, if $V < C$, the same constraint leads to a requirement for optimal behaviour to be based on estimates that are less than their true value. Only when $p_c = 0.5$ should no bias evolve in the constrained model. This is why the estimate of individual capability almost always evolves to be biased in the model of [4]. In other words, because of the constraints imposed, optimal behaviour in that model requires that a biased estimate of ability should be adopted.

In other words, there needs to be a link between the value of the parameter and the state of mind of the animal, such that the animal's subjective probability of winning depends on the value of the parameter. This highlights a problem for all attempts to model the evolution of cognitive biases. A complete model must address what the animal believes; this is a point we return to below.

For many populations, it seems unlikely that over evolutionary time every member of the population will have the same values of V and C in every contest. Costs and benefits are likely to depend on circumstances; for example, the value of gaining a unit of food depends on the animal's energy reserves [12] and the cost of injury depends on future expectations [13]. If an individual can form some estimate of costs and benefits, it would be

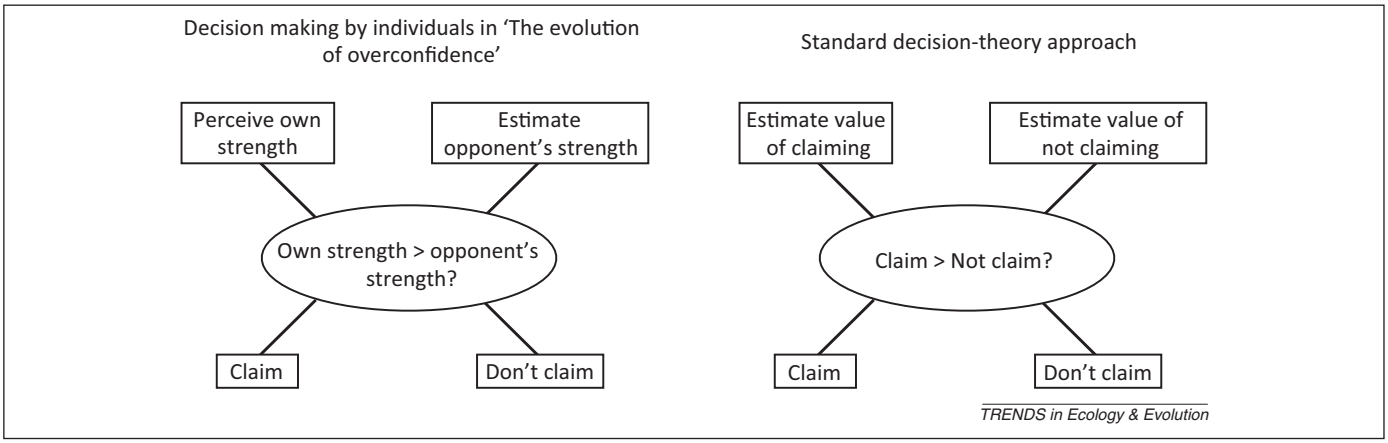

Figure 1. Johnson and Fowler [4] assume that rewards and costs are not explicitly taken into account when deciding whether to claim a resource (left). The particular form of decision rule used in [4] means that optimal decision making typically requires underconfidence or overconfidence, as described in Box 1. By contrast, standard decision theory takes account of expected costs and benefits when determining whether to perform some action (right). As described in Box 2, such decision rules cannot be outperformed.

Box 2. Optimal Bayesian decision making

Here we illustrate optimal decision making using a simple scenario in which an individual must decide whether to attempt to claim a resource, and where the world can be in one of two states: good or bad. Table I presents the possible payoffs to the decision maker according to its action and the true state of the world.

The decision maker is aware of the potential costs and benefits b and c, but does not have knowledge of the state of the world. Rather, the decision maker has two hypotheses: hypothesis 1 ($H1$) is that the world is good, whereas hypothesis 2 ($H2$) is that the world is bad. The decision maker has two further sources of information about the relative likelihoods of $H1$ and $H2$. First, the decision maker has access to a noisy signal that gives some uncertain information; in the case of a predator, this might be temperature or ambient light, which might predict whether airborne insect prey are likely to be found flying or not; let us call this x. Second, the decision maker also has prior probabilities for the world being in each of the two states; let us call these $P(H1)$ and $P(H2)$.

Standard decision theory [31] shows that a decision maker cannot do better on average than following the rule: only claim the resource when:

$$\frac{P(x|H2)}{P(x|H1)} < \frac{P(H1)(E(\text{payoff}|\text{claim, good}) - E(\text{payoff}|\text{don't claim, good}))}{P(H2)(E(\text{payoff}|\text{don't claim, bad}) - E(\text{payoff}|\text{claim, bad}))} \quad [III]$$

where $E(\text{payoff}|\text{claim, good})$ is the expected payoff from claiming a resource in a good environment and so on. For the scenario described in Table I, the optimal decision rule is thus:

$$\frac{P(x|H2)}{P(x|H1)} < \frac{P(H1)b}{P(H2)c}, \quad [IV]$$

which takes account of the likelihood ratio of the available evidence $P(x|H2)/P(x|H1)$, the prior probabilities $P(H1)$ and $P(H2)$, and the benefits b and costs c of correct and incorrect decisions.

A decision maker cannot do better on average than to use precisely the decision rule above; deviation from it can only reduce the expected payoff. As can be seen from Equation IV, this optimal decision depends on the prior probabilities of the states of the world and the benefits and costs of correct decisions and mistakes.

The decision scenario presented here is similar to that presented in [4], but is deliberately simplified for explanatory purposes. However, the logic of the optimal decision rule for J&F's scenario would be exactly that of Equation III above.

Table I. Payoffs for a decision maker attempting to claim a resource, or not, under different environmental states

		Environmental state	
		Good ($H1$)	Bad ($H2$)
Action	Claim	b	$-c$
	Don't claim	0	0

beneficial if its behavioural rule explicitly took these into account in addition to the probability of winning a fight [14] (Box 2). If the payoff for obtaining the resource is many times higher than the cost of failure, the odds of success must be very low for it not to be worth contesting the resource. In line with this, it has been shown that individuals do alter their competitive strategy according to the perceived resource value [15,16].

If there are variations in costs, benefits, and relative fighting ability, and these can be reliably estimated, an animal can do no better than to adopt the rule: contest the resource if the probability of winning satisfies $p > p_c$ where p_c is given by Equation II. Here p is the probability of winning the contest given the available information and is not distorted (biased), but should be computed based on prior probabilities and evidence on the individual's and the competitor's abilities (Box 2). Similarly, V and C are unbiased. Any other rule that did as well would have to make exactly the same decisions as this rule in all circumstances and thus cannot be distinguished on the basis of behaviour. Such a rule could be based on distorted values of V and p (where the distortion in p compensates for the distortion in V). This rule is still perfectly rational, but if V and p correspond to the appropriate aspects of an animal's subjective experience, the animal would be cognitively biased. In operational terms, the two rules are equivalent (and are both rational); they could be distinguished only if it were possible to examine the psychological mechanism used to reach a decision.

Overconfidence through self-deception?

Although unbiased estimates are best when analysing decision scenarios where the outcomes have fixed probabilities (Box 2), such as in betting on the outcome of a coin toss, perhaps things change if we consider the potential for biased estimates to alter the objective probabilities of decision outcomes? Specifically, perhaps acting according to a biased estimate of chances of success in some scenarios could change the probability of success and hence the expected reward. Consider the conflict over resources captured in the Hawk–Dove game, as discussed in the previous section. If acting overconfidently were sufficient to convince an opponent that their chances of success in a conflict were poor, a valuable resource might be gained without contest. Thus, acting overconfidently could be optimal in that the advantage of gaining the resource without a fight more than compensates for the cost of a fight should one occur. Selection for such deception in social situations underlies Trivers' proposals for the evolution of cognitive biases [7–9]. This is then coupled with an argument that deception can be more effective, and cognitively more efficient, when the deceiver believes the deception; in other words, individuals' ability to appear confident is subject to constraints. Constraints on the mechanisms underlying behaviour are thus supposed to be fundamental to the evolution of overconfident behaviour.

These ideas have attracted much attention, particularly from psychologists, whose critiques largely focus on issues such as the mental machinery involved in, or empirical support for, self-deception [17]. From the perspective of evolutionary biology, the main problem is the question of the evolutionary stability of 'self-deception', or bias. Trivers [8] notes the interest in evolutionary modelling of his theory, with populations including unbiased, non-self-deceiving individuals, but currently this problem does not seem to have attracted the attention of mathematical biologists. In fact, as some commentators on self-deception have noted (Frey and Voland, Gangestad in [17]), the theory of self-deception is closely related to the well-established field of signalling theory [18]. As this theory shows, animals should evolve to attend to honest signals, which convey reliable information, and ignore those that do not. In some ways, however, self-deception theory is too nuanced to allow a simple application of signalling theory.

In many applications of signalling theory, such as males signalling quality to potential mates, there are no disadvantages to being 'taken at your word'; if a poor-quality male manages to convince a high-quality female to mate with him, his fitness is greatly increased as a result. By contrast, consider the example of animal conflict. Acting overconfidently might discourage competitors from conflict in some instances, but eventually the protagonist's bluff is likely to be called; in this case, injury or death is a possible consequence of competing with a potentially much stronger opponent. Several of the commentaries on self-deception also make this point (Brooks and Swann, Frey and Voland, Funder in [18]). However, in replying to their critics, von Hippel and Trivers ignore this important issue [19]. A formal theory of self-deception would need to address the evolutionary stability of holding and acting on erroneous beliefs. Some recent work on 'persona' games, extending the evolution of preferences approach, has shown how deviations from Nash equilibria of games played by rational agents can be explained by assuming that agents signal binding commitments to particular personas in the forthcoming play [20]. Such an approach might be applied to explain the evolution of overconfidence; by signalling their commitment to deviate from rational play in the game by claiming the resource regardless of the relative costs and benefits of doing so, an individual could force a rational but stronger responding player to abandon a low-value resource rather than engage in an inevitable and overly costly contest. In some kinds of sequential game, the first player to choose can indeed determine the equilibrium outcome of the game, but not all games are of this type [21]. Most importantly, however, in this approach the evolutionary stability of making binding commitments to a particular persona is not considered.

Some approach to analysing the evolutionary stability of cognitive biases seems necessary, therefore. The model of Johnson and Fowler [4], discussed in the previous section, could easily be extended to begin to address the question of self-deception; in the original model, the evolved bias of individuals in their assessment of their own capability is private information. To model self-deception, a focal animal could simultaneously signal its own capability, including bias, as well as act on this biased assessment in making its decision whether to contest the resource. To study the evolutionary stability of such self-deception without the strong assumption of strategy-set restrictions, however, would require an additional category of individuals who are able to separately signal one biased level of capability, while simultaneously making decisions about whether to commit to a conflict, once conflict is clearly about to happen, based on a separate estimate of capability. Such a model is beyond the scope of this opinion article; however, the likely outcome of an evolutionary-stability analysis is that, if there is assumed to be no cost for doing so, individuals should dissociate their projected capability from their assessment of their own capability. Below, we consider the issue of costs and constraints and revisit the issue of what an apparent observation of cognitive bias might actually be diagnosing.

Concluding remarks: requirements for an evolutionary theory of cognitive bias

Optimality theory is used in evolutionary biology to determine how organisms should behave, with empirical deviations from optimality providing useful information to the experimenter or theoretician [22]. As argued above and elsewhere [23,24], the optimal approach to decision making is to use an unbiased Bayesian estimate of success probability, combined with costs and benefits of failure and success (Box 2). In game-theoretic situations in a population, the same logic applies, although the effects of adopting a particular strategy in potentially shifting the responses of other population members need to be taken into account in calculating the necessary estimate. However, as evolutionary biologists we do not suppose that organisms typically compute Bayesian posteriors, but simply that they (approximately) act as if they are doing so [23]. In asocial and in social decision problems, computing such estimates accurately is likely to be demanding; hence, constraints on computational speed or efficiency, or constraints on the evolvability of mechanisms, are likely to result in the use of heuristics, but these should give good approximations to the optimal estimate [4,23–25]. However, little or nothing can be learned from an unrealistic model in which the constraints mean that, to achieve optimal behaviour, a bias needs to be introduced.

In general, any theory about beliefs is likely to be difficult to verify. One might suppose that in humans the theory could be tested, because they can be interrogated about their subjective experience; however, humans are notoriously bad at objective introspection [26]. In this vein, it seems possible that the design of psychological studies into cognitive bias that focus on perceptions of the probability of events occurring without also considering the costs and benefits of different outcomes (e.g., [27]) might encourage subjects to conceptualise and report biased probabilities arising from an optimal system of decision making, where these biased probabilities might not be experienced subjectively during normal decisions.

In this opinion article, we have focussed on cognitive biases that are a means to implement optimal behaviour, given appropriate constraints. We have thus largely neglected departures from rational behaviour. A possible explanation of some departures is that they are not really departures, in that behaviour does not maximise some immediate currency but nevertheless maximizes fitness [28–30]. We argue that the use of terms such as optimism and overconfidence in interpreting such results does not provide additional insight [29].

One possible approach to developing an evolutionary understanding of overconfidence remains and is central to Trivers' proposals on the importance of self-deception. Part of Trivers' proposals hinge on evidence that it is cognitively less costly to deceive when the deceiver believes the information they are signalling. This suggests that considering the role of physiological constraints in understanding the evolution of apparently irrational behaviour might be crucial; under this assessment, such an understanding can be developed only by combining the study of mechanisms with the search for functional explanations [25].

Acknowledgements

The authors thank A.D. Higginson, T.W. Fawcett, two anonymous referees, and P. Craze for comments on earlier drafts. They also thank D. Johnson and J. Fowler for discussions. This work was supported in part by the European Research Council (Evomech Advanced Grant 250209 to A.I.H.).

References

1 Carver, C.S. *et al.* (2010) Optimism. *Clin. Psychol. Rev.* 30, 879–889
2 Sharot, T. (2012) *The Optimism Bias*. Constable and Robinson
3 Sharot, T. (2011) The optimism bias. *Curr. Biol.* 21, R941–R945
4 Johnson, D.D.P. and Fowler, J.H. (2011) The evolution of overconfidence. *Nature* 477, 317–320
5 Maynard Smith, J. and Price, G.R. (1973) The logic of animal conflict. *Nature* 246, 15–18
6 Maynard Smith, J. (1982) *Evolution and the Theory of Games*. Cambridge University Press
7 Trivers, R. (1985) Deceit and self-deception. In *Social Evolution*, pp. 395–420, Benjamin/Cummings
8 Trivers, R. (2011) *Deceit and Self-Deception: Fooling Yourself the Better to Fool Others*. Allen Lane
9 von Hippel, W. and Trivers, R. (2011) The evolution and psychology of self-deception. *Behav. Brain Sci.* 34, 1–16
10 Ramachandran, V.S. (1996) The evolutionary biology of self-deception, laughter, dreaming, and depression: some clues from anosognosia. *Med. Hypotheses* 47, 347–362
11 Kacelnik, A. (2006) Meanings of rationality. In *Rational Animals?* (Hurley, S. and Nudds, M., eds), pp. 87–106, Oxford University Press
12 Houston, A.I. and McNamara, J.M. (1988) Fighting for food: a dynamic version of the Hawk-Dove game. *Evol. Ecol.* 2, 51–64
13 Houston, A.I. and McNamara, J.M. (1991) Evolutionarily stable strategies in the repeated Hawk-Dove game. *Behav. Ecol.* 2, 219–227
14 Enquist, M. and Leimar, O. (1987) Evolution of fighting behaviour: the effect of variation in resource value. *J. Theor. Biol.* 127, 187–205
15 Leimar, O. *et al.* (1991) A test of the sequential assessment game: fighting in the bowl and doily spider *Frontinella pyramitela*. *Evolution* 45, 862–874
16 Mohamad, R. *et al.* (2010) Can subjective resource value affect aggressiveness and contest outcome in parasitoid wasps? *Anim. Behav.* 80, 629–636
17 Bandura, A. *et al.* (2011) Open peer commentary on 'the evolution and psychology of self-deception'. *Behav. Brain Sci.* 34, 16–41
18 Zahavi, A. and Zahavi, A. (1997) *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. Oxford University Press
19 von Hippel, W. and Trivers, R. (2011) Reflections on self-deception. *Behav. Brain Sci.* 34, 41–56
20 Wolpert, D.H. and Jamison, J. (2011) The strategic choice of preferences: the persona model. *Berk. Electron. J. Theor. Econ.* 11, 1
21 McNamara, J.M. *et al.* (2006) Is it better to give information, receive it, or be ignorant in a two-player game? *Behav. Ecol.* 17, 441–451
22 Parker, G.A. and Maynard Smith, J. (1990) Optimality theory in evolutionary biology. *Nature* 348, 27–33
23 McNamara, J.M. and Houston, A.I. (1980) The application of statistical decision theory to animal behaviour. *J. Theor. Biol.* 85, 673–690
24 McKay, R. and Efferson, C. (2010) The subtleties of error management. *Evol. Hum. Behav.* 31, 309–319
25 McNamara, J.M. and Houston, A.I. (2009) Integrating function and mechanism. *Trends Ecol. Evol.* 24, 670–675
26 Johannson, P. *et al.* (2005) Failure to detect mismatches between intention and outcome in a simple decision task. *Science* 310, 116–119
27 Weinstein, N.D. (1980) Unrealistic optimism about future life events. *J. Pers. Soc. Psychol.* 39, 806–820
28 Trimmer, P.C. *et al.* (2011) Decision-making under uncertainty: biases and Bayesians. *Anim. Cogn.* 14, 465–476
29 Houston, A.I. *et al.* (2011) Is optimism optimal? Functional causes of apparent behavioural biases. *Behav. Processes* 89, 172–178
30 McNamara, J.M. *et al.* (2011) Environmental variability can select for optimism or pessimism. *Ecol. Lett.* 14, 58–62
31 Green, D.M. and Swets, J.A. (1966) *Signal Detection Theory and Psychophysics*. Wiley and Sons