The Paranoid Optimist:

An Integrative Evolutionary Model of Cognitive Biases

Martie G. Haselton
University of California, Los Angeles,
Communication Studies and Department of Psychology

&

Daniel Nettle

Psychology, Brain and Behaviour,
Henry Wellcome Building,
University of Newcastle,
Framlington Place
Newcastle,
NE2 4HH, UK

April 8, 2005

Author Note: We are grateful to Clark Barrett, Daniel Fessler, Garth Fletcher, Joachim Krueger, Mark Schaller and one anonymous reviewer for helpful comments on an earlier draft of this article, and to Karthik Panchanathan and Randolph Nesse for providing useful background for the mathematical modeling. We thank David Buss for bringing the auditory looming effect to our attention and for insights on the fundamental attribution error and disease avoidance phenomena.

Abstract

Human cognition is often biased, from judgments of the time of impact of approaching objects all the way through to estimations of social outcomes in the future. We propose these effects and a host of others may all be understood from an evolutionary psychological perspective. In this paper we elaborate *error management theory* (Haselton & Buss, 2000). 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. This perspective integrates a diverse array of effects under a single explanatory umbrella and it yields new content-specific predictions.

The Paranoid Optimist:

An Integrative Evolutionary Model of Cognitive Biases

Better safe than sorry. (Folk Wisdom)

Nothing ventured, nothing gained. (Contradictory Folk Wisdom)

These two wisdoms seem contradictory. The first urges caution, whereas the second reminds us that we have nothing to lose and should throw caution to the wind. Yet both seem to capture aspects of human psychology. A person following both maxims would be a paranoid optimist, taking chances in some domains while simultaneously being very fearful of certain kinds of harm. We will argue, using insights from signal detection and error management theory, that there are good evolutionary reasons why the paranoid optimist mind could evolve. Furthermore, in which domains it is best to be paranoid and in which to be optimistic is predictable from the pattern of recurrent costs and benefits associated with decisions in that domain throughout our evolutionary history. This perspective suggests that one of the curiosities of human cognition—the fact that it seems riddled with biases—may be a functional feature of mechanisms for making judgments and decisions.

Human cognition has often been shown to be biased. Perceivers underestimate the time-to-impact of approaching sounds (Neuhoff, 1998, 2001), and overestimate the connection between pictures of snakes and unpleasant outcomes like electric shocks (Tomarken, Mineka & Cook, 1989). People also appear to have a variety of *positive illusions* (Taylor & Brown, 1988), which cause them to overestimate the likelihood that they will succeed in spite of the adversity they face. Evidence in these domains and many others suggest that humans possess a multitude of biases, or propensities to adopt one belief on the basis of more slender evidence than would be required to believe in an alternative.

Until recently, many psychologists have been content to describe these phenomena, their contexts of appearance, and possible implications, without much concern for their ultimate origin. As Krebs and Denton (1997) note, in as much as explanation is needed, it tends to be proximate in nature. Psychologists argue that cognition is performed by a set of simple heuristic procedures, which are effective in many circumstances but prone to error in others (e.g., Miller & Ross, 1975; Kahneman et al., 1982). Or, in social domains, biases in judgment serve the proximate function of preserving self-esteem or subjective happiness for the ego-centered human animal (Crocker & Park, 2004; Greenberg, Pyszczynski, Solomon & Pinel, 1993; Kunda, 1990). Researchers offer evoked biases as examples of just such imperfections.

A noteworthy exception exists in the domain of sexual inference. Haselton and Buss (2000) argued that the documented tendency for men to overestimate women's sexual intent could be an adaptive bias designed by natural selection. 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. (We discuss this example further below.) A second example occurs in the perceptual domain. Neuhoff (2001) argued that the perceptual bias towards thinking that incoming sources of sound will arrive sooner than they actually do may be adaptive, because it is better to be ready too early for an incoming object, than too late.

Here we extend the insight that biased systems can result in higher fitness relative to unbiased ones, and demonstrate that a wide variety of biases, both positive (optimistic) and negative (paranoid), may be brought under a single explanatory umbrella. We elaborate *error*

management theory (Haselton & Buss, 2000) by presenting a mathematical derivation of the model, broadening its potential domains of application, and presenting new predictions. We argue that a key parameter explaining the direction of biases is the relative effects on fitness of the different types of error.

This effort toward integration is useful because it provides clues about the circumstances in which reasoning in a biased way may have yielded fitness advantages in the ancestral past, thus providing guidance about where we should expect to find the classic biases, such as the fundamental attribution error, and their exceptions, and where as yet undiscovered biases may be found. Equally important, this perspective speaks to the ongoing debate about human rationality by demonstrating that biased reasoning need not be deemed a *design flaw* of human cognition; instead it may often be a *design feature*.

Error Management and Adaptive Bias

Noise and Uncertainty

The world of the perceiver is filled with uncertainty. In social inference, a judge must overcome the fact that a target's behavior is determined by multiple factors, many of which interact in complex ways to produce behavioral outcomes. Moreover, if the perceiver and target are engaged in strategic interaction, marked by competing interests, important social clues may be concealed or a target might stage interference by engaging in active deception. Social judgment and inference may also concern events that are not directly observable because they occurred in the past or might happen in the future.

These difficulties are not limited to social domains. The fact that in complex environments perception is always clouded by the presence of confounding noise was central to the development of signal detection theory in psychophysics (Green & Swets, 1966). All forms

of judgment under uncertainty will be prone to errors. Given the necessary existence of these errors, how should systems best be designed? Error management theory provides a potential answer.

Error Management Theory

Error management theory (Haselton & Buss, 2000) applies the principles of signal detection theory (Green & Swets, 1966; Swets, Dawes, & Monohan, 2000) to understanding how natural selection engineers psychological adaptations for judgment under uncertainty. 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 can not be adopted and not be true (a true negative or TN). Then there are two possible errors. A false positive error (FP) occurs when the subject adopts a belief that is not in fact true, and a false negative (FN) occurs when the subject fails to adopt a belief that is true. The same framework applies to actions. A false positive occurs when the subject does something though it doesn't produce the anticipated benefit, and a false negative when the subject fails to do something that, if done, would have provided a benefit.

The costs of the different outcomes, and in particular the two types of error, are rarely identical. In testing hypotheses, type I errors (false positives) are typically considered more costly by the scientific community than are type II errors (false negatives). Thus, scientists bias their decision-making systems (e.g., classical inferential statistics) toward making type II errors, because reducing type I errors necessarily increases type II errors. The reverse asymmetry characterizes hazard detection. Misses (false negatives) are often much more costly than false alarms (false positives). This asymmetry holds for humanly engineered devices such as smoke

detectors and for evolved hazard detectors such as anxiety, stress, and cough (Nesse, 2001), and hazard-detection systems are often biased toward false alarms.

Whenever the costs of errors are asymmetrical, humanly engineered systems should be biased toward making the less costly error. This bias sometimes increases overall error *rates*, but, by minimizing the more costly error, it minimizes overall *cost* (Green & Swets, 1966; Swets, Dawes, & Monohan, 2000). According to error management theory, 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.

Because the human environment is often very uncertain, and the costs of the two types of errors are likely to be recurrently asymmetric in most fitness-relevant domains, EMT predicts that human psychology contains evolved decision rules biased toward committing one type of error over another. In the following sections, we demonstrate how this model can account for a large number of biases that have been observed empirically. First, though, we derive the central claims of EMT formally, using a simple model based on signal detection theory (Green and Swets 1966). Consider the situation where the subject might or might not form some belief (for example, that there is a snake in the grass, or that a member of the opposite sex is sexually interested in him). The approach can be extended to cover judgements on a quantitative scale, giving essentially the same results, but here we consider only cases where there is a dichotomous choice to form a belief or not do so. The belief in question need not be a conscious one. By 'adopting a belief' is meant behaving or reasoning as if the corresponding proposition were true.

Let us call the state of the world that may or may not obtain lower case s, whereas the belief that the subject may form is capital S. That is to say, a subject with belief S believes that the world is in state s, which may or may not really be the case. As detailed above, there are four possible outcomes in such a situation. A true positive would be where the belief was formed and was in fact true (that is, S and s). A true negative would be where the belief was not formed, and was not true (¬S and ¬s). A false positive would be where the belief was erroneously formed (S and ¬s), whereas a false negative would represent the failure to form a belief which is in fact true (¬S and s).

The signal detection problem is the problem of how much evidence for the state s to require before adopting the belief S. For every degree of evidence e, it is possible to specify the probability of that evidence being observed if s, or p(e|s), and also the probability of that evidence being observed if ¬s, or p(e|¬s). If both p(e|s) and p(e|¬s) are non-zero, then there is some uncertainty in the world. That is, the observed evidence could have been generated if the world is in state s or in state ¬s . If this uncertainty is not present, then the signal detection problem is trivial and there is no scope for the evolution of bias according to our formulation.

Intuitively, it would seem that the subject should form the belief S if p(e|s) is greater than p(e|¬s). This is indeed an optimal rule if the a priori probabilities of s and ¬s are equal, and the organism's goal is to maximize the number of true beliefs (Green and Swets 1966, p. 23). The ratio p(e|s)/ p(e|¬s) is called the likelihood ratio. As the likelihood ratio increases, the relative probability that s is in fact the case given the observed evidence e increases. An unbiased decision would mean adopting S wherever the likelihood ratio is greater than 1. Any other threshold is a bias; a bias against S if it is greater than 1, and a bias towards S if it is less than 1.

From an evolutionary perspective, a decision rule is optimal not if it maximizes the number of true beliefs, but if it has the best possible effect on the organism's fitness. Let us assume that the four possible outcomes have different effects on fitness. vTP is the effect on fitness of believing S when s is in fact the case; vTN is the effect of believing ¬S when ¬s is in fact the case. vFP is the payoff for a baseless belief S. vFN is the effect of believing ¬S when s is actually the case.

The expected value of any decision is given by the following expression:

$$EV = p(s)\{\mathrm{p(S \mid s)vTP} + \mathrm{p(\neg S \mid s)vFN}\ \} + p(\neg s)\{\mathrm{p(\neg S \mid \neg s)vTN} + \mathrm{p(S \mid \neg s)vFP}\} \qquad (1)$$

The burden of expression (1) is simply that the expected value is the sum of the probability of a true positive times the payoff for a true positive, the probability of a true negative times the payoff for a true negative, the probability of a false positive times the payoff for a false positive, and the probability of a false negative times the payoff for a false negative. The optimal decision rule would be one that maximized expression (1). It can be shown that (1) is maximized by adopting the belief S wherever there degree of evidence is equal to e, where:

$$\frac{p(e \mid \mathrm{s})}{p(e \mid \neg \mathrm{s})} = \frac{p(\neg s)}{p(s)} \bullet \frac{(vTN + vFP)}{(vTP + vFN)} \qquad (2)$$

For economy, we do not provide a derivation of this expression in this paper, but is given in full in Green and Swets (1966, pp. 21-23). The left-hand term is the likelihood ratio of s given the evidence e, and the right-hand term is made of up of the relative frequencies of s and ¬s, and the payoffs for the four possible outcomes. Equation (2) has the satisfying property that if s and

¬s are equally likely a priori and the payoffs for all the possible states are equal, then the subject should believe S if p(e|s) is greater than p(e|¬s), as intuition would predict.

The type of case central to this paper is that in which the payoffs for the different outcomes are not all equal. Though there are four payoffs to consider, that can in principle vary independently, the situation can made conceptually clearer by holding vTP and vTN constant and equal, and defining vFP and vFN as the payoff of the deviation from the optimal outcome, including opportunity costs, caused by the two types of error. As long as the errors are measured in relation to the value of the true outcomes, and all payoffs are expressed in the same currency, no information is lost by this, and it means that only the costs of the two error terms need be considered in making predictions. For example, in the female investment-detection example used below, the vTP is substantial (value of male investment), but this may be conceptualized in terms of vFN as an opportunity cost (cost of missing out on male investment); thus the model can hold vTP and vTN constant and vary only vFP and vTN without loss of information.

First, consider the case where the cost of a false positive is rather small, and that of a false negative is rather large. This would be the situation, for example, for an animal detecting a snake. The payoff for a false positive would be the wasted energy of moving away when in fact there was no danger – say 1 unit. The payoff for a false negative would be allowing a potentially venomous snake too close. This negative effect of this could be very large – say 1 to 50 units. There is a certain amount of sensory evidence available, but that evidence is uncertain. For example, a stick has many of the properties of a snake. The question is how much evidence to require that the perceived object must belong to the class of snakes not the class of sticks before assuming that it is a snake.

Assume for convenience that p(s) = 0.1 and p(¬s) = 0.9 – that is, there are nine times as many sticks as snakes in the world. This assumption is arbitrary, but it affects only the scaling and not the shape of the relationships to be shown, since p(s)/p(¬s) is always a constant. Figure 1 plots the optimal point at which the subject should adopt belief S under these conditions, with vFN varying from -1 to -50. When vFN is very small, the organism should require a large likelihood ratio to adopt S, because if it does so it incurs the cost of moving away. In fact, when snakes are rare and not very dangerous, there is a bias against detecting them, since the optimum threshold is greater than 1. However, as vFN increases in magnitude, the optimal threshold for adopting S very rapidly declines. At vFN=-10, the optimal point is at 0.82, which is a bias towards detecting snakes. At vFN= -50, under optimal behaviour, the individual should adopt S even though ¬s is over 5 times more likely than s given the evidence. Total costs to fitness are still minimized, because the rare false negative is so much more damaging than even multiple false positives. It is far better to see a snake where there is only a stick than vice versa.

This first example illustrates the 'smoke detector principle' (Nesse, 2001, 2005)—if the cost of failing to detect something is relatively high, it is best to have a lot of false alarms if it means catching the real event when it does happen. Let us also consider another possibility. Imagine a female trying to detect whether a male is willing to make a significant post-reproductive investment if she mates with him. The value of this investment is positive, and a false negative involves missing out on it, so the opportunity cost vFN is significant, let us say -5. However, the value of the false positive is potentially higher (very costly), since if she mates and then is deserted she faces the possibility of raising an offspring alone, and may have trouble finding another partner in future. Thus vFP varies from -1 to -50. We assume 40% of all men are deserters. Again this is an arbitrary assumption that affects only the scaling.

Figure 2 shows the optimum threshold as vFP varies. If the cost of being deserted is low, the female should have a bias towards accepting the evidence for male commitment. However, as the cost of desertion increases, the optimum threshold soon exceeds 1, which means that she should not adopt S even when the available evidence is more likely to be have been generated by s than ¬s. If vFP = -30, then she should not accept S unless the objective likelihood of s is more than four times that of ¬s given the evidence she has been able to observe. Thus the model predicts the phenomenon of 'commitment skepticism', which has been empirically documented by Haselton and Buss (2000). (We discuss this example and the research evidence for it in more detail below.)

The model clearly shows that the optimal decision rule is based not on the objective likelihoods alone, but also on the payoffs of the different outcomes. EMT, as its name suggests, deals specifically with the relative costs of the two errors, FP and FN, but this does not restrict the conclusions that can be drawn, since the payoff of a veridical outcome can always be restated as an opportunity cost of the converse error, and vice versa. The optima generated by a signal detection model will be influenced by the a priori probabilities of the two states of the world, and also by the how well the evidence discriminates between the two states (that is, the distribution of the likelihood ratio). However, we do not explore those dynamics here (see Nesse 2001, 2005, for some further exploration of these models in the context of the 'smoke detector principle'; also see Swets, Dawes, & Monahan, 2000, for applications in diagnostic domains). Our central result is robust to permutations of these other parameters: where the relative costs of the two errors are asymmetric, the optimal thresholds are biased away from 1 and towards the less costly error.

Applications of EMT

We review three somewhat overlapping classes of biases. Our classification system is intended to provide a heuristic organizing scheme, rather than an exhaustive, mutually exclusive taxonomy. We argue that each case is an example of error management. The two possible errors are plausibly asymmetrical in cost, and in each case, the bias is towards making the less costly error. In each case, decision makers also face a significant degree of uncertainty about the true probability of an event.

Protective Effects in Perception, Attention, and Learning

Few failures are as unforgiving as failure to avoid a predator.

(Lima & Dill, 1990)

Auditory Looming. Neuhoff (1998, 2001) shows that there are biases in the perception of sounds that are rising and falling in intensity. Rising intensity is usually a cue that the source of the sound is approaching the listener. In a series of psychoacoustical experiments involving speakers moving on cables, Neuhoff and colleagues demonstrate that sounds rising in intensity are perceived as approaching faster than matched sounds that are falling in intensity (see Neuhoff, 2001, for a review). Moreover, they are judged to be closer than equidistant falling sounds. Neuhoff proposes an adaptive explanation; when a source is approaching, it is better to be prepared for it too early than too late, and so selection would favor neural mechanisms that detect approaching sounds in manner asymmetric to receding ones. This explanation is compatible with the error management model. Natural environments are filled with competing sources of sound that render auditory judgments susceptible to error. For approaching sounds, the relatively inexpensive false positive error would be to take preparatory action for an arriving sound source too early. The false negative would be to take such action too late, which could

well lead to such costly outcomes as being struck by a projectile, predator, or assailant. Thus, the optimal system is biased toward false positive errors. This is the familiar principle of the smoke detector: it is better to tune a smoke detector to always detect a genuine fire, even if the cost is the occasional false alarm (Nesse, 2001, 2005; also see Bouskila & Blumstein, 1992). We will argue that a whole host of biases fall into this same, self-protective smoke detector class (table 1).

Allergy, Cough, and Anxiety. Nesse (2001, 2005) argued for the 'smoke detector principle' in bodily systems designed to protect from harm. Nesse describes medical examples such as allergy and cough where a protective system is often mobilized in the absence of real threat. These defense systems appear to be over-responsive; dampening them with drugs or treatment actually results in few untoward effects on the subject (Nesse, 2001). Psychological defense mechanisms such as anxiety are also easily evoked, especially in connection with things likely to have been dangerous in the ancestral environment, such as spiders, snakes, and potentially dangerous persons, as we discuss below (Mineka, 1992; Seligman, 1971; Tomarken, et al., 1989). A tendency for anxiety mechanisms to produce false positives is a plausible explanation for the observed prevalence of phobias and anxiety disorders (Nesse, 2001).

Dangerous Animals. It has long been argued that humans are phylogenetically prepared to produce a fear response to snakes and spiders (Seligman, 1971). More recent evidence suggests not only a special sensitivity to acquire fears of these ancestrally dangerous animals, but also biases that serve to elicit fear, maintain it, and express it more often than it is needed. Mineka and colleagues have demonstrated that snake fear responses are more easily acquired and more difficult to extinguish than fears of other fear-relevant stimuli (see Mineka, 1992, for a review). Even when extinction is successful, it tends to be short-lived, as the fears are easily

reacquired (Mineka, 1992). In experiments, people overestimate the covariation between electric shocks and images of snakes and spiders but do not overestimate the covariation between shock and images of flowers or mushrooms, or even with images of damaged electrical outlets (de Jong & Merckelbach, 1991; Tomarken, Sutton, & Mineka, 1995; Tomarken et al., 1989). The covariation bias effect appears to be strongest in people with specific animal fears (Tomarken et al., 1995), but when the fear-relevant stimulus (e.g., snake photos) are raised in frequency in an experiment, low-fear individuals also exhibit the covariation bias effect (Tomarken et al., 1989, also see deJong & Merckelbach, 1991). Once a negative association with snakes and spiders is established in a person's mind, the fear response can be evoked by a much briefer presentation of the feared image than is required for pictures of other stimuli (Oehman & Soares, 1993).

Thus, there appear to be biases in expressing fears of snakes and spiders, and the specialized sensitivity that facilitates the acquisition of these fears may also be conceived of as a bias—snake and spider fears are acquired on the basis of slimmer evidence than are fears of other dangerous objects, even those that in contemporary terms are much more dangerous, such as electrical outlets, guns, and automobiles. In ancestral environments, the over-expression of fears of snakes and spiders was inconvenient but not overly costly, whereas failing to fear truly dangerous animals would have been extremely costly, given the presence of severely venomous snakes and spiders in tropical regions. Bouskila and Blumstein (1992) develop similar expectations about estimations of predation hazard in non-human animals.

Dangerous Persons. It is quite possible that the greatest threat to life in ancestral environments was other people. In modern environments, from traditional societies to industrialized nations, groups regularly wage deadly wars on one another (Keeley, 1996), young adult men who are at the peak stage of intrasexual conflict commit a disproportionate number of

murders, and competing reproductive interests result in spousal homicide (Daly & Wilson, 1988). Thus, a parallel analysis to that advanced for dangerous animals applies to dangerous persons. There is evidence that cues of interpersonal threats also tend to be processed in a biased fashion. For example, Fox, Russo, and Dutton (2002) showed that angry faces capture attention for longer than happy or neutral faces, even when participants are trying to ignore them. Similarly, Pratto and John (1991) found that words describing undesirable traits capture attention for longer and cause more task interference than words describing neutral or positive traits. In practice, the extremely undesirable traits are things that evoke interpersonal threat or violence, such as *hostile, mean* and *sadistic*. Thus, these effects may result from the operation of a threat detection system that is predisposed to bias attention toward ancestrally dangerous stimuli.

In ancestral environments, between-group differences in appearance and behavior, such as tribal markers, signaled differences in coalition membership; in modern environments, racial and ethic cues appear to activate the psychology of inter-group conflict (Kurzban, Tooby, & Cosmides, 2001; Sidanius & Veniegas, 2001). The assumption that members of one's own racial or ethic group are more generous and kind (Brewer, 1979), and less hostile and violent than out-group members (e.g., Quillian & Pager, 2001), is a bias that can be understood from an error management perspective. Inferences about relatively unknown out-group members are uncertain. For ancestral humans, the costly false negative was to miss aggressive intentions on the part of others, whereas the false positive of over-inferring aggressiveness was low, especially for members of competing coalitions. This asymmetry did not characterize inferences about in-group members, in which costly within-coalition conflict would have resulted from unwarranted inferences of hostility or aggressiveness. Consistent with this analysis, ambient darkness—a cue signaling increased risk of hostility from others—increases racial and ethic stereotypes connoting

violence, but has little effect on other negative stereotypes (e.g., laziness or ignorance) (Schaller, Park & Faulkner, 2003; Schaller, Park & Mueller, 2003).

Food Aversions. A single instance of gasterointestinal malaise following ingestion of a particular food is sufficient to induce a strong, long-lasting, avoidance of that food (Garcia, Ervin & Koelling, 1976; Rozin & Kalat, 1971). These aversions are likely the product of specialized associative biases designed to help organisms avoid ingesting toxins, even at the cost of a lost source of calories. As with snake fears, taste aversions are long-lived and hard to extinguish. Only taste and smell cues are effective at creating an aversion (auditory or visual cues are generally ineffective, Rozin & Kalat, 1971). And, in contrast to other conditioned associations, creating the aversion requires only one trial and the delay between ingestion and malaise can be quite prolonged (Garcia et al., 1966). These associative biases characterize omnivorous animals for whom they would be most beneficial; the ability to form conditioned taste aversions is lost in a species that relies on only one food that is always fresh as it is drunk straight from a live host: vampire bats (Ratcliffe, Fenton, & Galef, 2003). The ease with which food aversions are acquired and maintained, given relatively slim evidence of their toxicity, results in many false alarms—avoiding foods which are in fact safe.

Within the EMT framework, the false positive is the formation of a taste aversion to a food that is normally harmless. This has a non-trivial cost, since it may mean missing out for an entire lifetime on an available source of nutrition. On the other hand, this cost is low compared to the cost of eating a potentially fatal toxin or pathogen (a mistake one can make only once!), so the system is biased towards self-protection rather than calorific maximization.

Several other food choice phenomena are illustrative. Children, who are less able than adults to detoxify poisonous plant parts, tend to avoid leaves and vegetables and are notoriously

picky about what they eat (Cashdan, 1998). Pregnant women, whose immune system is suppressed in order to avoid attacking the fetus, develop a variety of pregnancy-specific food aversions (Fessler, 2002). As clever experiments by Rozin and colleagues demonstrate, even the mere suggestion that a food might be contaminated is sufficient to elicit avoidance or disgust. When given a choice between two containers of sugar, people opt for the container labeled "table sugar" over the one marked "NOT sodium cyanide" even though they had just watched the experimenter fill both bottles from the same box of Domino sugar (Rozin, Markwith, & Ross, 1990). Likewise, people refuse to eat otherwise tasty food products that are presented in the shape of a disgusting substance, like fudge in the shape of dog feces (Rozin & Fallon, 1987).

Avoiding the Ill. People may require little evidence of illness or contamination to avoid someone, whereas much stronger evidence is required warrant the inference that someone is safe or free from disease (Kurzban & Leary, 2001; Park, Faulkner, & Schaller, 2003). For example, although people understand that mere contact is insufficient for the transmission of AIDS, they physically distance themselves from AIDS victims, demonstrate *dose insensitivity* by expressing discomfort with even 5 minutes of contact, and exhibit *backward contagion* as evidenced by discomfort with the thought that an item of clothing they once wore would be worn by an AIDS victim in the future (Rozin, Markwith, & Numeroff, 1992; Bishop, Alva, Cantu, & Rittiman, 1991). As we discuss below in the implications of EMT section, disease avoidance may be broadly over-inclusive and people may also treat other disabilities or pheonotypic anomalies (e.g., obesity) as if they are produced by contagious disease. The error management interpretation of these phenomena is that the costs of false negatives (failing to avoid someone with a contagious disease) is high, whereas the cost of a false positive is relatively low (avoiding contact with a non-contagious person), so disease avoidance mechanisms will be over-inclusive

and will express many false alarms. This may account for the difficulty in reversing stigmas associated with both contagious and non-contagious physical afflictions (Bishop et al., 1991) as compared with more easily manipulated social stigmas (such as those surrounding homosexuality, Kurzban & Leary, 2001). This form of defense over-responsiveness might also explain the seemingly irrational local panic associated with outbreaks of SARS and Mad Cow disease in far away places.

Biases in Interpersonal Perception

Interpersonal perception is notoriously prone to bias and error. We propose that many of these documented biases can be interpreted within the framework of error management theory (table 2).

The Illusion of Animacy. Guthrie (2001) uses error management logic to explain one of the key features of religion—animism. He proposes that in ambiguous circumstances to falsely assume that an intentional agent (e.g., another human) has caused some event is less costly than to miss this fact. Given that agents often have interests that compete with those of the perceiver, it is important to have a low threshold for inferring their presence. For example, if one encountered a collection of twigs arranged in an improbably neat array, Guthrie proposes that it would be better to entertain the thought that a human or other intentional agent was responsible for the arrangement, and to increase one's vigilance to the possibility of the agent's presence, than to casually ignore it. Guthrie (2001) and Atran and Norenzayan (in press) propose that belief in gods may be a by-product of this adaptive bias. The proposed animacy bias is consistent with classic laboratory experiments conducted by Heider & Simmel (1944; see also Bloom and Veres, 1999). When participants view moving images of circles and squares, they find it difficult not to infer intentional states—chasing, wanting, and escaping. The tendency to

infer intentional states in these stimulus arrays emerges early (age 4), and there is preliminary evidence of cross-cultural universality of the bias (in Germans and Amazonian Indians, Barrett, Todd, Miller, & Blythe, in press), though its magnitude of expression may certainly be variable. Common features of religion across cultures (Atran & Norenzayan, in press) are also consistent with a universal animacy bias.

The Sinister Attribution Error, Overweighting of Social Gaffes, and Negative Forgiveness Bias. The sinister attribution error is ego's assumption that relatively trivial aspects of another's behavior indicate negative thoughts or intentions towards ego (Kramer, 1994, 1998). EMT would predict that such a bias could arise where the costs of failing to detect negative evaluations that in fact do exist are higher than the costs of inferring such evaluations where there are none in reality. Kramer has shown that the sinister attribution error and paranoid cognition are exhibited differentially by people under intense scrutiny, new to social groups, or low in status within an organization (see Kramer, 1998 for a review; also see Fenigstein, 1984).

In one sinister attribution study, first and second year students in a masters program at a prestigious business school were asked how they would interpret ambiguous interactions with their fellow students. They were asked, for example, what they would infer if they made an urgent phone call the evening before an exam that their fellow student did not return, or if they were telling a joke they thought was funny and one of their fellow students abruptly rose and left the table. First year students were more likely than second year students to interpret the interactions in a "personalistic" fashion by inferring that their call was not returned because the recipient did not wish to speak to them or that the person found their joke boring (rather than inferring, for example, that their phone message was never received). This effect was amplified when first year students imagined that the interaction took place with a second year student,

whereas second year students did not make differing attributions depending on the status of the person they imagined interacting with. In a second study with the business students, Kramer investigated whether participants in an economic coordination game believed their fellow participants were trying to sabotage them in order to earn more money. Those who believed that their reactions in the game revealed managerial skill and that they were being videotaped attributed a greater desire for sabotage to their fellow students than those who did not believe they were under scrutiny (Kramer, 1994).

Savitsky, Epley, and Gillovich (2001) documented related effects. Participants committed an experimentally-induced social gaffe—failing at a "simple" anagram test or being described in an embarrassing manner. In four studies, the participants believed that they were judged as less intelligent and less favorable in their general impression by strangers than they actually were.

In sum, when individuals are new to social groups or feel that they are under scrutiny, they become hypervigilant to the negative thoughts, intentions, or evaluations of others. These situations resemble ancestral environments where failing to detect negative social evaluations was highly costly, such as when entering into a new coalition or moving into a new village. Failing to detect negative intentions or evaluations could result in ostracism or direct aggression, and the consequences could literally have been deadly (Baumeister & Leary, 1995).

In the context of romantic relationships, Friedman, Fletcher, & Overall (in press) found that men and women underestimated the degree to which their partners had forgiven them after a transgression (e.g., insults, flirtation with others). With transgression severity controlled, this bias was strongest in partnerships characterized by less relationship satisfaction. Thus, as the

researchers proposed, a negative forgiveness bias may help to ensure that transgressions are fully mended or not further exacerbated, especially in relationships that are already on the rocks.

The Fundamental Attribution Error. When interpreting the behavior of others, people are prone to making the fundamental attribution error (FAE), which is the assumption that a person's behavior corresponds to their underlying dispositions to a greater extent than is logically warranted (e.g., Andrews, 2001; Nisbett, Caputo, Legant, & Marecek, 1973; Ross, 1977). The extent to which this bias is expressed varies between collectivist and individualist cultures with members of collectivist cultures tending to qualify dispositional inferences by referencing the social context to a greater extent than members of individualist cultures (Choi, Nisbett, & Norenzayan, 1999). When situational and dispositional inferences are disentangled, members of both collectivist and individualist cultures tend to display dispositional inferences to the same degree (Norenzayan, Choi, & Nisbett, 2002). Kurzban & Leary (2001) argue that many of our initial social judgments are designed to help us avoid poor social exchange partners. This is important because humans depend on social exchange based on reciprocity to a very great extent, and reciprocity is always vulnerable to cheating. Thus, it is plausible to argue that avoiding social cheaters has been a major selective pressure on human social cognition (Cosmides, 1989).

One effect of the fundamental attribution error is to cause observers to avoid social exchange partners who have once demonstrated some negative social behavior, because it entails the assumption that the person is disposed to do the same again on repeat interaction. This aspect of the FAE can be interpreted from an error management perspective (see Andrews (2001) for this and other, complementary explanations of the various manifestations of the FAE). The false negative is to assume that a person's behavior is not representative of their long-term dispositions, and thus not take it into account in future interactions. The risks of the false

negative are becoming involved with a person who has social cheating tendencies. The false positive is assuming someone is anti-socially disposed because of a behavior, which did not in fact represent his or her underlying dispositions, but was brought about by a more transient feature of the context. The cost of such a false positive might be the avoidance of people who would in fact be appropriate social partners. This cost might be significant, but often not as high as the cost of being exploited.

A recent series of studies is consistent with this interpretation of the FAE. One of the original demonstrations of the FAE showed that people infer the existence of more personality traits in others than they do in themselves (Nisbett et al., 1973). Using similar methods, Burkett, Cosmides, and Kirkpatrick (2003) demonstrated that this manifestation of the FAE occurred for negative traits such as dishonest, mean, rude, and inconsiderate, but not for positive trait attributions such as honest, fair, kind, and intelligent. Of the negative traits investigated, traits related to social exchange (dishonesty, cheater, liar, and deceitful) were those for which there was the largest FAE bias. In a similar vein, using the lexical decision paradigm, Ybarra, Chan, and Park (2001) found that adults were faster to identify trait words connoting interpersonal social costs (e.g., hostile, cruel, disloyal) than words connoting poor skill (e.g., stupid, weak, clumsy), or positive qualities (e.g., honest, friendly, gentle). Ybarra (2002) concluded that people tend to lean toward seeing the bad in others in "morality" domains in order to protect themselves from poor social partners.

The Social Exchange Heuristic. Standard economic principles predict that players in the one-shot prisoner's dilemma game should defect rather than cooperate. If one partner cooperates while the other defects, the cooperator suffers a greater loss than if he or she had defected. The interaction is not repeated, so there is no incentive to signal cooperativeness, and experiments are

carefully devised so that there is no information about reputation that might serve to provide clues about the partner's cooperative disposition at the start of the game. Yet, cooperation often occurs in the one-shot prisoner's dilemma game and in many other games in experimental economics (Sally, 1995; Caporael, Dawes, Orbell, & van der Kragt, 1989; Camerer & Thaler 1995; Henrich et al., 2001).

Yamagishi and colleagues hypothesized that cooperation in one-shot games results from the operation of a *social exchange heuristic* (Yamagishi, Terai, Kiyonari & Kanazawa, 2003). They propose that the costs of falsely believing one can defect without negative social consequences are often higher than cooperating when one could safely defect. This asymmetry should hold when the costs of "unneeded" cooperation are relatively low (e.g., a low dollar amount is lost) or when the social costs of failing to cooperate (potential ostracism) are high. The costs of ostracism may be particularly high in interdependent social contexts in which cooperation is either highly valued or especially necessary. And, as predicted, in Japanese collectivist samples where exchanges are often closed to outsiders, cooperation in one-shot experiments is higher than in the more individualist United States samples (Yamagishi, Jin, & Kiyonari, 1999).

We suggest that this bias can be conceptualized as a combination of error management and an artifact of modern living, since in an ancestral environment the probability of repeated encounters would have been high and social reputation effects especially potent. Thus, people may be predisposed to expect negative consequences of non-prosocial behavior even when, objectively, such consequences are unlikely to follow. The bias towards prosociality is the subject of competing explanations which take quite different explanatory stances (Price, Cosmides & Tooby, 2002; Gintis et al., 2002; Henrich & Boyd, 2001; Bowles & Gintis, 2002),

and it is as yet unexplored whether these are complementary or competing accounts to the social exchange heuristic.

Sex-Differentiated Biases in Decoding Courtship Signals. To the degree that the problems of judgment and social inference differed for the men and women over evolutionary history, or were associated with different cost asymmetries for the sexes, EMT predicts that biases will be sex differentiated. Haselton & Buss (2000) hypothesized a number of sex-specific biases in interpersonal perception.

The reproductive success of males is ultimately limited by the number of females they can inseminate, whereas for females there is no fitness return on increasing numbers of mating partners beyond a certain point (indeed additional matings may become costly, Rice, 1996, 2000; Symons, 1979). Thus, for males there is a higher cost to missing out on a mating opportunity than there is for females. For females, becoming pregnant is highly costly, and fitness is affected by the continued investment of the male. Given these asymmetric costs and benefits, Haselton and Buss argued that men would have adaptive cognitive mechanisms designed to avoid missed mating opportunities, whereas women would have cognitive mechanisms designed to avoid post-reproductive desertion.

The error management predictions in this case are that men should tend to overestimate the sexual interest of women with whom they interact, since the false negative (missing a sexual possibility that was in fact real) is more costly than the false positive (inferring a sexual interest where there is none). A number of empirical studies demonstrate that men do indeed overestimate women's sexual interest. In laboratory studies, when male partners in previously unacquainted male-female dyads are asked to infer their partner's sexual interest, they consistently rate it as higher than the female partner's report suggests, and higher than the ratings

provided by female third-party viewers of the interaction (Abbey, 1982; Saal, Johnson, & Weber, 1989). A similar effect occurs in studies using photographic stimuli (Abbey & Melby, 1986; Maner et al., in press), videos (Johnson, Stockdale, & Saal, 1991), short vignettes (Abbey & Harnish, 1995), ratings of courtship behaviors (Haselton & Buss, 2000), and in surveys of naturally occurring misperception events (Haselton, 2003). Importantly, evidence of sexual overperception does not appear in women (Haselton, 2003; Haselton & Buss, 2000; Maner et al., in press).

For women, in considering the commitment intentions of a potential partner, the false negative would be to miss signs of a genuine desire to commit. The false positive, on the other hand, would be assumption of a willingness to commit where in fact there was little or none. A woman making this error could be forced to raise a child without the help of an investing father, which in extant traditional societies can more than double the risk of offspring death (Hurtado & Hill, 1992). This error could also reduce her future mating potential because it decreases her residual reproductive value (Buss, 1994; Symons, 1979). Thus, EMT predicts a bias in women towards underperception of men's commitment intentions. Laboratory studies confirm that in the courtship context women underestimate men's commitment. Women infer that potential indicators of men's desire for a committed relationship (e.g., verbal displays of commitment and resource investment) indicate less commitment than men report that they intend them to indicate (Haselton & Buss, 2000). The same result appears when comparing women's and men's ratings of a third-party man's dating behaviors, demonstrating that the effect is not attributable to a simple self-other rating difference which might result from participants' concerns about self presentation (Haselton & Buss, 2000). Importantly, evidence of commitment bias does not appear in men's assessments of women's behaviors (Haselton & Buss, 2000).

Self-Related Biases

Positive Illusions. Some of the best-known cognitive biases concern beliefs about the self and the future. People have been shown to have unrealistically positive views of the self, unwarranted optimism about the future, and to believe that they control the flow of events to a greater extent than is logically warranted. These effects were grouped together by Taylor and Brown in their seminal review (Taylor & Brown, 1988), and dubbed 'the positive illusions'. Since the time of the review, some debate has arisen about the pan-cultural status of the positive illusions. In particular, members of East Asian cultures such as the Japanese and Chinese have sometimes been found not to self-enhance, but rather to self-criticize (Heine, Lehman, Markus, & Kitayama, 1999; Kitayama, Markus, Matsumoto, & Norasakkunkit, 1997; Yik, Bond, & Paulhus, 1998).

On the other hand, other investigators have found that both American and Japanese participants self-enhance (Sedikides, Gaertner, & Toguchi, 2003), but do so in different ways. The Japanese participants rated themselves as more positive than the mid-point on collectivistic attributes such as 'cooperative' and 'respectful', but did not self-enhance on individualistic attributes such as 'self-reliant' and 'unique'. American participants showed the reverse pattern, and were actually self-effacing on the collectivistic traits. The authors interpret this finding in terms of a universal propensity to self-enhancement, which is expressed in whatever domain excellence is rewarded in, in the local context. This interpretation would accord with the error management account that we develop below, which suggests some ways in which cultural differences could emerge. We return to this issue in the general discussion.

Taylor and Brown offer an explanation for the prevalence of the positive illusions that tacitly contains an error management argument. They argue that positive illusions motivate

people to persevere towards goals that would be beneficial but which have an objectively low probability of success (1988, p. 199). For example, HIV-positive men who are developing symptoms of AIDS have beliefs about the controllability of the disease, which are unrealistic, but nonetheless serve to motivate them towards active health-promoting behaviors (Taylor et al., 1992). Nettle (2004) provides a more formal evolutionary model of the Taylor and Brown argument. Accurately assessing the likelihood of obtaining some outcome in the real world is very difficult, because situations do not recur with exactly the same parameters. The two possible errors will lead to opposite behaviors; a false negative to passivity, and a false positive to over-sanguine behavior, with projects taken on that do not succeed. EMT predicts that if the cost of trying and failing is low relative to the potential benefit of succeeding, then an illusional positive belief is not just better than an illusional negative one, but also better than an unbiased belief (see figure 1 and table 2). This is the smoke detector principle applied to a positive outcome. It is better to believe that you can get something desirable even if you can't, as long as the cost of the false alarm is low relative to the opportunity cost of missing out on a fitness-enhancing opportunity.

The EMT approach does indeed seem to account for the domains where the positive illusions occur. People have unrealistically positive views of precisely those characteristics of themselves which are desirable or beneficial (Brown, 1986; Campbell, 1986), and when people judge third parties and thus derive no potential benefit from enhancement, the positive bias disappears (Campbell, 1986). People are unrealistically optimistic about the probability that fitness-enhancing outcomes such as finding an ideal partner and gaining professional status will happen to them (Weinstein, 1980). They have been argued to be unrealistically optimistic (that is, to underestimate the likelihood) of health problems (Weinstein, 1982), which at first would

seem opposite of what an error management account would predict. However, our interpretation of this phenomenon is that people are unrealistically optimistic about their effectiveness of their own efforts to avoid health problems (Taylor, Helgeson, Reed, & Skokan, 1991; Taylor et al., 1992). This makes sense from the EMT perspective, as trying to avoid health difficulties that are inevitable is a lower cost error than failing to avoid those that are avoidable.

The two different smoke detector biases predicted by EMT – excessive sensitivity to potential harms coming from outside, and excessive optimism about benefits that can be obtained by the self – predict that reasoning in domains controlled by the self may display different biases to reasoning in domains beyond the self's control. This is the essence of the paranoid optimism phenomenon, predicting paranoia about the environment but optimism about the self. There are phenomena in the literature which suggest such double standards. For example, a meta-analysis of over 70 life satisfaction studies from 9 countries shows that people tend to believe that their own life is getting better, while also believing that life in general in the country where they live is getting worse (Hagerty, 2003). Similarly, people feel they are less likely than average to be involved in an automobile accident when they are the driver, but not when they are the passenger (McKenna 1993). Such discrepancies are an area where EMT makes interesting predictions for further research.

The Illusion of Control. Finally, where events display some randomness, people judge that their behavior has a greater influence on the flow of events than is in fact warranted, resulting in the so-called illusion of control (Alloy & Abramson, 1979; Langer, 1975; Langer & Roth, 1975; Rudski, 2000; Vazquez, 1987). Given that the controlling behaviors in these experiments are usually rather low cost (pressing a key, for example), it is a less costly error to

continue the control behavior when it is in fact ineffective (the false positive) than it is to miss out on the chance to control events (the false negative).

Related to the illusion of control are superstitions. It was Skinner (1948) who first showed that if a pigeon is given food reinforcement every 15 seconds, regardless of its behavior, it may develop behavioral rituals, such as walking in a circle or rubbing its face on the floor. Skinner's explanation was in terms of adventitious reinforcement; a behavior that had once occurred before the delivery of food was 'assumed' to have caused the delivery of food. Very similar effects can be demonstrated in humans, who when presented with actually random patterns of reinforcement, develop superstitious beliefs about actions they must perform to produce the desired contingency (Catania & Cutts, 1963; Matute 1994, 1995; Ono, 1987; Rudski, 2001). Such effects are not confined to the laboratory; naturalistic surveys reveal that belief in lucky charms and lucky tricks is widespread (Vyse, 1997). Experiments by Matute (1994,5) show that the result of uncontrollable reinforcement in a human conditioning paradigm is not passivity or learned helplnessness, but instead, superstitious behaviour and a strong subjective illusion of control. Only when explicit feedback of the non-effectiveness of the superstitious behavior is provided does the illusion disappear, and under such conditions, learned helplessness ensues. There is a conceptual link with depression here, since depression has often been explained in terms of learned helplessness (e.g. Abramson, Seligman and Teasdale 1978), and depressed subjects are distinguished by the absence of illusion in control paradigms (Alloy and Abramson 1979, Vazquez 1987). Thus, the evidence suggests that superstitions and illusion of control, though strictly speaking irrational, are healthy responses to an uncertain world.

In the ancestral environment, accurate information about the true contingencies between people's behavior and events around them, such as the movements of game animals, would have

been scarce. As long as the cost of performing the superstitious behaviors was low relative to the benefit of actually controlling events, EMT would predict cognitive mechanisms biased towards superstition and the illusion of control to evolve.

Discussion

Adaptive Biases

We have reviewed a large number of cases where apparently irrational biases in cognition are explained by the existence of asymmetric error costs and significant uncertainty. Thus, bias in cognition is no longer a shortcoming in rational behavior, but an adaptation of behavior to a complex, uncertain world. Biased mechanisms are not design defects of the human mind, but rather design features. In view of the content specificity of these effects, and the absence of bias in many other types of cognition, a theory that held bias to be a generalized outcome of individual or cultural learning seems implausible. Rather, it seems likely that the mind is equipped with multiple, domain specific cognitive mechanisms, with specific biases appropriate to the content of the task and the particular pattern of costs, benefits and likelihoods. For example, we are predisposed to fear spiders and snakes rather than elements of our contemporary environment that are in fact much more dangerous, such as electrical outlets. We are predisposed to fear injured or diseased people and contamination of the food supply, when in fact road traffic and obesity are much more likely to kill us. We are prone to sex differences in the perception of sexual intent, and to assume social non-reciprocation has dispositional rather than situational causes. We are prone to believe that random events in the environment reflect the operations of some unseen intelligence.

The existence of these biased systems is an important link between psychology and culture. To persist in a culture, a pattern of information must capture the attention of individuals

such that they will remember and pass it on. Those elements of culture best able to exploit the inherent biases of the mind will have the greatest probability of being retained and transmitted. In fact, tales of invisible gods who orchestrate the natural world, legends of dangerous serpents, stories of plagues, and taboos about meat all abound in the world's cultures (Atran, 2002; Atran & Norenzayan, in press; Fessler, 2002; Guthrie, 2001).

Open versus Closed Developmental Systems

Our argument is not that all of the biases we have described are produced by the same cognitive mechanism, but rather than they have all been produced by the same evolutionary mechanism, that is selection to minimize overall error costs, acting on many different cognitive systems. Some of these systems are relatively closed. For example, the system of food aversions, or the predispositions to fear snakes and spiders, seems to have fixed content and require only triggering by the environment. Other biases, such as optimism about future fitness prospects, are much more open to environmental influence. In one culture the relevant domain for positive illusions might be hunting, in another success in college, and in still another, standing in the local community. The cognitive system leaves open the flexibility for the individual to identify those domains in the environment where success yields benefits, and those where failure is costly.

We would predict that biases produced by relatively closed systems, such as snake and spider fears, and food aversions, would show less cross-cultural variation. Biases such as the positive illusions, which are produced by open systems, would have the possibility of local variation. Such variation might arise for several reasons. It might be that in a collective cultural context, in which social rewards are contingent on cooperation and loyalty to the group, the benefits to, for example, earning extra money are diminished. In as much as such cultures disfavor individualists, there might actually be significant social costs to individual success in

competitive affairs. In such a culture, the costs of the two errors would actually be different compared to an individualistic culture, and so EMT would predict that positive biases would not appear. Indeed, EMT would predict that if it is true that East Asian cultures operate in a more collectivist way than Western ones, then the positive biases should be shifted in East Asia towards attributes related to excellence as a collective member and away from those to do with excellence in inter-individual competition. This is precisely the pattern found by Sedikides et al. (2003).

Differential Evocation of Bias

In many domains ancestrally, asymmetries in costs varied depending on context. The costs of missing threats are highest, for example, when individuals are vulnerable—when they are sick, alone, or otherwise unprotected. If moderating contexts were recurrent, consistent in their effects, and signaled by reliable cues, we should expect judgmental adaptations to respond to them with variable degrees of bias.

We have already discussed several cases in which biases differ by context: sinister attributions are more likely when people are new to social groups, negative forgiveness bias is more common in relationships at risk, and aggressive stereotypes about outgroups are enhanced in the dark. In each of these cases, a cue that was present in both ancestral environments and today—new social partners, relationship discord, and darkness—shifts the bias.

A complementary way to understand adjustments of bias may occur is through emotion. Emotion states are activated in response to threats and opportunities and they may adaptively channel us toward the specific thoughts and courses of action needed to respond to them (Cosmides & Tooby, 2000). Maner and colleagues (Maner et al., in press) hypothesized that fear would increase biases toward inferring aggressiveness in others, particularly members of

coalitional outgroups; sexual arousal, on the other hand, would increase men's bias toward overinferring sexual desire in women. They showed men and women clips of scary or romantically arousing films, and then asked them to interpret "micro-expressions" in photographs of people who had relived an emotionally-arousing experience but were attempting to conceal any facial expressions that would reveal it (the faces were actually neutral in expression). In the fear condition, the study participants, who were mostly White, "saw" more anger on male faces, especially the faces of outgroup (Black and Arab) males. The fear manipulation had no effect on perceptions of sexual arousal in the faces. In the romantically arousing film condition, men perceived greater sexual arousal in female faces, particularly when the faces were attractive. The arousal manipulation did not increase men's perceptions of sexual arousal in other men's faces, and the manipulation did not increase women's perceptions of sexual arousal in any of the faces. Thus, the effects were emotion and target specific, and for sexual arousal, sex specific. When fearful, men and women perceived greater threat from ethnic outgroup members; when aroused, men but not women perceived greater arousal in attractive opposite-sex faces.

Park, Schaller, and colleagues have documented parallel effects in the domain of disease-avoidance. They proposed that adaptations for disease avoidance are overinclusive and respond noncommunicable phenotypic anomalies and even a target's status as a "foreigner." They demonstrated that biased associations of phenotypic cues with disease increases when people are fearful of contamination. European-American students who read a news clip about a local hepatitis outbreak showed stronger associations between words like "disability" and "disease" and between "disability" and "unpleasant" on the implicit association test, as compared with controls (Park, Faulkner, & Schaller, 2003). In a subsequent study using the implicit association

test, participants were exposed to slides evoking pathogen risk (germs lurking in a kitchen sponge) or accidents (electrocution in a bathtub). Those who viewed the pathogen slides showed greater associations between slides of obese people and disease than those in the accident condition (Park, Schaller, & Crandall, 2004). Using related methods, the researchers found similar effects concerning immigrant groups that were unfamiliar to their Canadian participants. Participants in the pathogen condition had more negative attitudes about allowing immigration of unfamiliar immigrant groups (Nigerians in one study and Mongolians in another) than familiar immigrant groups (Scots and Taiwanese). In the accident condition, attitudes about these immigrant groups did not differ (Faulkner, Schaller, Park, & Duncan, in press). According to the logic of these studies, individuals whose immune systems are depressed might also be expected show increased bias toward these groups. Pregnant women experience reproductive immunosuppression to prevent rejection of the fetus, which shares only 50% of the mother's genes (Fessler, 2002); therefore we predict that pregnant women will experience enhanced disease avoidance biases.

Some emotional and motivational states are chronically present in some people, and therefore biases moderated by these states will also reliably differ between them. In the micro-expressions studies (Maner et al, in press), people who believed in general that the world is a dangerous place, saw more anger in male outgroup faces. People who tended toward a more promiscuous mating strategy saw more sexual arousal in opposite sex faces (Maner et al, in press). In the disease avoidance studies, individuals who scored high on an individual difference measure of germ aversion or vulnerability to disease also showed stronger disability-disease associations on the implicit attitudes measure (Park, Faulkner, & Schaller, 2003), greater dislike

of fat people (Park, Schaller, & Crandall, 2004), and more negative attitudes about unfamiliar immigrant groups (Faulkner, Schaller, Park, & Duncan, in press).

New Predictions

EMT predicts the evolution of biases wherever the problem involves significant uncertainty, has recurred and impacted fitness over evolutionary time, and where the two types of error have reliably had asymmetrical cost. EMT also predicts the direction of bias, which will be towards making the less costly of the two errors. Some of the effects we have reviewed were predicted in advance using error management logic. These include commitment underperception by women (Haselton & Buss, 2000), overinclusiveness of disease avoidance (e.g., Park et al., 2003; Faulkner et al., in press), the use of the social exchange heuristic (Yamagishi et al., 2003), negative forgiveness bias (Friesen et al., in press), and content effects in the fundamental attribution error (Burkett et al., 2003). Just as many such situations have already been studied, there may be many more that have not yet been the subject of empirical investigations.

We have already described several new predictions. We suggested that the personality domains in which the FAE is particularly likely to occur will be those that are most likely to impose fitness costs, such as aggressiveness and deceitfulness. In the previous section we proposed that the cultural differences in the domains in which positive illusions occur will be linked to cultural differences in the value of those domains. Qualities or outcomes that are universally valued, such as the preservation of health, will vary little across cultures. EMT also predicts that discrepancies between judgments about outcomes the subject controls will often show different biases to those not in the subject's control, as in the result that people believe their own life to be getting better but life in general to be getting worse. Such effects might be elicited in many different domains. For example, in a simulation of the transmission of a disease, EMT

predicts that people should be overly fearful that others are infectious, but overly optimistic that their own attempts to avoid contagion will be effective. We also suggested that pregnant women will express disease avoidance attitudes that are especially strong or overinclusive.

Haselton & Buss (2000) predicted two biases in the domain of sexuality and courtship. We suggest two more. The first is a bias in inferring the romantic or sexual interest of others in one's mate and the second a bias in inferring the interest of one's mate in others. First, the fitness costs of failing to recognize the interest of an interloper in one's mate and to lose one's mate as a result are high. One must reinitiate mate search, pay new costs associated with courtship and attraction, and risk the loss of investment from the mate in existing offspring. The costs of somewhat elevated vigilance, especially if activated only in situations presenting plausible threat, would be comparatively low. Thus we predict the *interloper effect*: a bias toward over-inferring the sexual interest of others in one's mate in ambiguous or mildly threatening situations. For example, at a cocktail party, if an attractive other behaves in a friendly and animated fashion toward ego's mate, ego will assume greater sexual interest on the part of the other than will an independent on-looker. This bias would function, we propose, to increase mate retention efforts and help to ward off defection. Smurda and Haselton (2002) documented evidence suggestive of the interloper bias. They found that people involved in committed relationships tended to rate the sexual interest of same sex others (e.g., based on a smile) more highly than people not involved in relationships. Maner and colleagues (Maner et al., 2003) found that women in committed relationships showed a greater attentional and memorial bias for attractive female faces than women not in relationships, providing additional suggestive evidence of the interloper bias.

Second, the fitness costs to a man of failing to detect partner infidelity are high. His own reproduction can be delayed for the course of a pregnancy, at minimum. He also risks investing time and resources in the offspring of a reproductive competitor. However, the costs of false alarms are also plausibly high. Undue suspicion can damage relationships and time spent on unneeded monitoring of the partner results in missed opportunities to pursue other fitness enhancing activities, such as the collection of food or providing care for kin. Thus, there is a delicate balance between the costs of errors in infidelity detection (also see Buss, 2000). This balance shifts, however, over the course of the woman's menstrual cycle. As ovulation nears, fertility increases, and the risks to a man of cuckoldry are at their highest. Therefore, we propose a bias in men toward over-inferring extra-pair sexual interest (and, in extreme cases, infidelity) when (a) his partner is nearing midcycle and (b) he is confronted with ambiguous cues to infidelity (such as his partner's expressed friendliness to another man). This general logic also predicts that the interloper bias discussed above may become acute for men when their partners are most fertile. These predictions are rendered plausible by recent evidence suggesting that men have adaptations sensitive to their partner's fertility status. For example, women's body scent, including scent samples taken from the torso and upper body and samples of vaginal secretions, is rated as most attractive during the high fertility phase of the cycle (Doty, Ford, & Preti, 1975; Singh & Bronstad, 1999; Thornhill et al, 2003). Women also report increased love, attraction, sexual proprietariness, and jealousy expressed by their partners near ovulation as compared with other cycle phases (Gangestad, Thornhill, & Garver, 2002; Haselton & Gangestad, 2004).[1]

[1] Note, however, that these studies do not demonstrate bias in men's inferences of women's proclivity toward infidelity. There are at least three possible explanations for this effect: (1) given that women's extra-pair interests are elevated at midcycle (Gangestad, Thornhill, & Garver, 2002; Haselton & Gangestad, 2004) men could be tracking actual risk through their partner's behavior and adjusting their mate guarding accordingly; (2) women's perceptions of their partners behaviors change with their cycle; or (3) men use ovulatory cues to adjust their mate guarding efforts when their partners are most fertile and hence they are at greatest risk of cuckoldry. The hypothesis we advance is a version of (3), that men use ovulatory cues to adjust mate guarding and that they become biased

One of the best-researched examples we have discussed is the easily elicited fear of snakes and spiders. Snakes and spiders were not the only dangerous animals in ancestral environments. Predatory cats and other large mammals, as well as large reptiles such as crocodiles, have likely played a role in the evolutionary history of human, and have shaped a predator avoidance psychology in humans (Barrett, 1999). Therefore, we predict that the same effects documented for snakes and spiders will be documented for these other ancestrally dangerous animals. Moreover, we hypothesize that the environmental cues that reliably increased susceptibility to injury should increase false alarm rates in the detection of these animals and in inferences of their dangerousness. One such cue is ambient darkness (Schaller et al, 2003). Darkness and states of fear should also amplify other protective biases, including auditory looming (estimating early arrival of objects traveling toward you).

Bias versus Accuracy

Krueger and Funder (2004) raised questions about the obsessional focus of many psychologists on bias and error, which has led to an unnecessarily dreary outlook on human cognition and a failure to study how accurate judgments are actually made. In the studies we have reviewed, our focus on documented biases does not imply that people are usually (or often) wildly off-base. In the Haselton and Buss studies (2000), men and women showed remarkable agreement about how much commitment or sexual interest each dating cue communicated (with correlations above .90). But, at the same time, men overestimated women's sexual interest and women underestimated men's commitment. Likewise, in the forgiveness bias studies, partners tended to agree on whether one partner had forgiven the other (with a maximum correlation of .44), but they still tended to underestimate how much they had been forgiven (Friesen et al, in

---

toward false alarms in their inferences of their partner's extra-pair sexual interest and behavior. Controlled laboratory experiments may be required to test the infidelity-bias hypothesis.

press). Thus, as Fletcher (2002) noted, bias and accuracy can vary quite independently, and systematic bias does not preclude reasonable accuracy.

The criteria for predicting bias are different from those for predicting accuracy. An error management bias is predicted when errors differ reliably in their costs. Accuracy, or judgmental *sensitivity*, is predicted when valid cues are available (Funder, 1995) and the fitness consequences of correct discrimination are large—for example, in judging the dominance or sociosexual orientation of others (Gangestad, Simpson, DiGeronimo, & Biek, 1992). If the fitness consequences of discrimination are large, and there is a differential cost of errors in one direction or the other, then a judgmental system should be both sensitive and biased. In courtship, it is important to recognize that some cues are greater indicators of sexual interest than others (smiling vs. stoking one's date on the thigh), but it also pays to overestimate the degree to which cues indicate interest if it helps a man to avoid a miss.

The Rational Actor?

An important reason for seeking an explanatory framework for biases concerns the adequacy of human reasoning. Much social theory, particularly in economics and political science, depends on conceptualizing the individual as a rational actor able to use information available to him in an optimal way given his aims and objectives. If people turn out not to be rational in the required sense, such models lack validity. Experimentally and observationally based research, such as that carried out by social psychologists, anthropologists, and experimental economists, has often cast doubt on the accuracy of the rational actor assumption (Bell, 1995; Davis & Holt, 1993; Kahneman et al., 1982). However, if observed departures from rationality are studied piecemeal and accepted as so many quirks of human beings without seeking deeper explanation, social science becomes balkanized between theorists who have a

powerful explanatory framework that lacks validity, and empiricists who have an accurate list of phenomena but no explanations (Hermann-Pillath, 1994; Nettle, 1997). Moreover, the outlook for human rationality is bleak, since the implication is simply that people, because of limitations in cognitive machinery, are not capable of optimal decision making.

The reinterpretation of many biases as design features rather than design defects suggests a different perspective. Both the content and direction of biases can be predicted theoretically and explained by optimality when viewed through the long lens of evolutionary theory. Thus, the human mind shows good design, though it is design for fitness-maximization, not truth preservation. This reorientation accords with other recent work in psychology. For example, the heuristics and biases tradition (Kahneman et al., 1982) saw the mind as made up of simple problem solving tools that, while functional over a restricted range of circumstances, are simply inadequate to produce optimal judgment in general, resulting in a wide range of 'cognitive illusions' or pervasive departures from optimality. More recent work, however, has questioned this bleak view. Some cognitive illusions disappear or greatly attenuate when the task is presented in an ecologically valid format (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995). Ecological validity, a long-standing but under-theorized term in psychology, may in effect be equated to the task format approximating some task that humans have performed recurrently over evolutionary time.

Moreover, many of the simple heuristics that people actually use perform just as well as complex normative models under real-world conditions of partial knowledge (Gigerenzer & Todd, 1999). There are even circumstances in which they perform *better* than normative models, the so-called 'less is more' effect. The less is more effect occurs because simple heuristics can exploit structural features of the decision-making environments which are noisy and uncertain

and contain multiple cues. EMT complements the 'less is more' principle with a 'biased is better' principle; under some circumstances, which can be predicted in a principled way, biased strategies are actually superior to non-biased ones.

Conclusion

EMT predicts that over a certain set of conditions, biased reasoning strategies can be adaptive. Most importantly, where error costs are known and asymmetric, and there is uncertainty about likelihoods, a biased reasoning strategy can actually do *better* than an unbiased one. It strikes us that these conditions are likely to characterize many of the real dilemmas that have faced us and our ancestors. Since dilemmas are never repeated with exactly the same parameters, likelihoods are very hard to infer accurately. However, the payoffs for various kinds of outcomes, from being bitten by a snake to obtaining a mate, are recurrently positive or negative over evolutionary time. Thus, EMT predicts that highly specific biases should evolve.

Error management theory is an additional element in a picture of the mind as a well-designed instrument for solving the kinds of problems that have faced human beings over their evolutionary history. Many apparent quirks of human thought, from our fear of harmless spiders, to our superstition and paranoia, to our eternal optimism, may be optimal adaptations to the worlds in which we have lived.

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.

Abbey, A. & Harnish, R. J. (1995). Perception of sexual intent: The role of gender, alcohol consumption, and rape supportive attitudes. *Sex Roles, 32*, 297-313.

Abbey, A., & Melby, C. (1986). The effects of nonverbal cues on gender differences in perceptions of sexual intent. *Sex Roles, 15*, 283-298.

Abramson, L.Y., Seligman, M.E.P. and Teasdale, J.D. (1978). Learned helplessness in humans: Critique and reformulation. *Journal of Abnormal Psychology*, *87,* 49-74.

Alloy, L. B. & Abramson, L. Y. (1979). Judgment of contingency in depressed and non-depressed subjects: Sadder but wiser? *Journal of Experimental Psychology: General, 108*, 443-479.

Andrews, P. W. (2001). The psychology of social chess and the evolution of attribution mechanisms: Explaining the fundamental attribution error. *Evolution and Human Behavior, 22*, 11-29.

Atran, S. (2002). *In Gods We Trust: The Evolved Landscape of Religion*. Oxford University Press, New York.

Atran, S., & Norenzayan, A. (in press). Religion's evolutionary landscape: Counterintuition, commitment, compassion, communion. *Behavioral and Brain Sciences.*

Barrett, H. C. (1999). *Human cognitive adaptations to predators and prey*. PhD Dissertation, University of California at Santa Barbara.

Barrett, H. C., Todd, P. M. Miller. G. F. & Blythe P. W. (in press). *Accurate judgments of intention from motion cues alone: A cross-cultural study*. Evolution and Human Behavior.

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117*(3), 497-529.

Bell, D. (1995). On the nature of sharing: Beyond the range of methodological individualism. *Current Anthropology, 36*, 826-830.

Bishop, G. D., Alva, A. L., Cantu, L., Rittiman, R. K. (1991). Responses to persons with AIDS: Fear of contagion or stigma? *Journal of Applied Social Psychology, 21*, 1877-1888.

Bloom, P. & Veres, C. (1999). The perceived intentionality of groups. *Cognition, 71*, B1-B9.

Bouskila, A. & Blumstein, D.T. (1992). Rules of thumb for predation hazard assessment: predictions from a dynamic model. *The American Naturalist, 139*, 161-176.

Bowles, S. & Gintis, H. (2002) Homo reciprocans. *Nature, 415*, 125-8.

Brewer, M. B. (1979). Ingroup bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin, 86*, 307-324.

Brown, J. D. (1986). Evaluations of self and others: Self-enhancement biases in social judgments. *Social Cognition, 4*, 353-376.

Burkett, B. N, Cosmides, L. & Kirkpatrick, L. A. (2003, July) *Evidence for domain-specificity of trait attribution: Replication and extension.* Paper presented at the Human Behavior and Evolution Society Conference, Lincoln, Nebraska.

Buss, D. M. (1994). *The evolution of desire: Strategies of human mating*. New York: Basic Books.

Buss, D. M. (2000). *The dangerous passion: Why jealousy is as necessary as love and sex*. New York: Free Press.

Camerer, C. & Thaler, R. (1995). Ultimatums, dictators and manners. *Journal of Economic Perspectives, 9*, 337-356.

Campbell, J. D. (1986). Similarity and uniqueness: The effects of attribute type, relevance and individual differences in self-esteem and depression. *Journal of Personality and Social Psychology, 50*, 281-294.

Caporael, L., Dawes, R.M., Orbell, J.M., & van der Kragt, A.J. (1989). Selfishness examined. *Behavioral and Brain Sciences 12*, 683-739.

Cashdan, E. (1998). Adaptiveness of food learning and food aversions in children. *Social Science Information, 37*, 613-632.

Catania, A. C. & Cutts, D. (1963). Experimental control of superstitious responding in humans. *Journal of the Experimental Analysis of Behavior, 6*, 203-208.

Choi, I., Nisbett, R.E., & Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. *Psychological Bulletin, 125*, 47-63.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? *Cognition, 31*, 187-276.

Cosmides, L. & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58*, 1-73.

Cosmides, L. & Tooby, J. (2000). Evolutionary psychology and the emotions. In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of Emotions, 2nd Edition*. (pp. 91-115.) New York: Guilford.

Crocker, J., & Park, L. E. (2003). Seeking self-esteem: Construction, maintenance, and protection of self-worth. In M. R. Leary, & J. P. Tangney (Eds.), *Handbook of self and identity; handbook of self and identity.* (pp. 291-313). New York: Guilford.

Daly, M. & Wilson, M.I. (1988). *Homicide*. Hawthorne, NY: Aldine de Gruyter.

Davis, D. D. & Holt, C. A. (1993). *Experimental Economics.* Princeton University Press, Princeton.

de Jong, P. & Merckelbach, H. (1991). Covariation bias and electrodermal responding in spider phobics before and after behavioural treatment. *Behaviour Research & Therapy, 29*, 307-314.

Doty, R.L., Ford, M., Preti, G., & Huggins, G. R. (1975). Changes in the intensity and pleasantness of human vaginal odors during the menstrual cycle. *Science 190*, 1316-1317.

Faulkner, J., Schaller, M., Park, J. A., & Duncan, L. A. (in press). Evolved disease avoidance mechanisms and contemporary xenophobic attitudes. *Group Processes & Intergroup Relations.*

Fenigstein, A. (1984). Self-consciousness and the overperception of self as a target. *Journal of Personality & Social Psychology, 47*, 860-870

Fessler, D. M. T. (2002). Reproductive immunosuppression and diet - An evolutionary perspective on pregnancy sickness and meat consumption. *Current Anthropology, 43*, 19-61.

Fletcher, G. (2002). *The new science of intimate relationships.* Oxford: Blackwell.

Fox, E., Russo, R., & Dutton, K. (2002). Attentional bias for threat: Evidence for delayed disengagement from emotional faces. *Cognition and Emotion, 16*, 355-379.

Friesen, M. D., Fletcher, G. J. O., & Overall N. C. (in press). A Dyadic Assessment of Forgiveness in Intimate Relationships. *Personal Relationships*.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*, 652-670.

Gangestad, S. W., Simpson, J. A., DiGeronimo, K., & Biek, M. (1992). Differential accuracy in person perception across traits: Examination of a functional hypothesis. *Journal of Personality & Social Psychology, 62,* 688-698.

Gangestad, S. W., Thornhill, R., & Garver, C. E. (2002). Changes in women's sexual interests and their partners' mate retention tactics across the menstrual cycle: Evidence for shifting conflicts of interest. *Proceedings of the Royal Society of London: B Biological Sciences*, *269*, 975-982.

Garcia, J., Ervin, F. R., & Koelling, R. A. (1966). Learning with prolonged delay of reinforcement. *Psychonomic Science, 5*, 121-122.

Garcia, J., Hankins, W. G., & Rusiniak, K. W. (1976). Flavor aversion studies. *Science, 192*, 265-266.

Gigerenzer, G. & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684-704.

Gigerenzer, G. & Todd, P. M. (1999). *Simple Heuristics that Make Us Smart*. Oxford: Oxford University Press.

Green, D. M. & Swets, J. A. (1966). *Signal detection and psychophysics*. New York: Wiley.

Greenberg, J., Pyszczynski, T., Solomon, S., & Pinel, E. (1993). Effects of self-esteem on vulnerability-denying defensive distortions: Further evidence of an anxiety-buffering function of self-esteem. *Journal of Experimental Social Psychology, 29*, 229-251.

Guthrie, S. (2001). Why Gods? A cognitive theory. In J. Andresen (Ed.) Religion in Mind: *Cognitive perspectives on religious belief, ritual, and experience*. Cambridge: Cambridge University Press.

Hagerty, M. R. (2003). Was life better in the 'good old days'? Inter-temporal judgments of life satisfaction. *Journal of Happiness Studies, 4*, 115-39.

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*, 43-47.

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

Haselton, M. G. & Gangestad, S. W. (2004). *Conditional expression of female desires and male mate retention efforts across the human ovulatory cycle*. Manuscript under review.

Heider, F. & Simmel, S. (1944). An experimental study of apparent behavior. *American Journal of Psychology, 57*, 243-259.

Heine, S. J., Lehman, D. R., Markus, H., & Kitayama, S. (1999). Is there a universal need for positive self-regard? *Psychological Review, 106*, 766-794.

Henrich, J. & Boyd, R. (2001). Why people punish defectors: conformist transmission stabilizes costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology, 208*, 79-89.

Henrich, J., Boyd, R., Bowles, S., Camerer, C. Gintis, H, McElreath, R, & Fehr, E. (2001). In search of Homo economicus: Experiments in 15 Small-Scale Societies. American Economic Review, 91, 73-79.

Hermann-Pillath, C. (1994). Ecological rationality, Homo Economicus, and the origins of the social order. *Journal of Social and Evolutionary Systems, 17*, 41-69.

Hurtado, A. M. & Hill, K. R. (1992). Paternal effect on offspring survivorship among Ache and Hiwi hunter-gatherers. In B. S. Hewlett et al. (Eds.), *Father-child relations: Cultural and biosocial contexts*. (pp. 31-55). New York: Aldine De Gruyter.

Johnson, C. B., Stockdale, M. S. & Saal, F. E. (1991). Persistence of men's misperceptions of friendly cues across a variety of interpersonal encounters. *Psychology of Women Quarterly, 15*, 463-475.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgement under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Keeley, L.H. (1996). *War Before Civilization: The Myth of the Peaceful Savage.* New York: Oxford University Press.

Kitayama, S., Markus, H., Matsumoto, H., & Norasakkunkit, V. (1997). Individual and collective processes in the construction of the self: Self-enhancement in the United States and self-criticism in Japan. *Journal of Personality and Social Psychology, 72*. CHECK pages

Kramer, R. M. (1994). The sinister attribution error: Paranoid cognition and collective distrust in organizations. *Motivation and Emotion, 18*, 199-230.

Kramer, R. M. (1998). Paranoid cognition in social systems: Thinking and acting in the shadow of doubt. *Personality and Social Psychology Review, 2*, 251-275.

Krebs, D. L. & Denton, K. (1997). Social illusions and self-deception: The evolution of biases in person perception. In J. A. Simpson & D. T. Kenrick (Eds.) pp. 21-47. *Evolutionary social psychology.* Hillsdale, NJ: Erlbaum.

Krueger, J. I. & Funder, D. C. (2004). Towards a balanced social psychology: Causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences*, *27*, 313–376

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*, 480-498.

Kurzban, R. & Leary, M. R. (2001). Evolutionary origins of stigmatization: The functions of social exclusion. *Psychological Bulletin, 123*, 187-208.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology, 32*, 311-328.

Langer, E. J. & Roth, J. (1975). Heads I win, tails it's chance: The illusion of control as a function of the sequence of outcomes in a purely chance task. *Journal of Personality and Social Psychology, 32*, 951-955.

Lima, S.L. & Dill, L.M. (1990). Behavioural decisions made under the risk of predation: A review and prospectus. *Canadian Journal of Zoology, 68*, 619-640.

Maner, J. K., Kenrick, D. T., Becker, D. V., Delton, A. W., Hofer, B., & Wilbur, C. J. et al. (2003). Sexually selective cognition: Beauty captures the mind of the beholder. *Journal of Personality & Social Psychology, 85*, 1107-1120.

Maner, J. K., Kenrick, D. T., Becker, V., Robertson, T. E., Hofer, B., Neuberg, S. L. Delton, A. W., Butner, J. & Schaller, M. (in press). Functional Projection: How Fundamental Social Motives Can Bias Interpersonal Perception. *Journal of Personality and Social Psychology.*

Matute, H. (1994). Learned helplessness and superstitious behavior as opposite effects of uncontrollable reinforcement. *Learning and Motivation, 25*, 216-232.

Matute, H. (1995). Human reactions to unavoidable outcomes: Further evidence for superstitions rather than helplessness. *Quarterly Journal of Experimental Psychology, 48*, 142-157.

McKenna, F.P. (1993). It won't happen to me: Unrealistic optimism or illusion of control? *British Journal of Psychology, 84*, 39-50.

Miller, D. T. and Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin, 82*, 213-225.

Mineka, S. (1992). Evolutionary memories, emotional processing, and the emotional disorders. *Psychology of Learning and Motivation, 28*, 161-206.

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.

Nesse, R. M. (2005). Natural selection and the regulation of defenses: A signal detection analysis of the smoke detector problem. *Evolution and Human Behavior 26*, 88-105..

Nettle, D. (1997). On the status of methodological individualism. *Current Anthropology, 38*, 283-286.

Nettle, D. (2004). Adaptive illusions: Optimism, control and human rationality. In D. Evans & P. Cruse (Eds.), *Emotion, Evolution and Rationality*, pp. 193-208. Oxford: Oxford University Press.

Neuhoff, J. G. (1998). Perceptual bias for rising tones. *Nature, 395*, 123-124.

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

Nisbett, R. E., Caputo, C., Legant, P., & Marecek, J. (1973). Behavior as seen by the actor and as seen by the observer. *Journal of Personality and Social Psychology, 27*, 154-164.

Norenzayan, A., Choi, I., & Nisbett, R.E. (2002). Cultural similarities and differences in social inference: Evidence from behavioral predictions and lay theories of behavior. *Personality and Social Psychology Bulletin, 28*, 109-120

Oehman, A. & Soares, J. J. (1993). On the automatic nature of phobic fear: Conditioned electrodermal responses to masked fear-relevant stimuli. *Journal of Abnormal Psychology, 102*, 121-132.

Ono, K. (1987). Superstitious behavior in humans. *Journal of the Experimental Analysis of Behavior, 47*, 261-271.

Park, J. H., Faulkner, J., & Schaller, M. (2003). Evolved disease-avoidance processes and contemporary anti-social behavior: Prejudicial attitudes and avoidance of people with disabilities. *Journal of Nonverbal Behavior, 27*, 65-87.

Park, J. H., Schaller, M., & Crandall, C. S. (2004). Obesity as a heuristic cue connoting contagion: Perceived vulnerability to disease promotes anti-fat attitudes. Manuscript under review.

Pratto, F. P. & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology, 61*, 380-391.

Price, M., Cosmides, L. & Tooby, J. (2002). Punitive sentiment as an anti-free rider psychological adaptation. *Evolution and Human Behavior 23*, 203-31.

Quillian, L. & Pager, D. (2001). Black neighbors, higher crime? The role of racial stereotypes in evaluations of neighborhood crime. *American Journal of Sociology 107,* 717-67

Ratcliffe, J. M., Fenton, M. B., & Galef, B. G. (2003). An exception to the rule: Common vampire bats do not learn taste aversions. *Animal Behaviour, 65*, 385-389.

Rice, W. R. (1996). Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. *Nature, 361*, 232-234.

Rice, W. R. (2000). Dangerous Liaisons. *Proceedings National Academy of Sciences, 97*, 12953-12955.

Ross, L. (1977). *The intuitive psychologist and his shortcomings.* In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 10, pp. 173–220). New York: Academic Press.

Rozin, P. & Fallon, A. E. (1987). A perspective on disgust. *Psychological Review, 94*, 23-41.

Rozin, P. & Kalat, J. W. (1971). Specific hungers and poison avoidances as adaptive specializations of learning. *Psychological Review, 78*, 459-486.

Rozin, P., Markwith, M., & Nemeroff, C. (1992). Magical contagion beliefs and fear of AIDS. *Journal of Applied Social Psychology, 22*, 1081-1092.

Rozin, P., Markwith, M., & Ross, B, (1990). The sympathetic magical law of similarity, nominal realism, and neglect of negatives in response to negative labels. *Psychological Science, 1*, 383-384.

Rudski, J. M. (2000). Illusion of control relative to chance outcomes. *Psychological Reports, 87*, 85-92.

Rudski, J. (2001). Competition, superstition and the illusion of control. *Current Psychology, 20*, 68-84.

Saal, F. E. Johnson, C. B. & Weber, N. (1989). Friendly or sexy? It may depend on whom you ask. *Psychology of Women Quarterly, 13*, 263-276.

Sally, D. (1995). Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and Society, 7*, 58-92.

Savitsky, K., Epley, N., & Gilovich, T. (2001). Do others judge us as harshly as we think? Overestimating the impact of our failures, shortcomings, and mishaps. *Journal of Personality & Social Psychology, 81*, 44-56

Schaller, M., Park, J. H., & Faulkner, J. (in press). Prehistoric dangers and contemporary prejudices. *European Review of Social Psychology, 14*, 105-137.

Schaller, M., Park, J. H., & Mueller, A. (2003). Fear of the dark: Interactive effects of beliefs about danger and ambient darkness on ethnic stereotypes. *Personality and Social Psychology Bulletin, 29*, 637-649.

Sedikides, C., Gaertner, L., & Toguchi, Y. (2003). Pancultural self-enhancement. *Journal of Personality and Social Psychology, 84*, 60-79.

Seligman, M. E. P. (1971). Phobias and preparedness. *Behavior Therapy, 2*, 307-320.

Sidanius, J. & Veniegas, R. C. (2000). *Gender and race discrimination: The interactive nature of disadvantage*. In S. Oskamp (Ed.), Reducing prejudice and discrimination. "the claremont symposium on applied social psychology"; reducing prejudice and discrimination. (pp. 47-69). Mahwah, NJ: Lawrence Erlbaum Associates.

Singh D. & Bronstad P. M. (2001). Female body odour is a potential cue to ovulation. *Proceedings of the Royal Society, London: B Biological Sciences, 268,* 797-801

Skinner, B. F. (1948). Superstition in the pigeon. *Journal of Experimental Psychology, 38*, 168-172.

Symons, D. (1979). *The evolution of human sexuality*. New York: Oxford University Press.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*, 1-26.

Taylor, S. E., Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103*, 193-201.

Taylor, S. E., Hegelson, V. S., Reed, G. M., Skokan, L. A. (1991). Self-generated feelings of control and adjustment to physical illness. *Journal of Social Issues, 47*, 91-109.

Taylor, S. E., Kemeny, M. E., Aspinwall, L. G., Schneider, S. G., Rodriguez, R., Herbert, M. (1992). Optimism, coping, psychological distress and high-risk sexual behavior among men at risk for acquired immunodeficiency syndrome (AIDS). *Journal of Personality and Social Psychology, 63*, 460-473.

Thornhill, R., Gangestad, S. W., Miller, R., Scheyd, G., McCollough, J., & Franklin, M. (2003). MHC, symmetry and body scent attractiveness in men and women (*Homo sapiens*). *Behavioral Ecology, 14,* 668-678.

Tomarken, A. J., Mineka, S., Cook, M. (1989). Fear-relevant selective associations and covariation bias. *Journal of Abnormal Psychology, 98*, 381-394.

Tomarken, A. J., Sutton, S. K., & Mineka, S. (1995). Fear-relevant illusory correlations: What types of associations promote judgmental bias? *Journal of abnormal psychology, 104*, 312-326.

Vazquez, C. (1987). Judgement of contingency: Cognitive biases in depressed and non-depressed subjects. *Journal of Personality and Social Psychology, 52*, 419-431.

Vyse, S. A. (1997). *Believing in Magic*. New York: Oxford University Press.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology, 39*, 806-820.

Weinstein, N. D. (1982). Unrealistic optimism about susceptibility to health problems. *Journal of Behavioural Medicine, 5*, 441-460.

Yamagishi, T., Jin, N. & Kiyonari, T. (1999). Bounded generalized reciprocity: Ingroup favoritism and ingroup boasting. *Advances in Group Processes, 16*, 161-197.

Yamagishi, T., Terai, S., Kiyonari, T., & Kanazawa, S. (2003). *The social exchange heuristic: Managing errors in social exchange*. Manuscript under review, Center for Advanced Study in the Behavioral Sciences, Stanford, CA.

Ybarra, O. (2002). Naive causal understanding of valenced behaviors and its implications for social information processing. *Psychological Bulletin, 128*, 421-441.

Ybarra, O., Chan, E., & Park, D. (2001). Young and old adults' concerns about morality and competence. Motivation and Emotion, 25, 85-100.

Yik, M. S. M., Bond, M. H., Paulhus, D. L. (1998). Do Chinese self-enhance or self-efface? It's a matter of domain. *Personality and Social Psychology Bulletin, 24*, 399-406.

Table 1: Protective Biases in Perception, Attention, and Learning

Domain	**False Positive (FP)**	**Cost of FP**	**False Negative (FN)**	**Cost of FN**	**Result**
Approaching sounds	Ready too early	Low	Struck by source	High	Auditory Looming: Bias towards underestimating time to arrival
Dangerous animals (e.g., snakes and spiders)	Fear harmless snakes and spiders	Low	Fail to fear venomous snakes and spiders	High	Easily elicited fear reaction to snakes and spiders
Dangerous persons	Fear harmless people	May be low, depending on the relationship	Fail to fear truly hostile others	High	Easily elicited fear and/or inferences of dangerousness
Food aversions	Avoid a food that is usually harmless	Non-zero but not too high	Eat a fatally toxic food	High	Avoidance of any food that may be associated with sickness
Diseased persons	Avoid a person who is not infectious	May be low, depending on the relationship	Become infected	Often high	Tendency to avoid persons with physical afflictions

Table 2: Biases in Social and Self Perception

Domain	False positive	Cost of FP	False negative	Cost of FN	Result
Unexplained changes in environment	Assume human agency	Vigilance against a conspecific that does not exist—low	Fail to detect competing or hostile group or individual	Suffer displacement competition or hostility—high	Illusion of animacy; agency bias
Sinister attribution or response to social scrutiny	Assume negative evaluation where there is none	Impairs social networks—could be significant	Fail to detect genuine negative evaluations	High if insecure or marginal within social network	Paranoid cognition in situations of marginality or low status; negative forgiveness bias
Dispositional Inference	Assume negative, enduring disposition	Lost opportunity for social exchange– could be significant	Fail to detect harmful, manipulative dispositions in others	High for certain negative traits	Fundamental attribution error for uncertain, negative traits (e.g., social cheating)
Cooperation with others	Believe one can defect or cheat without negative consequences	High—especially when costs of ostracism are great	Infer that one should cooperate even though one could safely defect or cheat (e.g., without detection by others)	Low—especially when resource amount given is small (e.g., small dollar amount)	Social exchange bias: tendency to cooperate when defection has greater payoff
Men's perception of women's sexual interest	Inferring interest where there is none	Wasted courtship effort—relatively low	Inferring no interest when there is	Missed reproductive opportunity—high	Overperception of women's sexual interest by men
Women's perception of men's commitment	Inferring willingness to commit where there is none	Desertion—high	Inferring unwillingness to commit where there is willingness	Delayed start to reproduction—relatively low	Underperception of men's commitment by women
Beliefs in personal control and efficacy	Assuming control or efficacy where there is none	Low as long as the costs of trying and failing are low	Assuming inability to control where control is possible	Passivity and opportunity costs—high	Positive illusions, illusion of control

Figure 1. The optimum threshold (likelihood ratio) for adopting a belief S where the cost of the false positive is fixed at 1 and the cost of the false negative varies, with the probability of s set at 0.1. A threshold less than 1 represents a bias towards adopting S. For explanations see text.

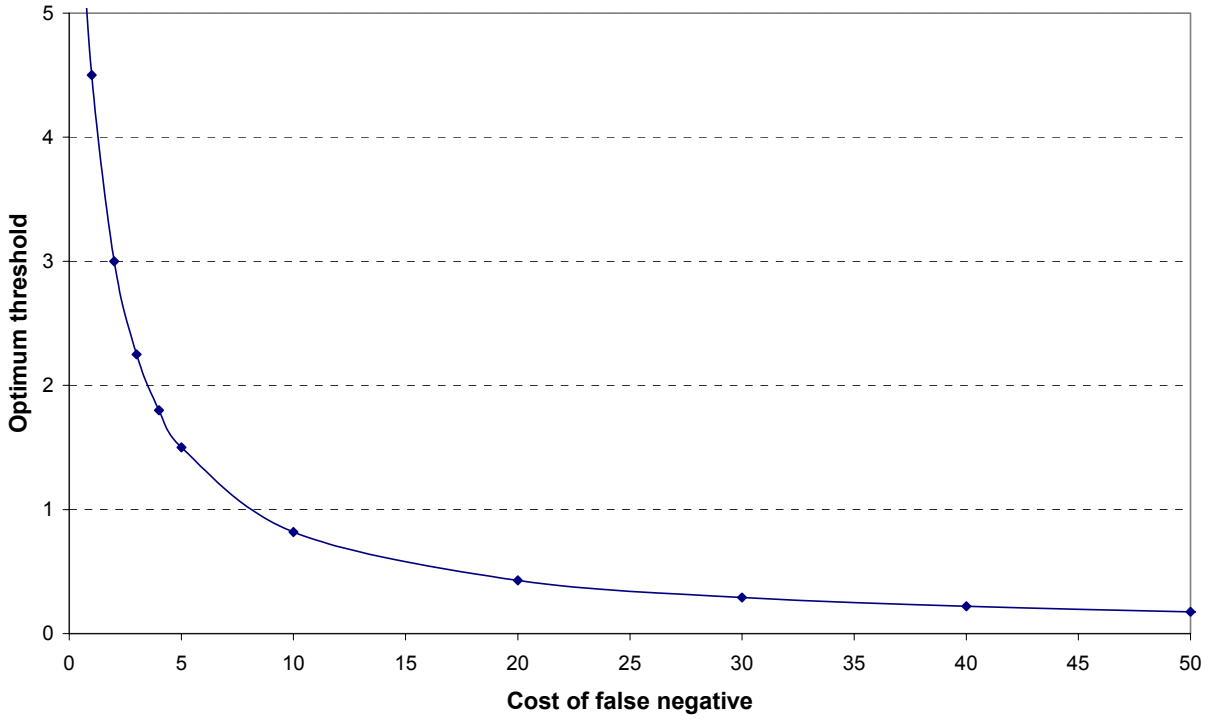

Figure 2. The optimum threshold (likelihood ratio) for adopting a belief S where the cost of the false negative is fixed at 5 and the cost of the false positive varies, with the probability of s set at 0.6. A threshold greater than 1 represents a bias against adopting S. For explanations see text.