The Subtleties of Error Management*

RYAN MCKAY[1,2,3] and CHARLES EFFERSON[1,2]

[1]Institute for Empirical Research in Economics, University of Zürich

[2]Laboratory for Social and Neural Systems, University of Zürich

[3]Centre for Anthropology and Mind, University of Oxford

April 25, 2010

Running title: The Subtleties of Error Management

Word count: 6,431

*__Corresponding author:__ Ryan McKay, Centre for Anthropology and Mind, University of Oxford, 51-53 Banbury Road, Oxford OX2 6PE, United Kingdom *ryantmckay@mac.com.*

Abstract: Error management theory is a theory of considerable scope and emerging influence. The theory claims that cognitive biases do not necessarily reflect flaws in evolutionary design, but that they may be best conceived as design features. Unfortunately, existing accounts are vague with respect to the key concept of bias. The result is that it is unclear that the cognitive biases that the theory seeks to defend are not simply a form of behavioral bias, in which case the theory reduces to a version of expected utility theory. We propose some clarifications and refinements of error management theory by emphasizing important distinctions between different forms of behavioral and cognitive bias. We also highlight a key assumption, that the capacity for Bayesian beliefs is subject to constraints. This assumption is necessary for what we see as error management theory's genuinely novel claim: that behavioral tendencies to avoid costly errors can rest on systematic departures from Bayesian beliefs, and that the latter can be adaptive insofar as they generate the former.

Key words: error management theory; cognitive biases; behavioral biases; expected utility theory.

1 Introduction

An influential theory in the recent evolutionary psychology literature is error management theory (Haselton, 2007; Haselton *et al.*, 2009; Haselton and Buss, 2000, 2003; Haselton and Nettle, 2006). Proponents of error management theory argue that biologically evolved systems of decision and judgment reveal a general engineering principle. Namely, when one type of error is consistently more costly than others, behavior that suppresses the rate at which individuals commit the more costly error will be favored. This principle is true even though the total number of errors may be higher than they are in the case where individuals commit all types of error at the same rate. To illustrate the general engineering principle, consider the winnower, a farm implement designed to separate grains of wheat from chaff. The winnower works by using a blower to blow away the less dense and unwanted chaff, while the heavier wheat falls out the bottom. A winnower can be viewed as making two types of error – it can blow away wheat and it can fail to blow away chaff. If the former type of error is more costly, the operator can set the blower speed so that the winnower blows away wheat less often than it fails to blow away chaff, and this is optimal even though a higher speed would reduce the overall error rate. In essence, the operator's task is not to minimize errors, but rather to minimize costs.

Error management theory has been used to explain diverse phenomena, including but not limited to phenomena associated with auditory perception, human courtship, food preferences, and interracial aggression. The error management perspective, moreover, appears to be a fertile source of novel empirical predictions. Our intention is to propose some clarifications and refinements of the theory. Although we are impressed by its simplicity, scope and power, we argue that existing accounts are vague with respect to a key concept: the notion of a bias. Error management theory is explicity a theory of cognitive bias. According to the theory, cognitive biases do not necessarily reflect flaws in evolutionary design, but may be best conceived as design features. Our concern is that the theory does not adequately distinguish

between cognitive biases and behavioral biases. Below we discuss different conceptions of bias in an effort to clarify the subtleties associated with this important issue.

2 Behavioral biases

We begin by describing two types of behavioral bias. These biases are behavioral in the sense that identifying them does not require access to the internal (e.g. cognitive) states of the agent. The first conception is trivial and simply captures the notion of a statistical tendency or inclination. Put abstractly, given N possible behaviors, a *trivial* behavioral bias exists when the probability distribution over these N behaviors is not uniform. For example, if in a given period of time an individual can engage in one of two possible behaviors, say choosing heads or tails, a trivial behavioral bias exists if over many periods the individual chooses heads more often than tails or vice versa. To use a second example that has become a staple of the error management literature, a male in a singles bar has two options with respect to the females he encounters. He can approach them or not. If over time he approaches females more often than not, or vice versa, he once again shows a trivial behavioral bias.

Our second conception of a behavioral bias is more interesting and incorporates the notion of an error. In order to motivate the idea, imagine that the world can take one of N possible states, and agents can exhibit one of N possible behaviors. States of the world are independently determined each period, and a one-to-one mapping between states of the world and optimal behaviors characterizes payoffs. Specifically, if the world is in state x, the optimal behavior is b. For any $b' \neq b$, b' is a sub-optimal behavior given state x, and we will call it an "error" because it results in a lower payoff than b. Altogether, $N(N-1)$ types of error are possible. On our definition, behavior exhibits an *interesting* bias if over a sufficiently large number of periods the empirical distribution over these $N(N-1)$ possible errors is not ap-

proximately uniform. If someone calls "heads" when the coin is tails more often than she calls "tails" when the coin is heads, or vice versa, she displays an interesting behavioral bias. A man in a bar can approach women who are not attracted to him, yielding slaps in the face, and he can fail to approach women who are attracted to him, yielding missed opportunities. If slaps in the face outnumber missed opportunies, or vice versa, he displays an interesting behavioral bias.

Notably, behavior that is trivially biased may not be interestingly biased. Nor is behavior that is trivially unbiased necessarily interestingly unbiased. Consider a man who approaches 50% of the women he encounters and hangs back shyly the rest of the time. His behavior is unbiased by our first, trivial, definition. Nonetheless, if the man in question is George Clooney, then his behavior may be extremely biased by our second, interesting, definition. Missed opportunities may vastly outnumber slaps in the face. Indeed, behavior can even be trivially biased in one direction and interestingly biased in the other direction. If George Clooney is an incorrigible womanizer, he may approach women at a high rate but, given his charms, still have more missed opportunities than slaps in the face.

3 Cognitive biases

We now consider two conceptions of cognitive bias. As with our first conception of behavioral bias, our first conception of cognitive bias is trivial. Given N possible states of the world, a *trivial* cognitive bias exists when an individual's subjective probability distribution over these N states is not uniform. Our roving male in the singles bar has a trivial cognitive bias if he believes that a certain female is more likely to accept his advances than to reject them or vice versa. The reason this conception of a cognitive bias is so trivial is that the biased belief in question may be justified by the evidence. In particular, justified beliefs based on reliable information about the state of the environment will be non-uniform, and thus they will count as trivially biased. Debonair characters fully cognizant of their charms

will count as cognitively biased on this trivial conception, as will unattractive men with no illusions about their limited appeal to females. Notably, however, extremely small departures from uniform beliefs can in principle generate extreme behavioral biases of both types, trivial and interesting.

Our second conception of cognitive bias is more interesting than the first. In the everyday sense of the word, a "bias" refers to a particular tendency or inclination, but especially one that merits reproach in some way. Our definition of an *interesting* cognitive bias captures this idea insofar as it involves violations of a rational standard. By our definition, *interesting* cognitive biases obtain when beliefs depart systematically from those of an agent with Bayesian beliefs. Such a cognitively biased individual does not have beliefs that are theoretically optimal given the available information. Moreover, the individual's beliefs depart from the theoretical optimum in a systematic, rather than random, fashion.

4 The minimum necessary belief

To clarify the relationship between interesting behavioral biases and the cost asymmetries so widely discussed in error management theory, we develop a simple binary model similar to that found in Haselton and Nettle (2006). The decision maker faces one of two possible situations. We refer to these situations as "states" of the decision maker's environment. The decision maker can make one of two choices. One choice is best in one environmental state, while the other choice is best in the other environmental state. Accordingly, the decision maker can be wrong in two possible ways. The costs of these two types of error are not equal, and this is why the decision maker's problem involves a cost asymmetry. Aside from this cost asymmetry, the other relevant part of the problem is the decision maker's subjective beliefs about which environmental state actually obtains.

The intuition. To capture the intuition behind the model, consider Mr. Clooney again. When Mr. Clooney encounters a woman, he faces one of two environmental

states. Either the woman will be receptive to any advances from Mr. Clooney, or she will not. If receptive, the error Mr. Clooney can make is hanging back and missing a fitness-enhancing opportunity. If not receptive, the error Mr. Clooney can make is approaching her anyway and receiving a slap in the face. A lost mating opportunity costs more than a slap in the face. Importantly, Mr. Clooney's objective is emphatically not to minimize the probability of an error, but rather to minimize the expected error cost. That means he should make choices in a way that lead him to miss an opportunity less often than they lead him to receive a slap in the face. Equivalently, the belief Mr. Clooney requires that the woman will be receptive before he approaches her is weaker than the belief he requires that she will slap him in the face before he hangs back. For example, maybe the cost asymmetry is such that Mr. Clooney will approach a woman if he thinks she will be receptive with probability 0.25 and unreceptive with probability 0.75. If the cost asymmetry is even stronger, he requires an even weaker belief that the woman will be receptive before approaching her. As the following model shows, this reduction in the required belief responds very dramatically to increases in the cost asymmetry. None of this, however, requires Mr. Clooney to have an interesting cognitive bias. Imagine an extreme case in which Mr. Clooney only requires the belief that a woman will be receptive with probability 0.01. If 2% of the women in the world are receptive to Mr. Clooney's advances, and he knows this fact with perfect accuracy, he will approach every woman he meets and receive slaps in the face 98% of the time. His behavior may appear exotic and even astonishing, but we do not need an interesting cognitive bias to explain it. He's doing exactly what a payoff maximizer with Bayesian beliefs would do.

The model. Let the environment take one of two states, where $X \in \{0, 1\}$ specifies the state according to $P(X = 1) = p$. For either environmental state, the decision maker makes a choice, $C \in \{0, 1\}$. Table 1 shows how payoffs depend on the realized state of the environment and the decision maker's behavior.

[Table 1 about here]

The function $\pi : \{0, 1\} \rightarrow \mathbb{R}$ assigns an expected payoff (e.g. expected fitness) to each behavior. The expected payoff of choosing $C = 0$ is

$$\pi(0) = (1 - p)b - pd, \tag{1}$$

and the expected payoff of $C = 1$ is

$$\pi(1) = -(1 - p)c + pa. \tag{2}$$

Now assume the decision maker does not necessarily choose the behavior that maximizes the expected payoff. She does, however, choose the behavior that maximizes expected payoffs with a probability that increases as the difference in expected payoffs associated with the two possible behaviors increases. This captures the idea that choices may be noisy but still sensitive to expected payoffs. If, for example, the expected payoffs associated with the two choices are similar, behavior will be more like flipping a coin than when the expected payoffs are very far apart. One can implement this idea in various ways (Camerer, 2003), and the choice is not critical for our purposes. We choose the logit transformation. Specifically, for some $\lambda \in \mathbb{R}_+$,

$$P(C = 1) = \frac{\exp\{\lambda\pi(1)\}}{\exp\{\lambda\pi(0)\} + \exp\{\lambda\pi(1)\}} \tag{3}$$

$$= \frac{1}{1 + \exp\{\lambda[\pi(0) - \pi(1)]\}} \tag{4}$$

$$= \frac{1}{1 + \exp\left\{\lambda(b + c)\left[(1 - p) - p\left(\frac{a+d}{b+c}\right)\right]\right\}}. \tag{5}$$

The quantities $\lambda(b + c)$, p, and $(a + d)/(b + c)$ are all non-dimensional (unit free), and this shows us that three non-dimensional parameter combinations control choice probabilities. The parameter combination $\lambda(b + c)$ is simply a measure of how sensitively choice probabilities respond to differences in the values of $\pi(0)$ and $\pi(1)$.

When $\lambda(b+c)$ is small, decision making is very noisy. When large, the decision maker almost always makes the choice that maximizes the expected payoff. More importantly, depending on the state of the environment and the realized payoff matrix, the cost of an error is either $b+c$ or $a+d$. To see this, assume that the environment is in the zero state, and so $X=0$. If the decision maker makes an error (i.e. $C=1$), she pays c and loses the b she would have gotten had she not made an error. This yields a total cost of $b+c$. Similar reasoning shows that the cost of an error when $X=1$ is $a+d$. Expression (5) indicates that $(a+d)/(b+c)$ is the relevant measure of cost asymmetry in the binary settings that figure so prominently in error management theory. Because this asymmetry will appear repeatedly, we will call it z. By assumption (see Table 1), $z>1$, which simply means we are restricting attention for the moment to cases involving a real asymmetry.

One approach to showing how cost asymmetries lead to interestingly biased behavior is to identify the minimum beliefs an agent must have before choosing the behavior that avoids the more costly error with a probability greater than some specified threshold. Call this threshold C_T. To focus on the case in which even a weak belief that $X=1$ leads to a strong tendency to choose $C=1$, assume that $C_T > 1/2$. In other words, because avoiding the more costly error requires the decision maker to choose behavior 1, we are asking how high p needs to be before she chooses behavior 1 with a probability that exceeds C_T. In this case, the decision maker exhibits a trivial behavioral bias that favors $C=1$. With a bit of algebra, one can show that the individual chooses such that $P(C=1) > C_T$ so long as

$$p > \left(\frac{1}{1+z}\right)\left\{1 - \frac{1}{\lambda(b+c)}\ln\left(\frac{1-C_T}{C_T}\right)\right\}. \tag{6}$$

Call the quantity on the right-hand side of this inequality $g(z)$. Taking derivatives,

the result is

$$\frac{dg}{dz} = \frac{-\left\{1 - \frac{1}{\lambda(b+c)} \ln\left(\frac{1-C_T}{C_T}\right)\right\}}{(1+z)^2} < 0 \tag{7}$$

$$\frac{d^2g}{dz^2} = \frac{2\left\{1 - \frac{1}{\lambda(b+c)} \ln\left(\frac{1-C_T}{C_T}\right)\right\}}{(1+z)^3} > 0. \tag{8}$$

This result shows that, if we focus on the case in which agents show a sufficiently strong behavioral tendency to avoid the costly error (i.e. $P(C = 1) > C_T$), then the minimum belief that $X = 1$ required to produce this behavioral tendency is a convex decreasing function of the cost asymmetry (Figure 1).

[Insert Figure 1 about here]

In particular, the minimum belief decreases as the asymmetry increases because the first derivative (7) is negative when $z > 1$, which is true by assumption. The decrease is convex because the second derivative (8) is positive when $z > 1$. Altogether, this tells us that the belief can be arbitrarily weak so long as the asymmetry is sufficiently large. Moreover, the fact that the minimum necessary belief decreases in a convex fashion illustrates the power of cost asymmetries. We do not typically need a dramatic cost asymmetry to generate a large behavioral tendency because the biggest marginal effects are associated with small asymmetries. As shown in Figure 1, the difference between no asymmetry and a relatively small asymmetry will often be huge, while the difference between the same small asymmetry and a huge asymmetry will often be comparatively small.

5 A profusion of cognitive pathways

To develop the link between cognition and behavior further, we would like to expand the model above. Our reading of the error management literature suggests to us that researchers often take an interesting behavioral bias as evidence for an interesting

cognitive bias (see section 6 below). Here we show that inferences about even trivial cognitive biases are underdetermined by evidence of an interesting behavioral bias.

The intuition. The basic idea is to decompose cognition into various possible mechanisms that could produce behaviors favored by selection. In particular, we consider a relatively simple case with three parts. The first part is the decision maker's belief that he is making a decision for which cost asymmetries obtain. Second, the decision maker thinks the asymmetry, if it obtains, has a specific magnitude. Finally, the decision maker has a belief that, if an asymmetry obtains, the environment is actually in the state in which she can make the more costly error.

For example, perhaps in certain social settings (on the red carpet at the Academy Awards, say) Mr. Clooney considers slaps in the face, because of the extreme embarrassment caused, just as costly as missed opportunities. In these social settings, he does not face a cost asymmetry. In other social settings, however, slaps in the face are less costly than missed opportunities, and he does face a cost asymmetry. In any given social setting, as a result, Mr. Clooney has some beliefs about whether he faces the former type of social setting or the latter. Mr. Clooney also thinks that a cost asymmetry, if present, has a particular magnitude. Maybe he thinks that missed opportunities, when in the relevant social settings, are 10 times as costly as slaps in the face. Moreover, when in social settings of either sort, Mr. Clooney also has some belief that a given woman will respond positively to his advances. These are the three parts of Mr. Clooney's decision making we consider. The model below addresses the case in which an increase in the rate at which Mr. Clooney approaches women is advantageous. In this case, given the three cognitive mechanisms we consider, there are infinitely many ways to accomplish the required behavioral change. This is an example of how inferences about cognition can be radically underdetermined when one observes an interesting behavioral bias.

The model. Specifically, assume the decision maker considers one of two possible payoff matrices, one with a cost asymmetry and the other without. As detailed in Table 2, the payoff matrix that denotes the map from the space of choices by

environments to payoffs takes one of these two forms, $M \in \{0, 1\}$, where $P(M = 1) = q$.

[Table 2 about here]

If matrix 0 holds (i.e. $M = 0$), then no cost asymmetry obtains. If matrix 1 holds ($M = 1$), then we have the same cost asymmetry characterized by z that we considered in Section 4. With this kind of uncertainty about the actual payoff structure of the decision-making situation, the expected payoff of choosing $C = 0$ is

$$\pi(0) = (1 - p)\{(1 - q)b + qb\} + p\{(1 - q)(-c) - qd\}, \tag{9}$$

and the expected payoff of $C = 1$ is

$$\pi(1) = (1 - p)\{(1 - q)(-c) - qc\} + p\{(1 - q)b + qa\}. \tag{10}$$

Analogous to the simpler model above (equations (3) - (5)), choice probabilities take the form

$$P(C = 1) = \frac{\exp\{\lambda\pi(1)\}}{\exp\{\lambda\pi(0)\} + \exp\{\lambda\pi(1)\}} \tag{11}$$

$$= \frac{1}{1 + \exp\left\{\lambda(b + c)\left[(1 - 2p + pq) - pq\left(\frac{a+d}{b+c}\right)\right]\right\}}. \tag{12}$$

Choice probabilities now depend on four non-dimensional parameter combinations, $\lambda(b + c)$, p, q, and $(a + d)/(b + c)$. We would like to focus on three of these (z, p, and q), so we denote $P(C = 1) = f(z, p, q)$. Given that $z > 1$ by assumption, one

can show that

$$\frac{\partial f}{\partial z} = \frac{\lambda(b+c)pq \exp\left\{\lambda(b+c)\left[(1-2p+pq)-pq\,(z)\right]\right\}}{\left[1+\exp\left\{\lambda(b+c)\left[(1-2p+pq)-pq\,(z)\right]\right\}\right]^2} > 0 \tag{13}$$

$$\frac{\partial f}{\partial p} = \frac{\lambda(b+c)(2+q(z-1)) \exp\left\{\lambda(b+c)\left[(1-2p+pq)-pq\,(z)\right]\right\}}{\left[1+\exp\left\{\lambda(b+c)\left[(1-2p+pq)-pq\,(z)\right]\right\}\right]^2} > 0 \tag{14}$$

$$\frac{\partial f}{\partial q} = \frac{\lambda(b+c)p(z-1) \exp\left\{\lambda(b+c)\left[(1-2p+pq)-pq\,(z)\right]\right\}}{\left[1+\exp\left\{\lambda(b+c)\left[(1-2p+pq)-pq\,(z)\right]\right\}\right]^2} > 0. \tag{15}$$

Although expressions 13 - 15 look formidable, all we really care about is the fact that all three partial derivatives are unambiguously positive. Because these derivatives tell us how the function changes in response to our three cognitive variables, the result tells us that we have three separate cognitive mechanisms for increasing the decision maker's tendency to avoid the more costly error. If selection favors a behavioral modification of this sort, then we have a profusion of cognitive pathways. An increase in the perceived payoff asymmetry given that one exists (i.e. z), an increase in the belief that the environment is in the state where a costly error is possible (i.e. p), or an increase in the belief that a cost asymmetry actually obtains (i.e. q) will all do the trick. So will infinitely many combinations of the three. Indeed, we can even reduce the forces favoring $C = 1$ in one or two of the cognitive dimensions we consider and still get a local increase in the probability of choosing $C = 1$. This unusual possibility simply requires that the remaining cognitive dimensions involve countervailing forces that overcompensate. Formally, for the probability of $C = 1$ to increase at the margin, all we need is a positive total differential,

$$df = \frac{\partial f}{\partial z}dz + \frac{\partial f}{\partial p}dp + \frac{\partial f}{\partial q}dq > 0. \tag{16}$$

As (16) makes clear, this condition can be satisfied in various ways, and this remains true even if some cognitive mechanisms, in isolation, are actually increasing the probability of choosing $C = 0$ and by extension increasing the decision

maker's tendency to make the *more* costly error. The net result is that, if we only have behavioral data showing a strong tendency to choose $C = 1$, we have an extremely restricted ability to draw inferences about cognitive processes - there are just too many free parameters. A strong behavioral tendency can arise from various combinations of at least three different quantities related to cost asymmetries. In essence, inferences about cognition are radically underdetermined. This is a version of the more general point that adaptationist thinking is often more useful for making predictions about adaptive phenotypes than predictions about the proximate mechanisms by which selection will fashion such phenotypes.

6 Error management biases

Having sketched several definitions of bias, we now consider the claims of error management theory in relation to this taxonomy. As noted above, error management theorists argue that cognition is biased, and they seek to recast cognitive biases as evolutionary adaptations to ancestral environments. What, however, do they mean by cognitive biases? Which, if either, of our two conceptions of cognitive bias do they have in mind? Given the trivial nature of the first conception, one might expect that the second, interesting conception is the only possibility. After all, probably every cognitive agent that has ever existed has had trivial cognitive biases. If you consider it relatively unlikely that the world will end in the next five minutes, or that a talking horse will be elected the next Pope, then you are cognitively biased in the trivial sense. No one has ever complained that humans have cognitive biases of this sort. Any attempt to defend or reframe cognitive biases, therefore, must presumably refer to cognitive biases of the second, more interesting sort, namely beliefs that depart systematically from the beliefs a hypothetical Bayesian would have.

The situation, however, is not so clear. In their article, "The Paranoid Optimist: An Integrative Evolutionary Model of Cognitive Biases," Haselton and Nettle (2006, p. 48) distinguish between errors of belief and errors of behavior:

> In general, there are four possible outcomes consequent on a judgment or decision. A belief can be adopted when it is in fact true (a true positive or TP), or it cannot be adopted and not be true (a true negative or TN). Then there are two possible errors. A false positive (FP) error occurs when a person adopts a belief that is not in fact true, and a false negative (FN) occurs when a person fails to adopt a belief that is true. The same framework applies to actions. An FP occurs when a person does something, although it does not produce the anticipated benefit, and an FN when a person fails to do something that, if done, would have provided a benefit.

Here the authors acknowledge a distinction between belief and behavior. Erroneous beliefs involve a mismatch between belief and reality. Perhaps the world is in state x, but the individual has an unjustifiably strong belief that it is in state x'. Erroneous behavior, instead, occurs when an individual takes an action that is not optimal given the state of the world. On the next page of the article, however, Haselton and Nettle (2006, p. 49) explictly conflate belief and behavior when they say, "By adopting a belief, we mean behaving or reasoning as if the corresponding proposition were true." In our view this conflation creates substantial confusion. After all, if approaching uninterested females in bars carries no meaningful costs but occasionally large benefits, selection will favor men who approach women *even if they only have an exceedingly weak belief that they will succeed with any given woman.* A man may be nearly convinced that a particular woman is not interested in him. Nonetheless, if he admits any hope, however slim, he will behave *as if* she is interested. He will do so precisely because the fitness costs of a missed encounter are so much larger than the costs of brief embarrassment. Crucially, this bias in behavior does not require an interesting bias in beliefs.

Given this conflation of belief and behavior, the possibility arises that the cognitive biases Haselton and Nettle (2006) seek to defend are not cognitive biases at

all. Indeed, much of what they discuss gives the impression that the biases they seek to recast as design features are not interesting cognitive biases but interesting behavioral biases. Consider, for example, this extract from their paper's abstract, "EMT [error management theory] predicts that if judgments are made under uncertainty, and the costs of false positive and false negative errors have been asymmetric over evolutionary history, selection should have favored a bias toward making the least costly error." Haselton and Nettle, of course, are exactly right; selection should favor a bias toward making less costly errors. This, however, is an interesting behavioral bias. Biased behavior simply requires a biased decision rule. It does not require a cognitive bias in the sense of beliefs that depart systematically from the theoretical optimum given the available evidence. Haselton and Nettle (2006, pp. 48-49) repeatedly mention decision rules, implying that humans have evolved to make decisions in a manner that maximizes reproductive success:

> According to EMT, certain decision-making adaptations have evolved through natural selection to commit predictable errors. Whenever there exists a recurrent cost asymmetry between two types of errors over evolutionary time, selection will fashion mechanisms biased toward committing errors that are less costly in reproductive currency... EMT predicts that human psychology contains evolved decision rules biased toward committing one type of error over another.

Without doubt, actors showing interesting behavioral biases will often end up making more errors than actors who minimize the overall error rate. Again, however, the task is not to minimize errors but rather to minimize costs. If this is the basic claim of error management theory, it is certainly persuasive. Nonetheless, the problem is that the same claim follows in a straightforward way from expected utility theory (von Neumann and Morgenstern, 1944), and the idea is central to the game theoretic solution concept known as "risk dominance" (Harsanyi and Selten, 1988). From this perspective, error management theory might seem somewhat derivative.

Fortunately, however, we think that error management theory provides a basis for a more novel claim, namely that interesting behavioral biases can rest on interesting cognitive biases, and that the latter can be adaptive insofar as they generate the former. More rigor is necessary, however, in formulating this claim precisely.

In what follows, we will attempt to bring this claim into sharper focus. Before proceeding, however, we would like to touch on another important point. Haselton and Nettle (2006, p. 49) put forward an unbiased decision rule for "adopting the belief S," where adopting the belief S is apparently viewed as a binary outcome equivalent to exhibiting the behavior S. If s is a state of the world, $\neg s$ is the complementary set of states, and e is available evidence, then the decision rule is to adopt or exhibit S when $P(e \mid s)/P(e \mid \neg s) > 1$. This is indeed an unbiased decision rule for behavior under our second, interesting definition of a behavioral bias, but only if the prior probabilities of s and $\neg s$ are equal (Bar-Hillel, 1980; Kahneman and Tversky, 1973). Although Haselton and Nettle acknowledge this assumption, they do not emphasize the fact that it rarely holds. Applying such a decision rule when the priors are not equal can lead to serious consequences. Consider a situation in which an airport security guard is trying to decide whether or not to subject a certain passenger to extra screening measures. Judging by his clothing, the passenger is a Muslim. The question of interest to the security guard is whether the passenger is also likely to be a terrorist. Let S be the decision to implement extra screening, and e be the empirical evidence available to the security guard. Further let s be the state in which the passenger is in fact a terrorist and $\neg s$ the state in which the passenger is not. The security guard estimates that $P(e \mid s) = 0.99$, i.e. 99% of the world's active terrorists have clothing that makes them seem Muslim. The guard also estimates that $P(e \mid \neg s) = 0.2$, which means that 20% of the world's peacable citizens have clothing that makes them seem Muslim. Because $0.99/0.2 > 1$, the guard subjects the passenger to additional screening, as indeed he does with all passengers who appear to be Muslim.

Whatever limited merits the security guard's decision rule might have (see Press,

2009), it is clearly biased. The reason, of course, is that the security guard fails to account for the base rates of terrorism and peacable tendencies in the general population. Given that the vast majority of Muslims who pass through airport security are not terrorists, the guard's decision rule generates a huge number of false positives. A more general decision rule is

$$\frac{P(s \mid e)}{P(\neg s \mid e)} = \frac{P(e \mid s)P(s)}{P(e \mid \neg s)P(\neg s)} > 1. \tag{17}$$

This rule incorporates the prior probabilities of s and $\neg s$. Because terrorists are vastly outnumbered by non-terrorists, this rule yields a very different decision from the rule above. Even if the security guard assumes that the proportion of active terrorists in the world is 0.1, which is surely a radical overestimate, then $P(s \mid e)/P(\neg s \mid e) = 0.55$, and the guard will let the passenger in question through unmolested. Indeed, she will let all passengers through unmolested. This new rule, of course, is also biased because it only generates false negatives. In fact, it is completely worthless as a rule because it never discriminates between terrorists and non-terrorists. In terms of the absolute numbers of errors, however, this rule generates far fewer than the previous rule.

7 The evolution of interesting cognitive biases

We now return to what we see as the interesting and novel claim that error management theory makes. We view the claim as having three parts.

- Interesting behavioral biases require patterns of behavior that result in a non-uniform distribution over the possible types of error. Such biases will be adaptive in many situations.

- Beliefs and decision-making rules that yield such biases under appropriate conditions will be favored by selection.

- In general, optimal beliefs will correspond to the beliefs of a Bayesian because, all else equal, we cannot do better than Bayesian beliefs in deliberative contexts. Nonetheless, in principle Bayesian beliefs could be unavailable because, say, they are too expensive neurologically. If something like this is the case, *systematic* departures from the beliefs of a Bayesian, which constitute interesting cognitive biases, may be adaptive provided they lead to behaviors that are interestingly biased in the adaptive direction.

The idea that some constraint may preclude Bayesian beliefs is crucial, but Haselton and Nettle (2006) do not mention it. For example, they argue (p. 48),

> Because men's reproduction is limited primarily by the number of sexual partners to whom they gain sexual access, a bias that caused men to err on the side of assuming sexual interest would have resulted in fewer missed sexual opportunities, and hence greater offspring number, than unbiased sexual inferences. Therefore, natural selection should favor sexual overperception in men.

It is true, to give an extreme example, that a man who pursues every woman he meets because he thinks they all want to sleep with him will have no more missed sexual opportunities than someone with Bayesian beliefs. Indeed, his number of missed opportunities will be zero, and thus it will probably be strictly less than the number of missed opportunities for someone with Bayesian beliefs. On the other hand, chasing all those women who in fact are not interested will probably bring costs in other fitness-relevant domains. As a result, departures from Bayesian beliefs are not obviously adaptive. Bayesian beliefs, after all, are distinguished by the fact that they are theoretically justified given all the available evidence. What Haselton and Nettle need is an additional assumption, namely the assumption that some kind of constraint limits the ability of selection to implement perfectly Bayesian beliefs. To see how *systematic* departures from Bayesian beliefs might be adaptive given such a constraint, consider the following.

As Haselton and Nettle point out, when some object making a sound is approaching, and one needs to prepare for the arrival of the object, estimating an arrival time that is systematically too early is better than a time that is systematically too late (Neuhoff, 2001). What Haselton and Nettle overlook is that, all else equal, it is better still to estimate an arrival time that is neither systematically too early nor too late. Such optimal estimates are Bayesian. Asymmetric costs may affect the optimal point in time for taking an action relative to the estimated arrival, but this is a separate issue entirely. Suppose, however, that some unspecified constraint limits our ability to have perfectly Bayesian beliefs. As a result, our subjective distribution of arrival times is not the same as a Bayesian would have. Two distributions, of course, can differ in several ways, but for simplicity we focus on the case in which two distributions differ only in terms of their means. In this case, systematically underestimating arrival times should be advantageous relative to systematic overestimation or unsystematic misestimations. This conclusion follows from the relatively weak assumption that being prepared a bit too early is usually better than being prepared a bit too late.

We have refined the claims of error management theory by providing precise definitions of behavioral and cognitive biases and by highlighting the importance of a key assumption. Namely, assuming constraints limit our ability to form Bayesian beliefs, some interesting cognitive biases, which are here defined as systematic departures from the beliefs of a Bayesian, may be adaptive in many situations. What the theory now needs is an account of why constraints might prohibit Bayesian beliefs. One possibility is that cognition is to a large extent general purpose, and performance in some domains is negatively correlated with performance in other domains. Dependent on the relevant cost functions, globally optimal performance in this case may entail sub-optimal performance in some or all of the individual domains.

From an evolutionary psychological perspective, however, this idea clashes with an important assumption of the discipline, namely the assumption that the human

mind is not a general-purpose cognitive system but rather a collection of functionally specialized modules (Cosmides and Tooby, 1997). Insofar as these modules interact, effects in one module may constrain optimization in other modules. Insofar as these modules operate independently, however, natural selection should be free to optimize in one domain without affecting other domains. Evolutionary psychologists are notable for positing a high degree of modularity, which would weaken the constraint argument above. An alternative explanation for constraints, which is perhaps more consistent with the hypothesized degree of modularity associated with evolutionary psychology, is that the neural circuitry required to implement Bayesian inference in all or most domains is prohibitively costly. We leave it to others to evaluate such possibilities.

8 Empirical evidence

The final point we want to make concerns the empirical evidence for interesting cognitive biases. We have claimed that, to argue for the adaptive nature of systematic departures from Bayesian inference, one needs to invoke constraints. But why make this argument at all? Why not just assume that evolved cognitive systems incorporate data in the optimal way by producing theoretically justifiable beliefs consistent with Bayesian updating? This seems to be the most parsimonious approach in the absence of contrary evidence. In our view, however, contrary evidence is available, but to date published error management accounts have not clearly discriminated between evidence for interesting cognitive biases and evidence for mere behavioral biases. We conclude by considering three different domains of evidence in an effort to demonstrate the interpretive complexities involved.

8.1 Protective biases in disease defense

In their overview of different biases, Haselton and Nettle (2006; see also Haselton et al., 2009) mention biological systems designed to protect the body from harm, via mechanisms such as allergy and cough. Such defense systems are often mobilized in the absence of any real threat. Coughs, for example, probably frequently represent false positives, and for this reason dampening them with medication often leads to few negative effects (Nesse, 2001). This phenomenon seems a plausible candidate for an interesting behavioral bias. Specifically, the more costly error, not coughing when a real threat exists, occurs less frequently than the less costly error, namely coughing when no threat is present. We see no need, however, to invoke an interesting cognitive bias here. Although interesting cognitive biases may reinforce adaptive interesting behavioral biases in some settings, evidence of an interesting behavioral bias is not evidence for an interesting cognitive bias. We may cough a lot because coughs are cheap and disease is expensive, but this observation does not require us to posit a distortion in the body's processing of information.

8.2 Hot hand behavior

The hot hand phenomenon occurs when research participants expect streaks in sequences of hits and misses, the probabilities of which are, in fact, independent. Does this phenomenon reflect an interesting cognitive bias? Wilke and Barrett (2009) have suggested that the phenomenon reflects a psychological adaptation for successful foraging. We consider this hypothesis in some detail here.

To begin, we define two relevant foraging errors: 1) *Over-foraging:* Having discovered a desired resource, an individual continues to forage when the reality is that continued foraging will prove fruitless; 2) *Under-foraging:* Having discovered a desired resource, the individual ceases foraging when continued foraging would have (perhaps literally) yielded further fruits. Now, imagine that an anthropologist visits a particular field location and observes interestingly biased foraging behavior in that

location. Specifically, she notes that foragers there are more likely to over-forage than to under-forage. What might this hypothetical researcher conclude? We distinguish two distinct possibilities, both compatible with the data she has collected.

The first possibility is that the foragers are flawless Bayesians who maximize expected utility under appropriate incentives. These cognitively laudable creatures reliably track the relevant environmental contingencies. The bias toward over-foraging that they exhibit may stem from the fact that they accurately perceive that their resources of interest are spatially clumped (positively autocorrelated at a given scale). Having discovered some of the resource, they may accurately infer that further foraging at this scale is statistically more likely to yield fruit than to not yield fruit. Alternatively, or additionally, the foragers may accurately perceive that the cost of the two foraging errors is asymmetric. If they consider that the cost of a missed foraging opportunity is greater than the cost of fruitless foraging, then they may adjust their foraging behavior so as to produce the observed non-uniform distribution over error types (i.e. they will tend to over-forage rather than under-forage).

The second possibility is that the observed interesting behavioral bias reflects an interesting cognitive bias in the foragers. It may be that some kind of constraint in the evolutionary past (e.g. a historical or ecological constraint) limited selection's ability to implement Bayesian beliefs. In this case, systematic departures from Bayesian updating (interesting cognitive biases) may have proven adaptive insofar as they generated adaptive tendencies to avoid costly errors. The present-day foragers retain this cognitive bias (which may or may not still be adaptive, depending on whether the relevant environmental contingencies are different to those that obtained in the environment of evolutionary adaptedness).

One of our main points in this paper is that the evidence the anthropologist has collected in this hypothetical example is consistent with both hypotheses. Having simply inferred interesting behavioral biases, therefore, researchers should be very cautious about inferring interesting cognitive biases – especially given that the latter possibility is somewhat unparsimonious (involving, as it does, the constraint

assumption).

A recent study by Wilke and Barrett (2009), however, represents a much more rigorous and controlled investigation of the issue than that conducted by our hypothetical anthropologist. Wilke and Barrett designed computer tasks in which participants could forage for different resources. Participants were presented with a sequence of hits (e.g. fruits) and misses (e.g. no fruits) and were required, after each event in the sequence, to predict whether the next event would be a hit or a miss. In this paradigm there were two relevant errors: guessing hit when the subsequent trial was a miss and guessing miss when the subsequent trial was a hit. Crucially, however, the (opportunity) cost of these errors was perfectly symmetric – participants were paid a standard amount for each correct response (whether true positive or true negative), and were not penalized for false alarms or misses. Observed behavior in this study thus transparently revealed cognition in a way that it did not in the hypothetical anthropological study above because there is every reason to suspect that a participant's button press for *hit* indicated a belief of at least $1/2$ that the next trial would be a hit.

Given 1) that responses appeared to indicate an assumption of clumpiness across all resource types, although distributions were in fact equivalent to a series of coin tosses, and 2) that there was no evidence of learning over time (participants were uniformly hot handed over the course of the computer task – if anything, hot-handedness *increased* as the experiments progressed), the conclusion that participants showed interesting cognitive biases in this domain does not initially seem unwarranted. However, there is an important caveat here. Assume, for a moment, that the participants were in fact flawless Bayesians. In this case the participants would soon correctly infer that the distribution of hits and misses was uniform. Given this knowledge, what guessing strategy should they have employed? The answer is that no strategy would have presented itself as preferable to any other. Participants would have fared just as well if they had guessed at random, if they had chosen hit on every trial, if they always chose the event that had just occurred, or any other

strategy whatsoever. Our point is that hot hand behavior had absolutely no payoff consequences here. The fact that participants displayed hot hand tendencies is not therefore at odds with their having had Bayesian beliefs. Although the fact that they displayed hot hand behavior is certainly suggestive, we cannot be sure that they had hot hand beliefs.

In order to demonstrate an interesting cognitive bias experimentally, it is necessary to set things up such that cognitive biases will yield one sort of behavior, and Bayesian beliefs will yield another. Because all behavior is equally rational in the situation where participants must, in effect, guess the outcomes of repeatedly flipping a fair coin, the above experiment does not ultimately accomplish this. Another possible approach would be to artificially constrain the distribution of hits and misses in the experiment such that there is an equal global proportion of each, and to tell the participants this in advance. Truly Bayesian participants in that situation would update the probabilities of hits and misses after each event, and at each point in time they would choose the option that had occurred least often up to that point. For example, if the participant knows that 50 hits and 50 misses will be presented in random order, then her subjective probability of a hit at the outset will be 50/50. If the first two events turn out to be hits, then her updated probability of a hit on the third event will be 48/98, and she will choose miss. Note that the Bayesian strategy here will not consist in simply choosing the event that has just occurred – so evidence of hot hand behavior here would be convincing evidence of interesting cognitive bias. Such behavior would have a detrimental effect on payoffs.

A different, and perhaps even better approach, would be to set things up such that the outcomes in the experiment are *negatively* autocorrelated. Wilke and Barrett (2009) found that participants displayed hot hand tendencies even for resources (such as bird nests) that are dispersed (rather than clumped) in the real world. However, these resources were not negatively autocorrelated in their experiment – as with the other resources they used, the distribution of bird nest hits and misses was equivalent to that generated by a series of coin tosses. One way of producing neg-

atively autocorrelated outcomes would be to use two different coins, one biased toward heads and the other biased toward tails, and to repeatedly flip these coins in turn. A Bayesian in that situation would end up exhibiting reverse hot hand behavior, so standard hot hand behavior in that situation would, again, be convincing evidence of interesting cognitive bias.

8.3 Sexual overperception by men

Our final example pertains to the domain of sexual inference discussed earlier. One claim is that natural selection has favored sexual overperception in men (Haselton and Buss, 2000; Haselton and Nettle, 2006). As we have indicated, this claim is underdetermined by evidence that men exhibit interesting behavioral biases in this domain. If wasted opportunities are costlier than rejections, as seems plausible, then expected fitness maximizers with Bayesian beliefs will behave so as to avoid wasted opportunities, and they will do so without any sort of interesting cognitive bias. Evidence in the domain of sexual overperception, however, goes beyond simply demonstrating an interesting behavioral bias. A number of studies do, in fact, suggest interesting cognitive biases. For example, Abbey (1982) had unacquainted men and women interact briefly in pairs. Hidden observers, both male and female, observed the interactions and rated the sexual intent of both parties. Male observers perceived greater sexual intent in the target women than did female observers. Other studies have documented similar effects, including interesting studies by Haselton (2003) and Haselton and Buss (2000). In these studies, given that men and women have the same data, so to speak, it would appear that they cannot both be cognitively unbiased (in the interesting sense).

However, consider one final caveat. It is possible that people are born with evolved domain-specific priors, priors that, in effect, encode the evolutionary history of humans. In principle, men and women could be congenitally equipped with the same prior beliefs about the sexual interest of a given woman. Men and women,

however, are socialized in systematically different ways. As a result they will have different evidence available to them as they proceed through life, and they may enter an experimental scenario with different prior beliefs about female sexual interest – beliefs that may be justifiable in each case, given the evidence they have, respectively, encountered. We make no claims about whether this is true or about what form differential socialization would take (see Haselton and Buss, 2000, who consider a related point and offer convincing evidence against it). We simply want to point out that different posterior beliefs in men and women could, in principle, reflect different prior beliefs that are justified in each case. If the experiment were continued for long enough, with male and female observers exposed to the same evidence, it is possible that their respective posteriors would converge through time and be empirically indistinguishable.

9 Discussion and conclusion

Error management theory is a theory of emerging influence and broad application. The theory claims that cognitive biases do not necessarily reflect flaws in evolutionary design, but that they may be best conceived as design features. Unfortunately, existing accounts of the theory are vague regarding the key concept of bias. The result is that it is unclear that the cognitive biases that the theory seeks to defend are not simply a form of behavioral bias, in which case the theory reduces to a version of expected utility theory. We have offered some refinements and clarifications in an effort to highlight what we see as error management theory's genuinely novel claim: that behavioral tendencies to avoid costly errors can rest on systematic departures from Bayesian beliefs, and that the latter can be adaptive insofar as they generate the former.

We have outlined a taxonomy of biases, comprising trivial and interesting conceptions of both behavioral bias and cognitive bias. We have shown that inferences about cognition are radically underdetermined by evidence of interesting behavioral

bias. On the one hand, evidence of a behavioral tendency to avoid costly errors is not, in itself, evidence of interesting cognitive bias. Individuals with perfectly Bayesian beliefs will display such tendencies to the extent that they maximize expected utility under appropriate incentives. On the other hand, we have noted that even inferences about trivial cognitive biases are underdetermined by evidence of behavioral bias. We have identified three different trivial cognitive biases that might result in behavioral bias – trivially biased beliefs about states of the world will do the trick, as will trivially biased beliefs about the existence and/or magnitude of cost asymmetries. Finally, we have highlighted a key assumption, that the capacity for Bayesian beliefs is subject to constraints. To date, published accounts of error management theory have not mentioned this assumption, but the claim that systematic departures from Bayesian beliefs can be adaptive insofar as they generate adaptive interesting behavioral biases requires it.

Acknowledgements

The first author was supported by a research fellowship as part of a large collaborative project coordinated from the Centre for Anthropology and Mind (http://www.cam.ox.ac.uk) at the University of Oxford and funded by the European Commission's Sixth Framework Programme ("Explaining Religion"). The second author was supported by the Swiss National Science Foundation, grant No. 105312-114107. Thanks to Daniel Fessler, Nick Chater, David Hugh-Jones, Christian Ruff and two anonymous reviewers for valuable input. Special thanks to Martie Haselton and Daniel Nettle for constructive and generous comments.

References

Abbey, A. (1982). Sex differences in attributions for friendly behavior: Do males misperceive females friendliness? *Journal of Personality and Social Psychology*,

42, 830–838.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, **44**(3), 211–233.

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton: Princeton University Press.

Cosmides, L. and Tooby, J. (1997). Evolutionary psychology: A primer. http://www.psych.ucsb.edu/research/cep/primer.html.

Harsanyi, J. C. and Selten, R. (1988). *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.

Haselton, M. G. (2003). The sexual overperception bias: Evidence of a systematic bias in men from a survey of naturally occurring events. *Journal of Research in Personality*, **37**(1), 34–47.

Haselton, M. G. (2007). Error Management Theory. In R. F. Baumeister and K. D. Vohs, editors, *Encyclopedia of social psychology*, volume 1, pages 311–312. Thousand Oaks, CA: Sage.

Haselton, M. G. and Buss, D. M. (2000). Error management theory: A new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology*, **78**(1), 81–91.

Haselton, M. G. and Buss, D. M. (2003). Biases in social judgment: Design flaws or design features? In K. D. W. J. P. Forgas and W. von Hippel, editors, *Responding to the social world: Implicit and explicit processes in social judgments and decisions*, pages 23–43. New York: Cambridge University Press.

Haselton, M. G. and Nettle, D. (2006). The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review*, **10**(1), 47–66.

Haselton, M. G., Bryant, G. A., Wilke, A., Frederick, D. A., Galperin, A., Frankenhuis, W. E., and Moore, T. (2009). Adaptive rationality: An evolutionary perspective on cognitive bias. *Social Cognition*, **27**(4), 732–762.

Kahneman, D. and Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, **80**(4), 237–251.

Nesse, R. M. (2001). The smoke detector principle: Natural selection and the regulation of defenses. *Annals of the New York Academy of Sciences*, **935**, 75–85.

Neuhoff, J. G. (2001). An adaptive bias in the perception of looming auditory motion. *Ecological Psychology*, **13**, 87–110.

Press, W. H. (2009). Strong profiling is not mathematically optimal for discovering rare malfeasors. *Proceedings of the National Academy of Sciences*, **106**(6), 1716–1719.

von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Wilke, A. and Barrett, H. C. (2009). The hot hand phenomenon as a cognitive adaptation to clumped resources. *Evolution and Human Behavior*, **30**(3), 161–169.

Table 1: The payoff matrix for the basic binary model. By assumption, a, b, c, $d > 0$, and $a + d > b + c$.

	$X = 0$	$X = 1$
$C = 0$	b	$-d$
$C = 1$	$-c$	a

Table 2: The payoff matrices for the binary model with two alternative payoff structures. The matrix on the left corresponds to $M = 0$, while the matrix on the right corresponds to $M = 1$.

	$X = 0$	$X = 1$
$C = 0$	b	$-c$
$C = 1$	$-c$	b

	$X = 0$	$X = 1$
$C = 0$	b	$-d$
$C = 1$	$-c$	a

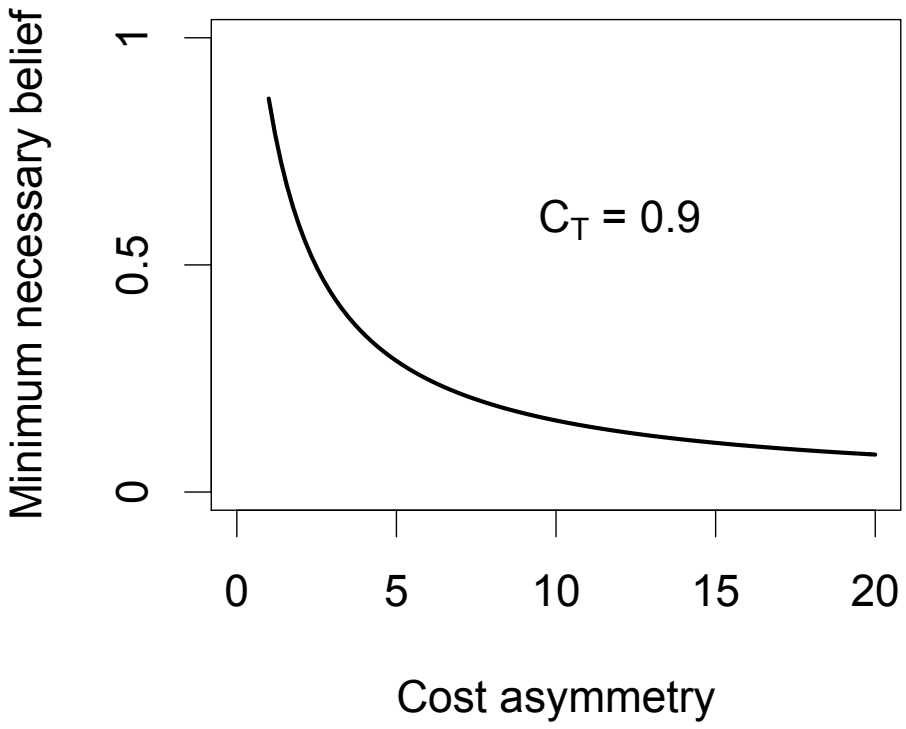

Figure 1: The minimum belief that $X = 1$ required to choose $C = 1$ with a probability greater than $C_T = 0.9$ shown as a function of z, the cost asymmetry. The graph shows this minimum required belief for values of z from 1 to 20 when $\lambda(b+c) = 3$.