To be published in *Behavioral and Brain Sciences* (in press)

The Evolution of Misbelief

Dr. Ryan T. McKay
Institute for Empirical Research in Economics
University of Zurich
Blümlisalpstrasse 10
Zurich 8006
Switzerland
ryantmckay@mac.com
http://homepage.mac.com/ryantmckay/

Prof. Daniel C. Dennett
The Center for Cognitive Studies
Tufts University
Medford MA 02155-7059
ddennett@tufts.edu
http://ase.tufts.edu/cogstud/incbios/dennettd/dennettd.htm

Abstract: From an evolutionary standpoint, a default presumption is that true beliefs are adaptive and misbeliefs maladaptive. But if humans are biologically engineered to appraise the world accurately and to form true beliefs, how are we to explain the routine exceptions to this rule? How can we account for mistaken beliefs, bizarre delusions and instances of self-deception? We explore this question in some detail. We begin by articulating a distinction between two general types of misbelief: those resulting from a breakdown in the normal functioning of the belief formation system (e.g. delusions) and those arising in the normal course of that system's operations (e.g. beliefs based on incomplete or inaccurate information). The former are instances of biological dysfunction or pathology, reflecting "culpable" limitations of evolutionary design. Although the latter category includes undesirable (but tolerable) by-products of "forgivably" limited design, our quarry is a contentious subclass of this category: misbeliefs best conceived as design features. Such misbeliefs, unlike occasional lucky falsehoods, would have been systematically adaptive in the evolutionary past. Such misbeliefs, furthermore, would not be reducible to judicious – but doxastically[i] noncommittal - action policies. Finally, such misbeliefs would have been adaptive in themselves, constituting more than mere by-products of adaptively biased misbelief-producing systems. We explore a range of potential candidates for evolved misbelief, and conclude that, of those surveyed, only *positive illusions* meet our criteria.

Keywords: Adaptive, Belief, Delusions, Design, Evolution, Misbelief, Positive illusions, Religion, Self-deception.

1. Introduction

A misbelief is simply a false belief, or at least a belief that is not correct in all particulars. We can see this metaphorically: If truth is a kind of target that we launch our beliefs at, then misbeliefs are to some extent wide of the mark. Of course, there is no philosophical consensus about just what a belief actually is. In what follows we intend to avoid this question, but we offer here the following working definition of belief, general enough to cover most representationalist and dispositional accounts: A belief is a functional state of an organism that implements or embodies that organism's endorsement of a particular state of affairs as actual.[ii] A misbelief, then, is a belief that to some degree departs from actuality, i.e. it is a functional state endorsing a particular state of affairs that happens not to obtain.

A prevailing assumption is that beliefs that maximise the survival of the believer will be those that best approximate reality (Dennett, 1971, 1987; Fodor, 1983, 1986; Millikan, 1984a, 1984b, 1993). Humans are thus assumed to have been biologically engineered to form true beliefs – by evolution. On this assumption, our beliefs about the world (about what is or isn't true) are essentially tools that enable us to act effectively in the world. Moreover, to be reliable, such tools must be produced in us (it is assumed) by systems designed (by evolution) to be truth aiming, and hence (barring miracles) these systems must be designed to generate *grounded* beliefs (a system for generating ungrounded but mostly true beliefs would be an oracle, as impossible as a perpetual motion machine). Grounded beliefs are simply beliefs that are (appropriately) founded on evidence and existing beliefs; Bayes' theorem (Bayes, 1763) specifies the optimal procedure for revising prior beliefs in the light of new evidence (assuming that veridical belief is the goal, and given unlimited time and computational resources; see Gigerenzer & Goldstein, 1996). Of course, just as we can have good grounds for believing propositions that turn out to be false, so can ungrounded beliefs be serendipitously true (others arguably lack truth values). To keep our exposition manageable, we will not consider such ungrounded beliefs to be misbeliefs, although we acknowledge that false and (serendipitously) true ungrounded beliefs (and perhaps those lacking truth values) may well be produced in much the same way – and by much the same types of mechanism (we return to this issue in section 14).

If evolution has designed us to appraise the world accurately and to form true beliefs, how are we to account for the routine exceptions to this rule - instances of misbelief?

Most of us at times believe propositions that end up being disproved, many of us produce beliefs that others consider obviously false to begin with, and some of us form beliefs that are not just manifestly but bizarrely false. How can this be? Are all these misbeliefs just accidents, instances of pathology or breakdown, or at best undesirable (but tolerable) by-products? Might some of them, contra the default presumption, be adaptive in and of themselves?[iii]

Before we can answer that, we must develop a tentative taxonomy of misbelief. We begin with a distinction between two general types: those that result from some kind of break in the normal functioning of the belief formation system and those that arise in the normal course of that system's operations. We take this to represent the orthodox (albeit unarticulated) view of misbelief. Part and parcel of this orthodox view is that irrespective of whether misbeliefs arise out of the normal or abnormal operation of the belief formation system, the misbeliefs *themselves* are maladaptive.

Our aim in this paper is to evaluate this claim. We will proceed by a process of elimination, considering and disqualifying various candidates until we arrive at what we argue are bona fide instances of adaptive misbelief. Some candidates will prove not to be directly adaptive, the falsity of others will not be obvious, and still others will be rejected on the grounds that they are not, in fact, beliefs. The process will highlight the theoretically important differences between the phenomena, which are interesting in their own right, and will clarify the hypothesis defended – that a subset of the misbeliefs that arise in the normal course of belief formation system operations are, in and of themselves, adaptive. But first we need to refine the distinction between abnormal functioning and normal functioning, as ambiguity on this topic has bedevilled the literature.

2. Manufacture and malfunction

First consider the domain of systems designed and manufactured by humans. Here we envisage the distinction as one (codified in warranty legislation) between "culpable design limitations" (=malfunctions) and "forgivable design limitations/features". When a given artifact fails to perform a particular task, this failure is always due to a limitation in the design of that artifact. The question is whether the design limitation concerned is – from the designer's perspective – a blameworthy, "culpable" limitation (a design flaw, or perhaps a flaw in the execution of the design), or whether it is a tolerable, "forgivable"

limitation. Examples of the former (with respect to the arbitrary task of "keeping time") include:

1) My $20,000 Bolex watch loses ten seconds every day (contra the advertisement).
2) My cheap Schmasio watch loses ten minutes every day (contra the advertisement).

Examples of the latter limitation include:

3) My toaster does not keep time at all.
4) My Bolex loses a second every day (within warranted limits).
5) My cheap Schmasio loses a minute every day (within warranted limits).
6) After putting it in a very hot oven for an hour, my Bolex does not keep time at all.

What we can see from these examples is that manufactured artifacts either work as intended (within a tolerable margin of error), or they don't work as intended (falling outside the tolerable margin). What's important is not how well the artifacts objectively work, but how well they work relative to how they were intended to work (and the intentions of the manufacturer will bear upon the advertised claims of the manufacturer). The Bolex and Schmasio examples reflect this, because the malfunctioning Bolex still works objectively better than the properly functioning Schmasio.

Of course, some apparent design limitations are in fact deliberate *design features*. To cite a single example, contemporary consumers are frequently frustrated by DVD region code restrictions. The fact that a region 1 DVD player (sold in North America) cannot play discs sold in Europe or Japan (region 2) is certainly (from the consumer's perspective at least[iv]) a limitation in the design of that DVD player – and often a frustrating limitation. In our terminology, however, the limitation is *forgivable* because such players are not designed to play DVDs from other regions, and indeed are deliberately designed *not* to do so. Region restrictions are, as software designers often say, "not a bug but a feature" of such machines, ostensibly to safeguard copyright and film distribution rights.

The essential lesson is that a manufactured artifact functions properly if it functions as its designer intended (and warranted) it to function, under the conditions in which it was intended (and warranted) to function. If the artifact fails to function under those conditions, then it has *malfunctioned*, which may be due to a flaw in the design or to a flaw in the execution of the design. Here "malfunction" is equated with "culpable design limitation" and is defined so as to exclude seeming breaks in function that occur outside the constraints specified by the manufacturer (i.e. if a watch falls to pieces a day after the

warranty expires, this is not a malfunction – not on our definition of malfunction, anyway – but a forgivable limitation).

Consider another example: Imagine a computer that is equipped with software for solving physics problems. The computer takes the problems as input, and produces purported solutions to the problems as output. Suppose, further, that the program that the computer implements when solving the problems utilizes Newtonian physics. Consider then three different possible scenarios:

1) The computer is assigned a problem about an apple falling from a tree on Earth. It produces the correct solution.
2) The computer is assigned a problem about an apple falling from a tree on Earth. Unfortunately, a low-level glitch occurs (a flaw in the execution of the program's design), causing the program to malfunction and to produce an incorrect solution.
3) The computer is assigned a problem about the mass of an apple as it approaches the speed of light. The program runs smoothly and predictably, but arrives at an incorrect solution.

Do the second and third scenarios here map onto the distinction between culpable and forgivable design limitations? Whether this is the case depends on the precise intentions of the program designer. If the designer had implemented a Newtonian program because it was easier and cheaper to do so, but was fully aware that Einsteinian problems would compute incorrectly, then the third limitation is forgivable, if it was so advertised. If, however, the designer intended his or her program to solve physics problems of all types, then this limitation is culpable (and constitutes a "malfunction", in this rather peculiar sense of the word).

Even such a common artifact as an electronic hand calculator produces output that may appear culpable:

> For instance, arithmetic tells us that 10 divided by 3 multiplied by 3 is 10, but hand calculators will tell you that it is 9.999999, owing to round-off or truncation error, a shortcoming the designers have decided to live with, even though such errors are extremely destructive under many conditions in larger systems that do not have the benefit of human observer/users (or very smart homunculi!) to notice and correct them. (Dennett, 1998, p. 315)

A manufactured object (or feature thereof) works as a model of adaptive misbelief if: 1) The object is a specific focus of deliberate design (not a mistake or a by-product); 2) The object appears, from a certain perspective, to be malfunctioning or limited insofar as it misrepresents information to the consumer of that information; and 3) Such

misrepresentation is actually beneficial to the consumer of that information. None of the cases of artifacts considered thus far would qualify as analogues of adaptive misbelief under these criteria, but here is one case that gets close: the automotive mirror that is designed such that objects appear farther away than they really are. That this is misrepresentation is made clear by the appended cautionary subtitle (required, no doubt, by the manufacturer's lawyers): "OBJECTS IN MIRROR ARE CLOSER THAN THEY APPEAR." The trade-off in the goal of this design is clear: to provide a wider field of view than a "veridical" mirror, which is deemed a benefit that outweighs the distortion, a cost that is diminished, presumably, by the attached warning. The reason this example ultimately fails as a model of adaptive misbelief is that the misrepresentation itself is not the specific focus of design, nor is it (in and of itself) beneficial to the consumer; rather, the misrepresentation is an unavoidable by-product of producing a wider field of view.

We don't know of other good candidates but can describe a possible device with a similar design rationale: an alarm clock designed to set itself ten minutes ahead in the middle of the night (and then to repair its "error" later in the day). Its design rationale would be to give its owner a little extra time to get going, but once the user figured this out, the device would of course lose effectiveness - a case of "the boy who cried wolf", a design complication that we will discuss in some detail below. Before we move on, we note that whereas artifacts designed to misrepresent information to their consumers may not exactly be thick on the ground,[v] there are certainly artifacts – such as shear pins and fuses - that are designed to *break*. In due course we will consider whether cognitive systems have evolved any parallel.

3. Evolutionary design and dysfunction

Commercial disputes notwithstanding, the distinction between abnormal and normal functioning seems intuitive enough in the case of systems designed and manufactured by humans. How neatly, however, does this distinction carve nature at the joints? Is it equally clear for evolved, biological systems? In such cases, our criterion for determining malfunction (the disparity between actual functioning and intended functioning) would seem invalid, because (unlike the good people at Bolex) evolution is a blind watchmaker (Dawkins, 1986), without intentions. What we would like here is some way of making a distinction that is equivalent to the distinction between culpable design limitations and forgivable design limitations/features. Whereas culpable misdesign in manufactured items is the essence of artifactual malfunction, the evolutionary equivalent would be the marker of biological *dysfunction*.

Consider the human immune system. What would count as an example of immune system dysfunction? Presumably if the immune system were to succumb to a run-of-the-mill pathogen, we could speak uncontroversially of immune dysfunction. In some instances, however, the immune system "errs" in attempting to defend the body. Thus one of the main problems in organ transplants is that the immune system tries to protect the body against foreign matter, even a new heart that would ensure its survival. Is the activity of the immune system in the latter case strictly in accordance with its normal function? Perhaps that depends on what function we choose to impose upon the system. Insofar as the system functions to attack foreign matter, it has performed well. Insofar as the system is construed with the more general function of safeguarding the health of the body, however, it becomes less clear whether it has functioned normally – and this is the problem of evolutionary intentions-by-proxy.[vi] Is all functionality just in the eye of the beholder? Millikan (1984a, 1993) proposes a more objective solution to this problem:

> Associated with each of the proper functions that an organ or system has is a Normal explanation for performance of this function, which tells how that organ or system… historically managed to perform that function. (1993, p. 243)

According to Millikan, in order to determine the function of an organ or system we should consider not its present properties, powers and dispositions, but should instead take into account its history.[vii] Given that organ transplants have not featured in the evolutionary history of immune systems, any contemporary immune system that attacks a donor heart is functioning in accordance with the adaptive functioning of immune systems historically. That system, therefore, is functioning normally – or more precisely, *Normally* (see below) – and its limitations are "forgivable".

Let us consider a further parallel with our proposed misbelief taxonomy, this time by examining two types of mis*perception*. Those of us who are short sighted perceive (without our corrective lenses) a somewhat distorted visual world. Due to a kind of breakdown or degeneration, our visual systems (broadly construed) misrepresent the facts – they cease to function properly. Consider, on the other hand, what happens when we – with eyeglasses at the ready - submerge a perfectly straight stick into a pool of water. Do we continue to perceive the stick as straight and unbroken? No – our visual systems fail to compensate for the optical effect of refraction (they do not compute and correct for Snell's law; Boden, 1984; Casperson, 1999),[viii] and the stick appears bent at the point where it meets the surface of the water. Our visual systems have again furnished us with misinformation, yet this time they have functioned Normally. The capital "N" here

denotes a normative, rather than statistical, construal of "normal" (Millikan, 1984a, 1993).

This is important because although our two examples of visual misperception (the short-sighted case and the stick-in-water case) can be distinguished on normative grounds (the first – being a case of visual dysfunction - is abNormal and the second Normal), they may *both* be normal on statistical ("small-n") grounds. After all, the prevalence of myopia varies across ethnic groups, and is as high as 70-90% in some Asian populations (Chow, Dhillon, Chew, & Chew, 1990; Wong et al., 2000). Millikan (1993), however, dismisses statistical construals of "normal" functioning. In a vivid example she points out that the proper function of sperm is to fertilise an ovum, notwithstanding the fact that, statistically speaking, it is exceedingly unlikely that any individual sperm will successfully perform that function (Millikan, 1984a). Proper, *Normal* functioning, therefore, is not what happens always or even on the average; sometimes it is positively rare. Unless otherwise indicated, our subsequent usage of "normal" will follow Millikan's capitalised, normative sense.

Now, back to beliefs and misbeliefs. We contend that all instances of misbelief can be roughly classified as the output of either a dysfunctional, abnormal belief formation system or of a properly functioning, normal belief formation system. The former category, to which we turn briefly now, would not include adaptive misbeliefs (although see section 10 below), but provides a necessary background for understanding the better candidates – which, if they exist, will form a subset (design *features*) of the latter category.

4. Doxastic dysfunction

In the first category, misbeliefs result from breakdowns in the machinery of belief formation. If we conceive of the belief formation system as an information processing system that takes certain inputs (e.g. perceptual inputs) and (via manipulations of these inputs) produces certain outputs (beliefs, e.g. beliefs about the environment that the perceptual apparatus is directed upon), then these misbeliefs arise from dysfunction in the system – doxastic dysfunction. Such misbeliefs are the faulty output of a disordered, defective, abnormal cognitive system.

This view of misbelief is prominently exemplified by a branch of cognitive psychology known as cognitive neuropsychiatry (David & Halligan, 1996). Cognitive

neuropsychiatrists apply the logic of cognitive neuropsycho*logy*, which investigates disordered cognition in order to learn more about normal cognition, to disorders of high-level cognition such as delusions (Coltheart, 2002; Ellis & Young, 1988). Notwithstanding objections to the so-called "doxastic conception" of delusions (see Section 9), delusions are misbeliefs *par excellence* - false beliefs that are held with strong conviction regardless of counter-evidence and despite the efforts of others to dissuade the deluded individual (American Psychiatric Association, 2000). They are first-rank symptoms of schizophrenia and prominent features of numerous other psychiatric and neurological conditions. Thematically speaking, delusions range from the bizarre and exotic (e.g. "I am the Emperor of Antarctica"; see David, 1999) to the more mundane and ordinary (e.g. "My husband is cheating on me"). Researchers in cognitive neuropsychiatry aim to develop a model of the processes involved in normal belief generation and evaluation, and to explain delusions in terms of damage to one or more of these processes.

To illustrate the cognitive neuropsychiatric approach to delusion, consider the case of "mirrored-self misidentification". Patients with this rare delusion misidentify their own reflected image, and may come to believe that a stranger is following them around. Breen, Caine and Coltheart (2001) investigated two cases of this delusion and uncovered two apparent routes to its development. The delusion of the first patient ("FE") appeared to be underpinned by anomalous face perception ("prosopagnosia"), as he demonstrated a marked deficit in face processing on neuropsychological tests. In contrast, the face processing of the second patient ("TH") was intact. This patient, however, appeared to be "mirror agnosic" (Ramachandran, Altschuler, & Hillyer, 1997), in that he evinced an impaired appreciation of mirror spatial relations and was unable to interact appropriately with mirrors. His delusion appeared to be underpinned by anomalous processing not of faces, but of reflected space (see Breen, Caine, Coltheart, Hendy and Roberts, 2000, for transcripts of interviews with the two patients; see Feinberg, 2001, and Feinberg & Shapiro, 1989, for descriptions of related cases).

An important question arising at this point is the question of whether prosopagnosia (or mirror agnosia) is a sufficient condition for the mirror delusion. The answer to this question is almost certainly No. Other cases of mirror agnosia have been reported without any accompanying misidentification syndrome (Binkofski, Buccino, Dohle, Seitz & Freund, 1999), and non-delusional prosopagnosia is quite common. Breen, Caine and Coltheart (2001) thus proposed that the delusion of mirrored-self misidentification results from the conjunction of *two* cognitive deficits, the first of which gives rise to some

anomalous perceptual data (data concerning either faces or reflected space), and the second of which allows the individual to accept a highly implausible hypothesis explaining these data. The first deficit accounts for the content of the delusion (the fact that it concerns a stranger in the mirror), while the second deficit accounts for why the stranger-in-the-mirror belief, once generated, is then adopted and maintained in the absence of appropriate evidence for that hypothesis. These deficits constitute breakdowns in the belief formation system, presumably underpinned by neuroanatomical or neurophysiological abnormalities. In both of the cases investigated by Breen et al. (2001), the mirror delusion occurred in the context of a progressive dementing illness.

Coltheart and colleagues (Coltheart, Menzies & Sutton, forthcoming; Davies & Coltheart, 2000; Davies, Coltheart, Langdon, & Breen, 2001; Langdon & Coltheart, 2000; McKay, Langdon, & Coltheart, 2007a, 2009) have suggested that a generalised framework of two concurrent cognitive deficits, or factors, might be used to explain delusions of many different types. In general, the first factor (*Factor-1*) is some endogenously generated abnormal data to which the individual is exposed. In addition to mirrored-self misidentification, *Factors-1* have been identified or hypothesised that plausibly account for the content of delusions such as thought insertion, Capgras delusion (the belief that a loved one has been replaced by an impostor) and Cotard delusion (the belief that one is dead).

The second factor (*Factor-2*), on the other hand, can be characterised as a dysfunctional departure from Bayesian belief revision (Coltheart, Menzies & Sutton, forthcoming), a departure that affects how beliefs are revised in the light of the abnormal *Factor-1* data. Bayes' theorem is in a sense a prescription for navigating a course between excessive tendencies toward "observational adequacy" (whereby new data is over-accommodated) and "doxastic conservatism" (whereby existing beliefs are over-weighted) (Stone & Young, 1997). McKay, Langdon and Coltheart (2009) have suggested that whereas some delusions – for example, mirrored-self misidentification - might involve the former tendency (see Stone and Young, 1997; Langdon & Coltheart, 2000; Langdon, Cooper, Connaughton, & Martin, 2006; also see Huq, Garety & Hemsley, 1988), others – for example, delusional denial of paralysis ("anosognosia") - might involve the latter (see Ramachandran, 1994a,b; 1995; 1996a,b; Ramachandran & Blakeslee, 1998). In general, therefore, *Factor-2* might be thought of as an acquired or congenital anomaly yielding one of two dysfunctional doxastic biases – a bias toward observational adequacy or toward doxastic conservatism.

The fact that we are not presently equipped with fail-safe belief-formation systems does not tell against an evolutionary perspective. This is because evolution does not necessarily produce optimally designed systems (Dawkins, 1982; Stich, 1990) and in fact often conspicuously fails to do so. It would be panglossian to think otherwise (Gould & Lewontin, 1979; Voltaire, 1759/1962):

> Brilliant as the design of the eye is, it betrays its origin with a tell-tale flaw: the retina is inside out… No intelligent designer would put such a clumsy arrangement in a camcorder. (Dennett, 2005, p. 11)

Evolutionary explorations in Design Space are constrained, among other things, by economic considerations (beyond a certain level, system improvements may exhibit declining marginal utility; Stich, 1990), historical vicissitude (the appropriate mutations must occur if selection is to act on them) and the topography of the fitness landscape (selection cannot access optimal design solutions if it must traverse a fitness valley to do so; Dennett, 1995a). Because evolution is an imperfect design process, the systems we have evolved for representing reality are bound to be limited – and sometimes they will break.

5. Misbeliefs as the output of a properly functioning system

Even if evolution were in some sense a "perfect" design process, there would still be limitations; only a violation of the laws of physics would permit, say, beliefs to be formed instantaneously, with no time lag whatsoever, or for an individual, finite believer to carry around in her head beliefs about the lack of prime factors of each specific prime number (only a brain of infinite volume could represent each individual prime).[ix] The result is that even the beliefs of textbook Bayesians will frequently be false (or at least incomplete) – and such misbeliefs cannot be considered "culpable".

Perhaps the most obvious examples of commonplace, forgivable misbelief occur when we are victimised by liars. Although extreme gullibility might be seen as dysfunctional (perhaps involving a *Factor-2* bias toward observational adequacy), most of us (Bayesians included) are vulnerable to carefully crafted and disseminated falsehood. However adaptive it may be for us to believe truly, it may be adaptive for *other* parties if we believe falsely (Wallace, 1973).[x] An evolutionary arms race of deceptive ploys and counterploys may thus ensue. In some cases the "other parties" in question may not even be animate agents, but cultural traits or systems (Dawkins, 2006a,b; Dennett, 1995a, 2006a). Although such cases are interesting in their own right, the adaptive misbeliefs we

pursue in this paper are beneficial to their consumers - misbeliefs that evolve to the detriment of their believers are not our quarries.

So, given inevitable contexts of imperfect information, even lightning-fast Bayesians will frequently misbelieve, and such misbeliefs must be deemed forgivable. We briefly consider now whether certain departures from Bayesian updating might also be considered forgivable. Gigerenzer and colleagues (e.g. Gigerenzer & Goldstein, 1996; Gigerenzer, Todd et al., 1999) have argued that some such departures, far from being defective, comprise "ecologically rational" decision strategies that operate effectively given inevitable limitations of time and computational resources. These researchers have documented and investigated a series of such "fast and frugal" heuristics, including the "take the best" heuristic (Gigerenzer & Goldstein, 1999) and the "recognition heuristic" (Goldstein & Gigerenzer, 2002).

Some departures from normative rationality standards, however, result from perturbations in belief formation machinery and are not "heuristic" in any sense. As we have noted, bizarre misbeliefs like mirrored-self misidentification and Cotard delusion may occur subsequent to neuropsychological damage. For example, Young, Robertson, Hellawell, de Pauw & Pentland (1992) described a patient who was injured in a serious motorcycle accident and subsequently became convinced that he was dead. Computerised tomography (CT) scans revealed contusions affecting temporo-parietal areas of this patient's right hemisphere as well as some bilateral damage to his frontal lobe. Misbeliefs, however, may also arise from less acute disruptions to the machinery of belief formation. For example, lapses in concentration due to fatigue or inebriation may result in individuals coming to hold erroneous beliefs, at least temporarily. Are such misbeliefs "culpable"? Do they reflect dysfunction in the belief formation system?

Although misbelief might always reflect the limitations of the system in some sense, it is not always easy to tell (absent a warranty) where imperfect proper doxastic functioning (forgivably limited) ends and where (culpably limited) doxastic dysfunction begins. This fuzziness is reflected in the literature on certain putative psychological disorders. As an example, consider the phenomenon of disordered reading. There are debates in the literature about whether there is a separate category of individuals who are disordered readers (e.g. see Coltheart, 1996). Opponents of this view argue that so-called "disordered readers" are just readers at the lower end of a Gaussian distribution of reading ability. Similarly, one of the most controversial psychiatric diagnoses in recent years has been the diagnosis of Attention-Deficit Hyperactivity Disorder (ADHD), which

some commentators insist is a figment, arguing that putatively ADHD children are just children at the extreme ends of Gaussian distributions of attention and activity (for a discussion, see Dennett, 1990a).

Controversies such as these are difficult to resolve. While we consider that Millikan's distinction between Normal and abNormal functioning provides a useful rule of thumb, we are not confident that this distinction – or *any* distinction – can be used to decisively settle disputes about forgivable versus culpable limitations in the biological domain. In this domain these categories are not discrete, but overlapping. Culpable misdesign in nature is always ephemeral - where design anomalies are rare or relatively benign, we will observe "tolerated" (forgivable) limitations; where anomalies begin to proliferate, however, they raise the selection pressure for a design revision, leading to either adaptive redesign or extinction. The upshot is that it may be difficult, if not impossible, to adjudicate on intermediate cases. How fatigued does an individual actually need to be before his doxastic lapses are deemed (evolutionarily) forgivable? And if alcohol did not feature in the evolutionary history of the belief formation system, are false beliefs formed while tipsy forgivable? Perhaps dousing one's brain in alcohol is akin to baking one's Bolex in a hot oven – both are forced to labour "under *external* conditions not Normal for performance of their proper functions" (Millikan, 1993, p.74, emphasis in original).

We acknowledge the overlap between our two broad categories of functioning. Such overlaps, however, characterise most biological categories: the boundaries - between, for example, species, or territories, or even between life and death - are porous and often violated. In any case, establishing a means of settling disputes about forgivable versus culpable limitations of the belief formation system is not crucial to our project. Although it is useful to be able to distinguish, crudely, between normal and abnormal doxastic functioning, the prevailing view is that misbeliefs formed in either case will themselves be abnormal. We will now begin to question this assumption. Contra the prevailing view, might there be certain situations in which misbelief can actually be adaptive (situations in which the misbeliefs themselves, not just the systems that produce them, are normal)? In those situations, if such there be, we would expect that we would be evolutionarily predisposed to form some misbeliefs. In short, *misbelief would evolve*.

6. Adaptive misbelief?

O who can hold a fire in his hand
By thinking on the frosty Caucasus?

Or cloy the hungry edge of appetite
By bare imagination of a feast?
Or wallow naked in December snow
By thinking on fantastic summer's heat?

~ Shakespeare (Richard II 1.3. 294-303)

How does religion fit into a mind that one might have thought was designed to reject the palpably not true? The common answer—that people take comfort in the thought of a benevolent shepherd, a universal plan, of an afterlife—is unsatisfying, because it only raises the question of *why* a mind would evolve to find comfort in beliefs it can plainly see are false. A freezing person finds no comfort in believing he is warm; a person face-to-face-with a lion is not put at ease by the conviction that it is a rabbit.

~ Pinker (1997, pp. 554-5, emphasis in original)

We are anything but a mechanism set up to perceive the truth for its own sake. Rather, we have evolved a nervous system that acts in the interest of our gonads, and one attuned to the demands of reproductive competition. If fools are more prolific than wise men, then to that degree folly will be favored by selection. And if ignorance aids in obtaining a mate, then men and women will tend to be ignorant.

~ Ghiselin (1974, p. 126)

How could it ever be beneficial to believe a falsehood? Granted, one can easily imagine that in many circumstances it might *feel* better to misbelieve (more on this in section 10). Thus in *Richard II*, Bolingbroke, who has been banished, is urged by his father to imagine that he is not banished but rather has left of his own volition. Bolingbroke's father appreciates that there may be psychological comfort in such a false belief. Bolingbroke's reply, however ("O who can hold a fire in his hand…"), speaks both to the difficulty of deliberately misbelieving as well as to the apparent absence of tangible benefits in thus misbelieving. How could misbelief aid survival?

We note that it is easy to dream up anomalous offbeat scenarios where true beliefs are in fact detrimental for survival:

> [Harry] believed that his flight left at 7:45am… Harry's belief was true, and he got to the airport just on time. Unfortunately, the flight crashed, and Harry died. Had Harry falsely believed that the flight left at 8:45, he would have missed the flight and survived. So true belief is sometimes less conducive to survival than false belief. (Stich, 1990, p. 123)

As Stich (1990) notes, cases such as this are highly unusual, and do little to obviate the claim that true beliefs are generally adaptive (see also Millikan, 1993). After all, natural

selection does not act on anomalous particulars, but rather upon reliable generalizations. Our question, then, is whether there might be cases where misbelief is *systematically* adaptive.

7. The boy who cried wolf

You've outdone yourself—as usual!

> ~ *Raymond Smullyan*

Theoretical considerations converging from several different research traditions suggest that any such systematic falsehood must be unstable, yielding ephemeral instances, at best, of misbelief. Recognition of the problem is as old as Aesop's fable of the boy who cried wolf. Human communication between agents with memories and the capacity to check on the reliability of informants creates a dynamical situation in which systematic lying eventually exposes and discredits itself. As Quine (1960), Davidson (1994, 2001), Millikan (2004) and other philosophers have noted, without a prevailing background of truth-telling, communication will erode, a practice that cannot pay for itself. That does not mean, of course, that individual liars will never succeed for long, but just that their success depends on their being rare and hard to track. A parallel phenomenon in evolutionary biology is Batesian mimicry, in which a non-poisonous species (or type within a species) mimics the appearance of a poisonous species (telling a falsehood about itself), getting protection against predators without paying for the venom. When mimics are rare, predators avoid them, having had more encounters with the poisonous variety; when mimics are common, the mimicry no longer works as well.

Quine and Ullian (1978) note an important wrinkle:

> If we could count on people to lie most of the time, we could get all the information from their testimony that we get under the present system [of predominant truth-telling]. We could even construe all their statements as containing an understood and unspoken 'not', and hence as predominantly true after all. Utterly random veracity, however, meshed with random mendacity, would render language useless for gathering information. (p. 52)

Isolated cases of the tacit negation suggested in this passage actually occur, when what might be called systematic irony erodes itself with repetition. "Terrific" no longer means "provoking terror" but almost its opposite, and if somebody calls your lecture "incredible" and "fantastic", you should not take offence; they almost certainly don't

mean that they don't believe a word of it and deem it to be out of touch with reality. A related phenomenon is "grade inflation" in academia. "B+" just doesn't mean today what it used to mean several decades ago. When everybody is declared "better than average" the terms of the declaration are perforce diminished in meaning or credibility or both.

What, if anything, would prevent similar accommodations from diluting the effect of systematic falsehoods within the belief formation system of an individual organism? We know from many experiments with subjects wearing inverting or distorting lenses (for a recent summary see Noë, 2004) that the falsehoods the eyes send the brain lead initially to false beliefs that seriously disable the subject, but in remarkably short time - a few days of accommodation - subjects have made an adjustment and can "get all the information from their testimony," as Quine and Ullian (1978) say, just as if they had inserted a tacit "not" or switched the meaning of "right" and "left" in the visual system's vocabulary. For a *systematic* falsehood-generating organ or tissue or network to have any staying power, it must send its lies to something that has little or no source memory or little or no plasticity in its evaluation of the credibility of the source.

Something like that may well be the case in some sensory systems. Akins (1996) discusses "narcissistic" biases built into sensory systems in order to optimize relevance and utility for the animal's behavioural needs. Instead of being designed to have their output states vary in unison (linearly) with the input conditions they are detecting (like thermometers or fuel gauges, which are designed to give objectively accurate measurements), these are designed to "distort" their responses (rather like the rear view mirror). She notes: "when a sensory system uses a narcissistic strategy to encode information, there need not be any counteracting system that has the task of decoding the output state" (p. 359). No "critics" or "lie detectors" devalue the message, and so the whole organism lives with a benign illusion of properties in the world that "just happen" to be tailor-made for its discernment. For instance, feedback from muscle stretch receptors needs to be discriminating over several orders of magnitude, so the "meaning" of the spike trains varies continuously over the range, the sensitivity being adjusted as need be to maintain fine-grained information over the whole range. "What is important to realize, here, is that there need not be any further device that records the 'position' of the gain mechanism." (p. 362). In other words, no provision is made for reality-checking on what the stretch-receptors are "telling" the rest of the system, but the effect of this is to permit "inflation" to change the meaning of the spike frequency continuously.

Here, then, are two distinct ways in which our nervous systems can gracefully adjust the use to which they put signals that one would brand as false were it not for the adjustment. In the phenomena induced by artificially distorting the sensory input, we can observe the adjustment over time, with tell-tale behavioural errors and awkwardness giving way to quite effective and apparently effortless responses as the new meanings of the input signals get established. In the sort of cases Akins discusses, there is no precedent, no "traditional meaning," to overcome, so there is no conflict to observe.

8. Alief and belief

Sometimes, however, the conflicts are not so readily resolved and the inconsistencies in behaviour do not evaporate. Gendler (2008) notes the need for a category of quasi-beliefs and proposes to distinguish between *alief* and *belief*:

> Paradigmatic alief can be characterized as a mental state with associatively-linked content that is representational, affective and behavioral, and that is activated – consciously or unconsciously – by features of the subject's internal or ambient environment. Alief is a more primitive state than either belief or imagination: it directly activates behavioral response patterns (as opposed to motivating in conjunction with desire or pretended desire.) (Gendler, abstract)

A person who trembles (or worse) when standing on the glass-floored Skywalk that protrudes over the Grand Canyon does not believe she is in danger, any more than a moviegoer at a horror film does, but her behaviour at the time indicates that she is in a belief-like state that has considerable behavioural impact. The reluctance of subjects in Paul Rozin's (Rozin, Millman, & Nemeroff, 1986) experiments with disgust to come in contact with perfectly clean but disgusting looking objects does not indicate that they actually believe the objects are contaminated; in Gendler's terms, they *alieve* this. In a similar vein, patients with Obsessive Compulsive Disorder generally don't *believe* that the repetitive behaviours they feel compelled to engage in are necessary to prevent some dreaded occurrence – but they may well *alieve* this. (The Diagnostic and Statistical Manual of Mental Disorders [DSM-IV-TR; American Psychiatric Association, 2000, p. 463] contains a specifier for OCD with "poor insight", which denotes patients who fail to recognise that their obsessions and compulsions are "excessive or unreasonable". In such patients alief may be overlaid with belief.)

Are such aliefs adaptive? Probably not. They seem to join other instances of "tolerated" side effects of imperfect systems, but in any case they are not beliefs proper. The question before us now is whether we ever evolve systems for engendering false *beliefs*:

informational states of global and relatively enduring (inflation-proof) significance to the whole organism that miss the usual target of truth and do so non-coincidentally.

9. Error Management Theory

[B]elief-formation systems that are maximally accurate (yielding beliefs that most closely approximate external reality) are not necessarily those that maximize the likelihood of survival: natural selection does not care about truth; it cares only about reproductive success.

~ Stich (1990, p. 62)

[T]he human mind shows good design, although it is design for fitness maximization, not truth preservation.

~ Haselton and Nettle (2006, p. 63)

The brain functions so that we *tend to construe as true* what is reproductively efficacious – true or not.

~ Schloss (2006, p. 191, emphasis in original)

Beliefs are notoriously hard to count. Is the belief that 3+1=4 distinct from the belief that 1+3=4 or are these just one belief? Can you have one without the other? (See Dennett, 1982, for an analysis of the problems attendant on such questions.) No matter how we individuate beliefs, we might expect that optimal systems of belief and decision would be maximally accurate. Given the contexts in which decisions are made, however, trade-offs may arise between overall accuracy and accuracy in certain situations. Dennett illustrates this point:

> [I]t might be better for beast *B* to have some false beliefs about whom *B* can beat up and whom *B* can't. Ranking *B*'s likely antagonists from ferocious to pushover, we certainly want *B* to believe it can't beat up all the ferocious ones and can beat up all the obvious pushovers, but it is better (because it "costs less" in discrimination tasks and protects against random perturbations such as bad days and lucky blows) for *B* to extend "I can't beat up *x*" to cover even some beasts it can in fact beat up. *Erring on the side of prudence* is a well-recognized good strategy, and so Nature can be expected to have valued it on occasions when it came up. (Dennett, 1987, p. 51, fn. 3, emphasis in original)

Stich echoes the logic of this scenario with an example of his own:

> Consider, for example, the question of whether a certain type of food is poisonous. For an omnivore living in a gastronomically heterogeneous environment, a false positive on such a

> question would be relatively cheap. If the organism comes to believe that something is poisonous when it is not, it will avoid that food unnecessarily. This may have a small negative impact on its chances of survival and successful reproduction. False negatives, on the other hand, are much more costly in such situations. If the organism comes to believe that a given kind of food is not poisonous when it is, it will not avoid the food and will run a substantial risk of illness or death. (1990, pp. 61-62)

What these examples suggest is that when there are reliable "asymmetries in the costs of errors" (Bratman, 1992), i.e. when one type of error (false positive or false negative) is consistently more detrimental to fitness than the other, then a system that is biased toward committing the less costly error may be more adaptive than an unbiased system. The suggestion that biologically engineered systems of decision and belief formation exploit such adaptations is the basis of Error Management Theory (EMT; Haselton, 2007; Haselton & Buss, 2000, 2003; Haselton & Nettle, 2006). According to EMT, cognitive errors (including misbeliefs) are not necessarily malfunctions reflecting (culpable) limitations of evolutionary design; rather, such errors may reflect judicious systematic biases that maximise fitness *despite* increasing overall error rates.

Haselton and Buss (2000) use EMT to explain the apparent tendency of men to overperceive the sexual interest and intent of women (Abbey, 1982; Haselton, 2003). They argue that, for men, the perception of sexual intent in women is a domain characterised by recurrent cost asymmetries, such that the cost of inferring sexual intent where none exists (a false-positive error) is outweighed by the cost of falsely inferring a lack of sexual intent (a false-negative). The former error may cost some time and effort spent in fruitless courtship, but the latter error will entail a missed sexual and thus reproductive opportunity – an altogether more serious outcome as far as fitness is concerned.

For women, the pattern of cost asymmetries is basically reversed. The cost of inferring a man's interest in familial investment where none exists (a false-positive error) would tend to outweigh the cost of falsely inferring a lack of such interest (a false-negative). The former error may entail the woman consenting to sex and being subsequently abandoned, a serious outcome indeed in arduous ancestral environments. The latter error, on the other hand, would tend merely to delay reproduction for the woman – a less costly error, especially given that reproductive opportunities are generally easier for women to acquire than men (Haselton, 2007). In view of such considerations, proponents of EMT predict that women will tend to underperceive the commitment intentions of men, a prediction apparently supported by empirical evidence (Haselton, 2007; Haselton & Buss, 2000).

Other EMT predictions that have received apparent empirical support include the hypotheses that recurrent cost asymmetries have produced evolved biases toward overinferring aggressive intentions in others (Duntley & Buss, 1998; Haselton & Buss, 2000), particularly members of other racial and ethnic groups (Haselton & Nettle, 2006; Krebs and Denton, 1997; Quillian & Pager, 2001); toward overinferring potential danger with regard to snakes (see Haselton & Buss, 2003; Haselton & Nettle, 2006); toward underestimating the arrival time of approaching sound sources (Haselton & Nettle, 2006; Neuhoff, 2001); and - reflecting Stich's (1990) example above – toward overestimating the likelihood that food is contaminated (see Rozin & Fallon, 1987; Rozin, Markwith, & Ross, 1990). The error management perspective, moreover, appears to be a fecund source of new predictions. In the realm of sexuality and courtship, for example, Haselton and Nettle (2006) predict biases toward over-inferring the romantic or sexual interest of a) others in one's partner (what they term the "interloper effect"); and b) one's partner in others. These predictions complement a series of other already confirmed predictions stemming from evolutionary analyses of jealousy (see Buss & Haselton, 2005, for a brief review).

One objection that might be raised at this point is that the above examples need not actually involve *misbelief*. Stich's omnivore need not *believe* that the food in question is poisonous – it might remain quite agnostic on that score. Similarly, jealous individuals need not harbour *beliefs* about partner infidelity – they might just be hypervigilant for any signs of it. The issue here is what doxastic inferences can be drawn from behaviour. After all, we always look before crossing a road, even where we are almost positive that there is no oncoming traffic. Our actions in such a case should not be read as reflecting a belief that there is an oncoming vehicle, but rather as reflecting a belief that there *might* be an oncoming vehicle (and the absence of a vehicle does not render that latter belief false). If we had to bet our lives one way or another on the matter, we might well bet that there isn't an oncoming vehicle (Bratman, 1992). Betting our lives one way or the other, however, is a paradigm case of error symmetry (if we're wrong, we die – no matter which option we choose). In everyday cases of crossing the road, however, the errors are radically asymmetrical – an error one way may indeed mean serious injury or death, but an error the other way will entail only a trivial waste of time and energy.

The upshot of this criticism is that tendencies to "overestimate" the likelihood that food is contaminated, to "overperceive" the sexual interest of women, or to "overinfer" aggressive intentions in others, may reflect judicious decision criteria for action rather

than misbeliefs. Nature may well prefer to create a bias on the side of prudence, but she does not always need to instil erroneous *beliefs* to accomplish this. She may instead make do with cautious action *policies* that might be expressed as "when in doubt [regarding some state of affairs relevant to current welfare], do *x*." Errors, therefore, may not need to be managed doxastically (see McKay & Efferson, in preparation, for a thorough treatment of these issues). Some authors, however, have suggested that certain *delusions* also involve error management processes. Schipper, Easton and Shackelford (2007), for example, conceptualise delusional jealousy (also known as morbid jealousy or Othello syndrome) as the extreme end of a Gaussian distribution of jealousy, and hypothesise that the same sex-specific patterns that characterise "normal" jealousy - stemming from recurrent divergence in the adaptive problems faced by each gender - will also characterise delusional jealousy: "hypersensitive jealousy mechanisms… may serve the adaptive purpose of preventing partner infidelity" (p. 630; see also Easton, Schipper, & Shackelford, 2007). Whereas it may be true, therefore, that errors are not ordinarily managed doxastically, surely *delusions* involve genuine belief?

There are, however, serious objections to the notion that delusions are beliefs (Hamilton, 2007; Stephens & Graham, 2004; see Bayne & Pacherie, 2005, for a defence of the "doxastic conception"). One objection stems from the observation that although some individuals act on their delusions - and sometimes violently (see Mowat, 1966; Silva, Ferrari, Leong and Penny, 1998) - other deluded individuals frequently fail to act in accordance with their delusions. Individuals with Capgras delusion, for example, rarely file missing persons reports on behalf of their replaced loved ones, and those who claim to be Napoleon are seldom seen issuing orders to their troops (Young, 2000). In response to such objections, some authors have provided characterisations of delusions that dispense with the doxastic stipulation. Jaspers (1913/1963) and Berrios (1991), for example, have each proposed "non-assertoric" accounts of delusions (Young, 1999). Jaspers (1913/1963) held that schizophrenic delusions are not understandable, while for Berrios (1991) the verbalizations of deluded patients are empty speech acts, mere noise masquerading as mentality. Other authors have put forward "metacognitive" accounts of delusions, whereby delusions are conceived as higher order meta-evaluations of standard, lower order mental items. For example, Currie and colleagues (Currie, 2000; Currie & Jureidini, 2001; see Bayne & Pacherie, 2005 for a critique) argue that delusions are in fact imaginings misidentified as beliefs. On this account, the delusional belief of a Cotard patient is not the belief that she is dead, but rather the belief that she *believes* she is dead – when in fact she only imagines that she is dead (see Stephens & Graham, 2004, for a variant of the metacognitive thesis).

In any case, it may be misguided to invoke delusions in attempting to link error management with adaptive misbelief. The reason is simple: even if one overlooks objections to the doxastic conception and insists that delusions *are* beliefs, a serious problem remains – the issue of whether delusions can, in any sense, be regarded as adaptive. We consider this question below.

10. Doxastic shear pins

In this paper we have distinguished two broad categories of misbelief – on the one hand a category of misbeliefs resulting from breaks in the belief formation system, and on the other a category of misbeliefs arising in the normal course of belief system operations. Here we briefly consider an intriguing intermediate possibility: misbeliefs enabled by the action of "doxastic shear pins". A shear pin is a metal pin installed in, say, the drive train of a marine engine, that locks the propeller to the propeller shaft and that is intended to "shear" should the propeller hit a log or other hard object. Shear pins are mechanical analogues of electrical fuses – each is a component in a system that is *designed to break* (in certain circumstances) so as to protect other, more expensive parts of the system. When a shear pin breaks (or a fuse blows), the system ceases its normal function. However, the action of the shear pin or fuse is not itself abnormal in these situations – in fact it is functioning perfectly as designed.

What might count as a doxastic analogue of shear pin breakage? We envision doxastic shear pins as components of belief evaluation machinery that are "designed" to break in situations of extreme psychological stress (analogous to the mechanical overload that breaks a shear pin or the power surge that blows a fuse). Perhaps the normal function (both normatively and statistically construed) of such components would be to constrain the influence of motivational processes on belief formation. Breakage of such components,[xi] therefore, might permit the formation and maintenance of comforting misbeliefs – beliefs that would ordinarily be rejected as ungrounded, but that would facilitate the negotiation of overwhelming circumstances (perhaps by enabling the management of powerful negative emotions) and that would thus be *adaptive* in such extraordinary circumstances.

Insofar as these misbeliefs were delusions, they would have a different aetiology to the more clear-cut cases of "deficit delusions" discussed earlier (mirrored-self misidentification and the like), because the breakage permitting their formation would

serve a defensive, protective function. In short, they would be *motivated* (see Bayne & Fernández, 2009; McKay, forthcoming; McKay, Langdon, & Coltheart, 2007a, 2009). Psychoanalytically inclined authors have proposed motivational interpretations of delusions such as the Capgras and Cotard delusions (e.g. see Enoch & Ball, 2001), but in the wake of more rigorous cognitive neuropsychiatric models such interpretations tend to be viewed with disdain as outlandish and anachronistic (Ellis, 2003).

Claims about motivational aetiologies for delusions are more plausible in other domains, however. Consider, for example, the following case of *reverse* Othello syndrome (Butler, 2000). The patient in question, "BX", was a gifted musician who had been left a quadriplegic following a car accident. BX subsequently developed delusions about the continuing fidelity of his former romantic partner (who had in fact severed all contact with him and embarked on a new relationship soon after his accident). According to Butler, BX's delusional system provided a "defense against depressive overwhelm… [going] some way toward reconferring a sense of meaning to his life experience and reintegrating his shattered sense of self. Without it there was only the stark reality of annihilating loss and confrontation with his own emotional devastation" (2000, p. 89). Although this seems a plausible motivational formulation, this is an isolated case study and Butler's theorising is unavoidably post hoc. Moreover, the fact that BX had sustained severe head injuries in his accident opens up the possibility that any breakage in his belief evaluation system was, as it were, ateleological - adventitious, not designed. More general (plausible) motivational interpretations exist for other delusions, however – especially for so-called "functional" delusions, where the nature and role of underlying neuropathy (if any) is unspecified (Langdon & Coltheart, 2000; Langdon, McKay, & Coltheart, 2008). In particular, there are well worked out motivational formulations for persecutory delusions (see Bentall & Kaney, 1996; Kinderman & Bentall, 1996, 1997), interpretations that have garnered recent empirical support (McKay, Langdon, & Coltheart, 2007b; Moritz, Werner, & von Collani, 2006; although see Vazquez, Diez-Alegria, Hernandez-Lloreda & Moreno, 2008).

It seems, therefore, that certain delusions might serve plausible defensive functions. Whether this implies that such delusions are adaptive, however, is a different question. To be sure, it might plausibly be argued that delusions are *psychologically* adaptive in certain scenarios (as the above reverse Othello case suggests). But this does not establish a case for *biological* adaptation. Here we must be careful to honour a distinction, often complacently ignored, between human happiness and genetic fitness. If the most promising path, on average, to having more surviving grandoffspring is one that involves

pain and hardship, natural selection will not be deterred in the least from pursuing it (it is well to remind ourselves of the insect species in which the males are beheaded in the normal course of copulation, or - somewhat closer to home - the ruthless siblingcide practiced by many bird species). Perhaps the most that can presently be claimed is that delusions may be produced by extreme versions of systems that have evolved in accordance with error management principles, i.e. evolved so as to exploit recurrent cost asymmetries. As extreme versions, however, there is every chance that such systems manage errors in a maladaptive fashion. As Zolotova and Brüne conclude, "[T]he content of delusional beliefs could be interpreted as *pathological variants* of adaptive psychological mechanisms…" (2006, p. 192, our emphasis; see also Brüne, 2001, 2003, in press).

In view of these caveats, it is unclear whether delusions could form via the teleological "shearing" of particular belief components under stressful circumstances. *Non*-delusional misbeliefs, however, *might* potentially be formed in something like this way (see section 13 for a discussion of health illusions). To an extent the issue here is merely stipulative, hinging on the definition of "delusion" one adopts. If delusions are dysfunctional by definition, then they cannot be adaptive. Moreover, many have reported that, in times of great stress, faith in God has given them "the strength to go on". It may be true that there are no atheists in foxholes (although see Dennett, 2006b), but if delusions are defined so as to exclude conventional religious beliefs (American Psychiatric Association, 2000), then even if foxhole theism is biologically adaptive it will not count as an instance of biologically adaptive *delusion*.

Accounts of religious belief as an adaptation in general have been proposed by a number of commentators (e.g. Johnson & Bering, 2006; Wilson, 2002; but see Dennett, 2006a, for a critique and an alternative evolutionary account). Given the costs associated with religious commitment (see Bulbulia, 2004; Sosis, 2004; Dawkins, 2006a; Ruffle & Sosis, 2007), it seems likely that such commitment is accompanied by bona fide belief of one sort or another (it might be only bona fide *belief in belief*–see Dennett 2006a). We therefore consider now whether in religion we have a candidate domain of adaptive misbelief.

11. Supernatural agency

Interestingly, error management logic pervades contemporary thinking about the origin of religion, and is also apparent in some less contemporary thinking:

> "God is, or He is not" … What will you wager? ... Let us weigh the gain and the loss in wagering that God is. Let us estimate these two chances. If you gain, you gain all; if you lose, you lose nothing. Wager then without hesitation that He is.

Pascal's famous wager provides perhaps the quintessential statement of error management logic, although it is important to note that the wager is an outcome of domain general rationality, whereas error management as implemented by evolved cognitive mechanisms is always domain specific (Haselton & Nettle, 2006). One such domain relevant to religion is the domain of agency detection. Guthrie (1993) has argued that a bias toward inferring the presence of agents would have been adaptive in the evolutionary past: "It is better for a hiker to mistake a boulder for a bear, than to mistake a bear for a boulder" (1993, p. 6). He argues further that religious belief may be a by-product of evolved cognitive mechanisms that produce such biases – mechanisms that Barrett (2000) has termed "Hyperactive agent-detection devices" ("HADDs"). As a by-product theory of religion (see below), this account provides little suggestion that religious belief is adaptive misbelief. Other authors, however, have proposed accounts of religion as an adaptation that incorporate error management logic.

For example, Johnson, Bering and colleagues (Bering & Johnson, 2005; Johnson, 2005; Johnson & Bering, 2006; Johnson & Krueger, 2004; Johnson, Stopka, & Knights, 2003) have advanced a "supernatural punishment hypothesis" regarding the evolution of human cooperation. The nature and extent of human cooperation poses a significant evolutionary puzzle (Fehr & Gaechter, 2002). Human societies are strikingly anomalous in this respect relative to other animal species, as they are based on large-scale cooperation between genetically unrelated individuals (Fehr & Fischbacher, 2003, 2004). Classic adaptationist accounts of cooperation such as kin selection (Hamilton, 1964) and direct reciprocity (Trivers, 1971) cannot explain these features of human cooperation. Moreover, the theories of indirect reciprocity (Alexander, 1987) and costly signalling (Gintis, Smith, & Bowles, 2001; Zahavi, 1995)**,** which show how cooperation can emerge in larger groups when individuals have the opportunity to establish reputations, struggle to explain the occurrence of cooperation in situations that preclude reputation formation - such as in anonymous, one-shot economic games (Fehr & Gaechter, 2002; Gintis, Bowles, Boyd, & Fehr, 2003; Henrich & Fehr, 2003).

Johnson, Bering and colleagues (Bering & Johnson, 2005; Johnson, 2005; Johnson & Bering, 2006; Johnson & Krueger, 2004; Johnson, Stopka, & Knights, 2003) argue that belief in morally interested supernatural agents – and fear of punishment by such agents -

may sustain cooperation in such situations. The argument they put forward is based explicitly on error management theory. They suggest that the evolutionary advent of language, on the one hand, and Theory of Mind (ToM; Premack & Woodruff, 1978), on the other (specifically, the evolution of the "intentionality system", a component of ToM geared toward representing mental states as the unseen causes of behaviour; Bering, 2002; Povinelli & Bering, 2002), occasioned a novel set of selection pressures. In particular, the evolution of these cognitive capabilities increased the costs associated with social defection (because one's social transgressions could be reported to absent third parties), and thus increased the adaptiveness of mechanisms that inhibit selfish actions.

Belief in supernatural punishment – an incidental by-product of the intentionality system - is one such mechanism. Theirs is thus an argument for supernatural belief as exaptation (Gould & Vrba, 1982), a fact that is important for the plausibility of their model. Their central claim is that selection would favour exaggerated estimates of the probability and/or consequences of detection, and thus would favour belief in morally interested supernatural agents. It is not clear, however, that the latter would be necessary to drive the former. Selection might simply implement biased beliefs regarding the probability and/or consequences of detection (cutting out the middle man, as it were). Even more parsimoniously, selection might favour accurate beliefs and implement appropriately judicious action policies vis-à-vis social situations (*cf.* the social exchange heuristic of Yamagishi, Terai, Kiyonari, Mifune, & Kanazawa, 2007). As per our earlier observations regarding evolutionary explorations in Design Space, however, such 'simpler' solutions might be unavailable to selection; it may be that the most direct means of inhibiting selfish behaviour is via supernatural punishment beliefs. If such beliefs were already on the evolutionary scene as by-products of pre-existing intentionality system structures, then they could be conveniently co-opted without any need for the engineering of novel neuro-cognitive machinery (see Bering, 2006).

The argument depends on a crucial error management assumption – that the costs of the two relevant errors in this novel selection environment are recurrently asymmetric, i.e. the cost of cheating and being caught reliably exceeds the cost of cooperating when cheating would have gone undetected. Provided that this inequality obtains, the theory claims that a propensity to believe in morally interested supernatural agents would have been selected for, because individuals holding such beliefs would tend to err on the (cooperative) side of caution in their dealings with conspecifics. "Machiavellian" unbelievers would not therefore gain an advantage, as they would lack important

"restraints on self-interested conduct" and thus be "too blatantly selfish for the subtleties of the new social world" (Johnson, 2005, p. 414).

What is the evidence for this theory? Johnson (2005) utilized data from Murdock and White's (1969) Standard Cross-Cultural Sample (SCCS) of 186 human societies around the globe to test whether the concept of supernatural punishment – indexed by the importance of moralizing "high gods" – was associated with cooperation. He found "high gods" to be "significantly associated with societies that are larger, more norm compliant in some tests (but not others), loan and use abstract money, are centrally sanctioned, policed, and pay taxes" (p. 426; see also Roes & Raymond, 2003). As Johnson acknowledges, his measures of supernatural punishment and cooperation were imprecise (a limitation of the data set employed), and his evidence is correlational at best – the causal relationship between supernatural punishment beliefs and cooperation remains obscure. The same criticisms apply to Rossano's (2007) argument that the emergence (in the Upper Palaeolithic) of certain ancient traits of religion (involving belief in "ever-vigilant spiritual monitors"; p. 272) coincides with evidence for a dramatic advance in human cooperation (see Norenzayan & Shariff, 2008, for a review of further studies reporting correlational evidence of religious prosociality).

In view of this criticism, studies that elicit *causal* evidence for the supernatural punishment hypothesis are crucial. The findings of a recent study by Shariff and Norenzayan (2007) are worth considering in this regard. These authors used a scrambled-sentence paradigm to implicitly prime "God" concepts, and found that participants primed in this manner gave significantly more money in a subsequent (anonymous, one-shot) economic game (the Dictator Game; see Camerer, 2003) than control participants. In discussing these results, Shariff and Norenzayan made appeal to a "supernatural watcher" interpretation of their findings, suggesting that their religious primes "aroused an imagined presence of supernatural watchers, and that this perception then increased prosocial behavior" (p. 807). As Randolph-Seng and Nielsen (2008) note, however, this interpretation may be less parsimonious than a behavioural-priming or ideomotor-action account (which Shariff and Norenzayan also considered), in which the activation of specific perceptual-conceptual representations increases the likelihood of behaviour consistent with those representations (see Dijksterhuis, Chartrand, & Aarts, 2007). Thus, much as people walk more slowly when the concept "elderly" is primed (Bargh, Chen, & Burrows, 1996), priming words that are semantically associated with prosocial behaviour (including words such as "God" and "prophet", both of which were utilised as "religious

primes" by Shariff and Norenzayan) may lead to such behaviour simply by virtue of that association.

The behavioural-priming or ideomotor-action explanation is buttressed by the results of Shariff and Norenzayan's second study, which showed that implicitly primed "secular" concepts were comparable to implicitly primed "God" concepts in terms of their effect on giving in a subsequent Dictator Game. As Randolph-Seng and Nielsen (2008) point out, it is not clear why secular primes such as "civic" and "contract", that contain no reference to God, should enhance prosocial behaviour if such behaviour results from the activation of "supernatural watcher" concepts. Nevertheless, we feel that the research design of Shariff and Norenzayan (and that of comparable recent studies; see Pichon, Boccato, & Saroglou, 2007; Randolph-Seng & Nielsen, 2007) is insufficient to adequately discriminate between the supernatural watcher and behavioural-priming interpretations. What is needed is a study that clearly separates the influence of an "agency" dimension (whether natural or supernatural) from a "prosociality" dimension. The appeal of the supernatural punishment hypothesis is that it shows how reputational concerns might influence behaviour in situations that preclude actual reputation formation. It is true that the "religious prime" and "secular prime" categories utilized by Shariff and Norenzayan both included words potentially associated semantically with prosocial behaviour. We note, however, that both word categories also include words potentially associated with agency ("God" and "prophet" in the former category, "jury" and "police" in the latter). It may be that the surveillance connotations of a word such as "police" may mean that priming with this word enhances prosocial behaviour by activating reputational concerns – *not* by semantic association with prosociality! Future studies would do well to tease these factors apart.

Recent research by Bering, McLeod and Shackelford (2005) employed a different paradigm to elicit causal evidence regarding the effect of a supernatural watcher (albeit a supernatural watcher without obvious moral interests). In one condition of their third study, undergraduate students were casually informed that the ghost of a dead graduate student had recently been noticed in the testing room. These participants were subsequently less willing than control participants to cheat on a competitive computer task, despite a low apparent risk of social detection. This result is intriguing, and not obviously susceptible to explanation in terms of behavioural-priming effects (*cf.* Randolph-Seng & Nielsen, 2007). As the relevant information was not collected, however, it is not clear to what extent the effect of the ghost prime in this study was mediated by participants' belief in ghosts. This is an important point, as it raises the

possibility that if behavioural effects *are* reliably elicited by supernatural primes, they may be elicited not by belief but by *alief*! (Gendler, 2008). Perhaps suitably primed participants *alieve* that a supernatural agent is watching, but *believe* no such thing. If this is the case, then such effects, although interesting, will have little bearing on the question of whether misbelief can be systematically adaptive.

It turns out that the evidence is mixed regarding whether supernatural belief mediates the effect of supernatural primes on behaviour. In the first of Shariff and Norenzayan's (2007) studies, the religious prime increased generosity for both theists and atheists. In their second study, however, the effect of the religious prime was stronger for theists than atheists (and in fact non-significant for atheists). It may be that this difference is attributable to the more stringent atheist criterion employed in the latter study, in which case belief *may* be crucial. Recent work by Bushman, Ridge, Das, Key and Busath (2007), which found that scriptural violence sanctioned by God increased aggression, especially in religious participants, is consistent with this proposition. However, Randolph-Seng and Nielsen (2007) found that whereas participants primed with religious words cheated significantly less on a subsequent task than control participants, the intrinsic religiosity of participants did not interact with the prime factor.

At present, therefore, there is no strong evidence that religious belief is important for the efficacy of religious primes, nor any strong evidence that such primes exert their effects by activating reputational concerns involving supernatural agents. Other approaches notwithstanding (e.g. Wilson, 2002; Sosis, 2004; Dawkins, 2006a), the currently dominant evolutionary perspective on religion remains a by-product perspective (Atran, 2004; Atran & Norenzayan, 2004; Bloom, 2004, 2005, 2007; Boyer, 2001, 2003, 2008; Hinde, 1999). On this view, supernatural (mis)beliefs are side-effects of a suite of cognitive mechanisms adapted for other purposes. Such mechanisms render us hyperactive agency detectors (Guthrie, 1993; Barrett, 2000), promiscuous teleologists (Kelemen, 2004), and intuitive dualists **(**Bloom, 2004); collectively (and incidentally), they predispose us to develop religious beliefs – or at least they facilitate the acquisition of such beliefs (Bloom, 2007). Meanwhile, advocates of "strong reciprocity" (Fehr, Fischbacher, & Gaechter, 2002; Gintis, 2000) argue that the puzzle of large-scale human cooperation may be solved by invoking cultural group selection (Boyd, Gintis, Bowles, & Richerson, 2003; Henrich & Boyd, 2001) or gene–culture coevolution (Bowles, Choi, & Hopfensitz, 2003; see Fehr & Fischbacher, 2003; Gintis, 2003).

12. Self-deception

When a person cannot deceive himself the chances are against his being able to deceive other people.

~ Mark Twain

[T]he first and best unconscious move of a dedicated liar is to persuade himself he's sincere.

~ Ian McEwan, "Saturday"

Arguments that systematic misbelief may have been selected for its ability to facilitate the successful negotiation of social exchange scenarios are not confined to the domain of religion. In his foreword to the first edition of Richard Dawkins' book *The Selfish Gene*, for example, the evolutionary biologist Robert Trivers outlined an influential theory of the evolution of *self-deception*:

> [I]f (as Dawkins argues) deceit is fundamental in animal communication, then there must be strong selection to spot deception and this ought, in turn, to select for a degree of self-deception, rendering some facts and motives unconscious so as not to betray – by the subtle signs of self-knowledge – the deception being practiced. Thus, the conventional view that natural selection favors nervous systems which produce ever more accurate images of the world must be a very naïve view of mental evolution. (2006, p. xx; see also Trivers, 1985, 2000; Alexander, 1979, 1987; Lockard, 1978, 1980; Lockard & Paulhus, 1988)

In the intervening years the notion that self-deception has evolved because it facilitates *other*-deception appears to have become something of a received view in evolutionary circles. The notion is not without its critics, however. Both Ramachandran and Blakeslee (1998) and Van Leeuwen (2007) have pointed out that deceivers who believe their own lies (regarding, say, the whereabouts of a food source) will not themselves be able to take advantage of the truth. Deception is thus clearly possible without self-deception. Van Leeuwen (2007) also claims the converse - that self-deception frequently occurs in the absence of any intention to deceive. On the basis of such considerations, Van Leeuwen argues that self-deception is not an adaptation but a by-product of other features of human cognitive architecture.

In any case, Trivers' theory has received surprisingly little empirical attention, and we know of no direct empirical evidence that the theory is valid. Indeed, a recent study by McKay, Novello and Taylor (in preparation) found preliminary evidence that high self-deceivers were, if anything, *less* likely to be trusted in a cooperative exchange situation than low self-deceivers. These authors recruited groups of previously unacquainted

participants, had them interact briefly with one another, and then invited each participant to play an anonymous, one-shot Prisoners Dilemma game with each other participant. Participants were subsequently told that they could double the stakes for one of these games. Individuals higher in self-deception (measured using the Self-Deceptive Enhancement [SDE] scale of the Balanced Inventory of Desirable Responding [BIDR; Paulhus, 1988]) were less likely to be nominated for such double-stakes exchanges, suggesting that such individuals appeared less trustworthy than individuals lower in self-deception.

In a variant of Trivers' dictum, Krebs and Denton (1997) state that "Illusions about one's worth are adaptive because they help people deceive others about their worth" (p. 37; see also Smith, 2006). Given the lack of evidence that others *are* deceived about the worth of self-deceptive individuals, it is questionable whether "illusions about one's worth" do in fact serve this function. Might such illusions serve other adaptive functions, however? Having peeled the onion down, and set aside a variety of inconclusive candidates for adaptive misbelief, we turn finally to an investigation of this question.

13. Positive illusions

> The perception of reality is called mentally healthy when what the individual sees corresponds to what is actually there.
>
> *~ Jahoda (1958, p. 6)*

> [T]he healthy mind is a self-deceptive one.
>
> *~ Taylor (1989, p. 126)*

In parallel with the prevailing evolutionary view of adaptive belief, a number of psychological traditions have regarded close contact with reality as a cornerstone of mental health (Jahoda, 1953, 1958; Maslow, 1950; Peck, 1978; Vaillant, 1977). A substantial body of research in recent decades, however, has challenged this view, suggesting instead that optimal mental health is associated with *unrealistically positive* self-appraisals and beliefs.[xii] Taylor and colleagues (e.g. Taylor, 1989; Taylor & Brown, 1988) refer to such biased perceptions as "positive illusions", where an illusion is "a belief that departs from reality" (Taylor & Brown, 1988, p. 194). Such illusions include unrealistically positive self-evaluations, exaggerated perceptions of personal control or mastery, and unrealistic optimism about the future.

For example, evidence indicates that there is a widespread tendency for most people to see themselves as better than most others on a range of dimensions. This is the "better-than-average effect" (Alicke, 1985) - individuals, on the average, judge themselves to be more intelligent, honest, persistent, original, friendly and reliable than the average person. Most college students tend to believe that they will have a longer-than-average lifespan, while most college instructors believe that they are better-than-average teachers (Cross, 1977). Most people also tend to believe that their driving skills are better than average – even those who have been hospitalised for accidents (e.g. McKenna, Stanier, & Lewis, 1991; Williams, 2003). In fact, most people view themselves as better than average on almost any dimension that is both subjective and socially desirable (Myers, 2002). Indeed, with exquisite irony, most people even see themselves as less prone to such self-serving distortions than others (Friedrich, 1996; Pronin, Gilovich, & Ross, 2004; Pronin, Lin, & Ross, 2002).

Positive illusions may well be pervasive, but are they adaptive, evolutionarily speaking? For example, do such misbeliefs sustain and enhance *physical* health? Our positive illusions may "feel good" and yet contribute nothing to - or even be a tolerable burden upon - our genetic fitness, a side effect that evolution has not found worth blocking. On the other hand, they may be fitness-enhancing, in either of two quite different ways. They may lead us to undertake adaptive actions; or they may more directly sustain and enhance health, or physical fitness in the everyday sense. We consider each of these prospects in turn.

First, let's look at what happens when positive illusions affect the decisions we make in the course of deliberate, intentional action. Do these rosy visions actually lead people to engage in more adaptive behaviours? According to Taylor and Brown (1994), they do. These authors note that individuals with strong positive perceptions – and in particular, *inflated* perceptions - of their abilities are more likely to attain success than those with more modest self-perceptions. In this connection they quote Bandura:

> It is widely believed that misjudgment produces dysfunction. Certainly, gross miscalculation can create problems. However, optimistic self-appraisals of capability that are not unduly disparate from what is possible can be advantageous, whereas veridical judgments can be self-limiting. When people err in their self-appraisals, they tend to overestimate their capabilities. This is a benefit rather than a cognitive failing to be eradicated. If self-efficacy beliefs always reflected only what people could do routinely, they would rarely fail but they would not mount the extra effort needed to surpass their ordinary performances. (Bandura, 1989, p. 1177)

Haselton and Nettle (2006) note the tacit error management perspective in Taylor and Brown's conception of positive illusions:

> [I]f the [evolutionary] cost of trying and failing is low relative to the potential [evolutionary] benefit of succeeding, then an illusional positive belief is not just better than an illusional negative one, but also better than an unbiased belief… (Haselton & Nettle, 2006, p. 58; see also Nettle, 2004)

Although the link here with error management is interesting and relevant, it is worth pausing to consider the precise wording of this quote. Haselton and Nettle speak of an illusional positive belief as being better than an unbiased belief, when presumably what they mean is that a belief *system* geared toward forming illusional positive beliefs – assuming that such beliefs are consistently less detrimental to fitness than illusional negative beliefs - may be more adaptive than an unbiased belief *system*. Even if the misbeliefs arising through the operation of the former system arise through the normal operation of that system, the misbeliefs *themselves* must surely count as abnormal (Millikan, 2004). After all, it's not clear that there is anything adaptive about trying and failing (but see Dennett, 1995b). Smoke detectors biased toward false alarms are no doubt preferable to those biased toward the more costly errors (failures to detect actual fires); but that doesn't mean that a false alarm is a cause for celebration. If a smoke detector came onto the market that detected every actual fire without ever sounding a false alarm, that would be the one to purchase. Even if they spring from adaptively biased misbelief-producing systems, therefore, individual misbeliefs about success are arguably more of a tolerable by-product than an adaptation. (Possible exceptions to this might be cases where individuals falsely believe that they will attain great success, yet where the confident striving engendered by such misbelief leads to greater success than would have been attained had they *not* falsely believed. Perhaps it is sometimes necessary to believe that you will win gold in order to have any chance of winning silver or bronze; see Krebs & Denton, 1997; Benabou & Tirole, 2002).

Might there be evidence, however, that misbeliefs *themselves* can propel adaptive actions? Here we note that positive illusions need not be merely about oneself. Perhaps the most compelling indication that positively biased beliefs lead people to engage in biologically adaptive behaviours is when such beliefs concern other people – in particular, those we love. Gagné and Lydon (2004; see also Fowers, Lyons, & Montel, 1996; Fowers, Lyons, Montel, & Shaked, 2001; Murray, Holmes, & Griffin, 1996) have found that the better-than-average effect applies for people's appraisals not just of themselves but also of their partners - 95% judge their partners more positively than the

average partner with respect to intelligence, attractiveness, warmth and sense of humour. Such biased appraisal mechanisms may be crucial to ensure the completion of species-specific parental duties: “The primary function of love is to cement sexual relationships for a period of several years, in order to ensure that the vulnerable human infant receives care from its mother, resources from its father, and protection from both” (Tallis, 2005, p. 194; see also Fisher, 2006). Note, in this connection, that biased appraisals of one’s children may also facilitate parental care: “[T]he ability of parents to deny the faults of their children sometimes seems to border on delusion” (Krebs & Denton, 1997, p. 34). Wenger and Fowers (2008) have recently provided systematic evidence of positive illusions in parenting. Most participants in their study rated their own children as possessing more positive (86%) and less negative (82%) attributes than the average child. This better-than-average effect, moreover, was a significant predictor of general parenting satisfaction.

Finally, we consider evidence that positive illusions can directly sustain and enhance health. Research has indicated that unrealistically positive views of one’s medical condition and of one's ability to influence it are associated with increased health and longevity (Taylor, Lerner, Sherman, Sage, & McDowell, 2003). For example, in studies with HIV-positive and AIDS patients, those with unrealistically positive views of their likely course of illness showed a slower illness course (Reed, Kemeny, Taylor, & Visscher, 1999) and a longer survival time (Reed, Kemeny, Taylor, Wang, & Visscher, 1994; for a review see Taylor, Kemeny, Reed, Bower, & Gruenewald, 2000).

Taylor et al. (2000) conjectured that positive illusions might work their medical magic by regulating physiological and neuroendocrine responses to stressful circumstances. Stress-induced activation of the autonomic nervous system and the hypothalamic-pituitary-adrenocortical (HPA) axis facilitates “fight or flight” responses and is thus adaptive in the short-term. Chronic or recurrent activation of these systems, however, may be detrimental to health (see McEwen, 1998), so psychological mechanisms that constrain the activation of such systems (perhaps doxastic shear pins that break - or even just bend a little - in situations of heightened stress) may be beneficial. Consistent with the above hypothesis, Taylor et al. (2003) found that self-enhancing cognitions in healthy adults were associated with lower cardiovascular responses to stress, more rapid cardiovascular recovery, and lower baseline cortisol levels.

Results linking positive illusions to health benefits are consistent with earlier findings that patients who deny the risks of imminent surgery suffer fewer medical complications

and are discharged more quickly than other patients (Goleman, 1987, cited in Krebs & Denton, 1997), and that women who cope with breast cancer by employing a denial strategy are more likely to remain recurrence-free than those utilising other coping strategies (Dean & Surtees, 1989). In such cases the expectation of recovery appears to facilitate recovery itself, even if that expectation is unrealistic. This dynamic may be at work in cases of the ubiquitous *placebo effect*, whereby the administration of a medical intervention instigates recovery before the treatment could have had any direct effect and even when the intervention itself is completely bogus (Benedetti et al., 2003; Humphrey, 2004).

Placebos have been acclaimed, ironically, as "the most adaptable, protean, effective, safe and cheap drugs in the world's pharmacopoeia" (Buckman & Sabbagh, 1993; cited in Humphrey, 2004). They have proven effective in the treatment of pain and inflammation, stomach ulcers, angina, heart disease, cancer and depression, among other conditions (Humphrey, 2002, 2004). From an evolutionary perspective, however, the placebo effect presents something of a paradox:

> When people recover from illness as a result of placebo treatments, it is of course their own healing systems that are doing the job. Placebo cure is *self-cure*. But if the capacity for self-cure is latent, then why is it not used immediately? If people can get better by their own efforts, why don't they just get on with it as soon as they get sick – without having to wait, as it were, for outside permission? (Humphrey, 2004, p. 736, emphasis in original.)

Humphrey (2002, 2004) considers the placebo effect in an evolutionary context and suggests an ingenious solution to this paradox. Noting that immune system functioning can be very costly, Humphrey construes the human immune response as under the regulation of an evolved administrative system that must manage resources as efficiently as possible. Because resources are limited, there is adaptive value to limiting resource expenditure just as there is value in the expenditure itself.

Sound economic management requires forecasting the future, and thus the health management system would need to take into account any available information relevant to future prospects. Such data would include information about the nature of the threat itself (including the likelihood of spontaneous remission), the costs of mounting an appropriate defence, and evidence relating to the course of the illness in other victims. Paramount among such sources of information, however, would be information about the availability of medical care: "People have learned… that nothing is a better predictor of how things will turn out when they are sick… than the presence of doctors, medicines,

and so on" (Humphrey, 2004, p. 736). To put a military gloss on Humphrey's economic resource management metaphor, there is less need for caution and conservation of resources once reinforcements arrive. Only then can one "*spare no expense* in hopes of a quick cure" (Dennett, 2006a, p. 138; emphasis in original).

The placebo effect seems at first to be a case where misbelief in the efficacy of a particular treatment regimen (which, after all, may be a sham with zero direct efficacy) facilitates health and physical fitness. Is this, however, a case of evolved misbelief? If Humphrey's account of the placebo effect is along the right lines, what evolved was a bias to attend to and wait for signs of security before triggering a full bore immune response, and these signs would, in the main, have been true harbingers of security (otherwise the bias would not have been adaptive and would not have evolved). As drug trials and placebos did not figure in our evolutionary history, they represent a later, artificial "tricking" of this evolved system, similar to the way calorie-free saccharine tricks our sweet tooth or pornography tricks our libido. Placebo misbelief, therefore, is not adaptive misbelief – it's a by-product of an adaptation. In Humphrey's words, the "human capacity for responding to placebos is… an emergent property of something else that *is* genuinely adaptive: namely, a specially designed procedure for 'economic resource management'… Unjustified placebo responses, triggered by invalid hopes, must be counted a *biological mistake*" (2002, pp. 261 & 279, emphasis in original).

Do similar remarks apply to the instances of positive illusion and health discussed above? Are the "unjustified" expectations and "invalid hopes" of some AIDS and cancer patients biologically mistaken? One might argue that if "unrealistic" optimism facilitates happy outcomes, then – in retrospect – such optimism was not so unrealistic after all! However, it seems clear that optimism in the relevant studies is not *realistic* optimism (even allowing that this is not an oxymoronic concept). For example, Reed et al. (1994) recruited gay men, who had been diagnosed with AIDS for about a year, for an investigation into the effect of positive illusions on physical health. As the data for this particular study were collected in the late 1980s, life expectancy for these men was not long, and two thirds of the men had died at the completion of the study. Realistic acceptance of death (measured by items including the reverse-scored item "I refuse to believe that this problem has happened") was found to be a significant negative predictor of longevity, with high scorers on this measure typically dying nine months earlier than low scorers. This relationship remained significant when a variety of potential predictors of death were controlled for, including age, time since diagnosis, self-reported health

status, and number of AIDS-related symptoms. It does seem, therefore, that the relevant beliefs here were unrealistically positive. “Foxhole” beliefs of a sort.

In positive illusions situations such as those outlined above, the benefits accrue from misbelief directly – not merely from the systems that produce it. To return to the terminology we introduced earlier, such doxastic departures from reality – such apparent limitations of veridicality - are not culpable but entirely forgivable: design *features*, even. These beliefs are “Normal” in the capitalised, Millikanian sense. In such situations, we claim, we have our best candidates for *evolved misbelief.*

14. Ungrounded beliefs

Although “natural selection does not care about truth; it cares only about reproductive success” (Stich, 1990, p. 62), true beliefs *can* have instrumental value for natural selection - insofar as they facilitate reproductive success. In many cases (perhaps most), beliefs will be adaptive by virtue of their veridicality. The adaptiveness of such beliefs is not independent of their truth or falsity. On the other hand, the adaptiveness (or otherwise) of *some* beliefs is quite independent of their truth or falsity. Consider, again, supernatural belief: If belief in an omniscient, omnipotent deity is adaptive because it inhibits detectable selfish behaviour (as per the Johnson & Bering theory that we discussed in section 11), this will be the case whether or not such a being actually exists. If such a being does not exist, then we have adaptive misbelief. However, were such a being to suddenly pop into existence, the beliefs of a heretofore false believer would not become maladaptive – they would remain adaptive.

The misbeliefs that we have identified as sound candidates for adaptive misbelief are like the supernatural (mis)beliefs in the example above – although we claim that they were adaptive in themselves (not merely by-products of adaptively biased misbelief-producing systems), we do not claim that they were adaptive *by virtue of their falsity:* “Falseness itself could not be the point” (Millikan, 2004, p. 86). It may be adaptive to believe that one’s partner and one’s children are more attractive (…etc.) than the average, but such adaptive beliefs are only adaptive *mis*beliefs, on our definition, if they happen to be false. Good grounds may arise for believing these things (success in beauty pageants, excessive attention from rivals etc.), but such grounds will not render these beliefs any less adaptive. Their adaptiveness is independent of their truth or falsity. Any given adaptive misbeliever is thus an adaptive *mis*believer because of contingent facts about the world – because her children are not actually as intelligent as she believes they are; because his

prospects for recovery are not as good as he believes they are, etc. The upshot is that we do not expect adaptive misbeliefs to be generated by mechanisms specialised for the production of beliefs that are false per se. Instead, there will be evolved tendencies for forming specific *ungrounded* beliefs in certain domains. Where these beliefs are (contingently) false, we will see adaptive misbelief.

Dweck and colleagues (Dweck, 1999; Blackwell, Trzesniewski, & Dweck, 2007) have shown a subtle instance of ungrounded belief (not necessarily false) that propels seemingly adaptive action. These authors distinguish two different "self-theories" of intelligence as part of the implicit "core beliefs" of adolescents: an "entity" theory (intelligence is a thing that you have a little or a lot of) and an "incremental" theory (intelligence is a malleable property that can develop). Those who hold an incremental theory are better motivated, work harder, and get better grades, and if students are taught an incremental theory in an intervention, they show significant improvement, and significantly more than a control group that is also given extra help but without the incremental theory. In fact, if students are told (truly or falsely) that they are particularly intelligent (intelligence is an entity and they have quite a lot of it), they actually do worse than if not told this. Note that these results are independent of the issue of whether or not an entity theory or an incremental theory is closer to the truth (or the truth about particular students). So whether or not one's intelligence is malleable, a belief that one's intelligence is malleable seems to have a strong positive effect on one's motivation and performance. It is tempting to conjecture that evolution has discovered this general tendency and exploited it: whenever a belief about a desirable trait is "subjective" (Myers, 2002), not likely to be rudely contradicted by experience, evolution should favour a disposition to err on the benign side, whatever it is, as this will pay dividends at little or no cost. Such an evolved bias could have the effect of installing a host of unrealistically positive beliefs about oneself or about the vicissitudes to be encountered in the environment. What would hold this tendency in check, preventing people from living in fantasy worlds of prowess and paradise? As usual, the tendency should be self-limiting, with rash overconfidence leading to extinction in the not very long run (see Baumeister, 1989, regarding the "optimal margin of illusion").

If psychologists like Dweck can discover and manipulate these core beliefs today, our ancestors, with little or no theory or foresight, could have stumbled onto manipulations of the same factors and been amply rewarded by the effects achieved, turning their children into braver, more confident warriors, more trustworthy allies, more effective agents in many dimensions. Cultural evolution can have played the same shaping and pruning role

as genetic evolution, yielding adaptations that pay for themselves - as all adaptations must - in the differential replication of those who adopt the cultural items, *or* in the differential replication of the cultural items themselves (Dawkins, 2006b; Dennett, 1995a), or both. This in turn would open the door to gene-culture co-evolution such as has been demonstrated with lactose tolerance in human lineages with a tradition of dairy herding (Beja-Pereira et al., 2003; Feldman & Cavalli-Sforza, 1989; Holden & Mace, 1997). Culturally evolved practices of inculcation could then create selective forces favouring those genetic variants that most readily responded to the inculcation, creating a genetically transmitted bias, a heightened susceptibility to those very practices (Dennett, 2006a; McClenon, 2002).

15. Conclusion

The driving force behind natural selection is survival and reproduction, not truth. All other things being equal, it is better for an animal to believe true things than false things; accurate perception is better than hallucination. But sometimes all other things are not equal.

~ Bloom (2004, pp. 222-223)

[S]ystematic bias does not preclude a tether to reality.

~ Haselton and Nettle (2006, p. 62)

Simple folk psychology tells us that since people use their beliefs to select and guide their actions, true beliefs are always better than false beliefs - aside from occasional unsystematic lucky falsehoods. But because our belief states have complex effects beyond simply informing our deliberations - they flavour our attitudes and feed our self-images - and complex causes that can create additional ancillary effects - such as triggering emotional adjustments and immune reactions - the dynamics of actual belief generation and maintenance create a variety of phenomena that might be interpreted as evolved misbeliefs. In many cases these phenomena are better seen as prudent policies or sub-personal biases or quasi-beliefs (Gendler's "aliefs"). Of the categories we consider, one survives: positive illusions.

What is striking about these phenomena, from the point of view of the theorist of beliefs as representations, is that they highlight the implicit holism in any system of belief-attribution. To whom do the relevant functional states represent the unrealistic assessment? If only to the autonomic nervous system and the HPA, then theorists would

have no reason to call the states misbeliefs at all, since the more parsimonious interpretation would be an adaptive but localized tuning of the error management systems within the modules that control these functions. But sometimes, the apparently benign and adaptive effect has been achieved by the maintenance of a more global state of falsehood (as revealed in the subjects' responses to questionnaires, etc.) and this phenomenon is itself, probably, an instance of evolution as a tinkerer: in order to achieve this effect, evolution has to misinform the whole organism.

We began this paper with a default presumption - that true beliefs are adaptive and misbeliefs maladaptive. This led naturally to the question of how to account for instances of misbelief. The answer to this question is twofold: First, the panglossian assumption that evolution is a perfect designer - and thus that natural selection will weed out each and every instance of a generally maladaptive characteristic - must be discarded. Evolution, as we have seen, is not a perfect design process, but is subject to economic, historical, and topographical constraints. We must therefore expect that the machinery that evolution has equipped us with for forming and testing beliefs will be less than "optimal" – and sometimes it will break. Moreover, we have seen a variety of ways in which these sub-optimal systems may generate misbeliefs not by malfunctioning but by functioning normally, creating families of errors that are, if not themselves adaptive, apparently tolerable. But beyond that, we have explored special circumstances where, as Bloom writes, "things are not equal" – where the truth hurts so systematically that we are actually better off with falsehood. We have seen that in such circumstances falsehood can be sustained by evolved systems of misbelief. So, in certain rarefied contexts, misbelief itself can actually be *adaptive*. Nevertheless, the truism that misinformation leads in general to costly missteps has not been seriously undermined: although survival is the only hard currency of natural selection, the exchange rate with truth is likely to be fair in most circumstances.

---

[i] Doxastic = of or pertaining to belief.

[ii] We set aside, on this occasion, the important distinction between probabilistic and all-or-nothing conceptions of belief (e.g., Dennett's, 1978, distinction between *belief* and *opinion*), as the issues explored here apply ingenerate to both conceptions.

[iii] For ease of exposition, we will tend to conflate "adaptive" and "adapted" throughout this paper. Because ecological niches change over time, these categories are overlapping but not equal: although all adapted traits must have been adaptive in the evolutionary past, they need not be adaptive in modern environments; likewise, traits that are currently adaptive are not necessarily adapted (they are not necessarily adaptations). This is, of course, an important distinction, but not much will turn on it for our purposes.

[iv] Naturally, manufacturers and consumers do not always see eye to eye. Limitations that appear culpable from a consumer perspective will frequently be judged forgivable by the manufacturer. They may even be deliberate features, as in the DVD region code case. Conversely, some instances of culpable misdesign from the manufacturer's perspective may actually be *welcomed* by consumers. One thinks of the popular

myth of the super-long-lasting incandescent light bulb. According to this myth, the technology exists to manufacture light bulbs that last thousands of times longer than regular bulbs – but to produce such bulbs would kill the light bulb industry, so nobody does! From the perspective of this (mythical) manufacturer, bulbs that last *too* long evidence culpable misdesign (though no consumer would complain).

[v] Stephen Stich (personal communication) provides another imaginary example: "Suppose there were a culture… for whom one specific number is regarded as particularly unlucky, the number 88888888. Designers of calculators know this. So they start with an ordinary calculator and build in a special small program which displays a random number whenever the rest of the calculator says that the answer is 88888888. They advertise this as a special selling point of their calculator. When the answer is really 'that horrible unlucky number' the calculator will tell you it is something else. It will lie to you. Sales of the 'lucky calculator' get a big boost."

[vi] Note that narrow-or-broad construals of function are also possible with respect to artifacts. To cite an example analogous to the immune system case, an electric sabre saw will cut right through its own power cord if the operator lets it. Is this a malfunction? The saw is designed to saw through whatever is put in its way, and so it does! The difference is that we can consult the designer for his/her intentions where artifacts are concerned. Most likely the designer will say "of course the sabre saw hasn't 'malfunctioned' – no artifact need be idiot-proof!" But we can still solicit the information, whereas that option is closed to us for evolved systems. See Section 10 for a related point.

[vii] Fodor (2007) has vigorously challenged not just Millikan's claim, but the family of related claims made by evolutionary theorists. According to Fodor, the historical facts of evolution, even if we knew them, could not distinguish function from merely accompanying by-product. Fodor's position has been just as vigorously rebutted (see e.g. Coyne & Kitcher, 2007; Dennett, 1990b, 2007a, 2008). It is perhaps worth noting that an implication of Fodor's position, resolutely endorsed by Fodor, is that biologists are not entitled to say that eyes are for seeing, or bird wings for flying - though airplane wings, having intelligent human designers, *are* for flying.

[viii] Other animals may have evolved methods of compensating for this distortion. For instance, Casperson (1999) suggests that in a certain class of birds that plan underwater foraging from wading or perched positions above the water, a characteristic vertical bobbing motion of the head may allow them to compensate for refraction: "…the refraction angles change as a bird moves its head vertically, and with suitable interpretation these angular variations can yield unambiguous information about water-surface and prey locations" (p. 45). See also Katzir and Howland (2003), Katzir and Intrator (1987), and Lotem, Schechtman and Katzir (1991).

[ix] Nevertheless, evolved cognitive systems are remarkably supple, as researchers in Artificial Intelligence are forever discovering. Among the holy grails of AI are systems that are "robust" under perturbation and assault, and that will at least "degrade gracefully" - like so many naturally evolved systems - instead of producing fatal nonsense when the going gets tough.

[x] In some cases these "other parties" may potentially be our close kin. This is not to suggest, however, that misbeliefs evolve via kin selection (Hamilton, 1964). Voland and Voland (1995) have suggested that the human "conscience" is an extended phenotype (Dawkins, 1982) of parental genes that evolved in the context of parent/offspring conflict (Trivers, 1974) over altruistic tendencies. In a particular "tax scenario" of this conflict (Voland, 2008; see also Simon, 1990), it may be adaptive for parents to raise some of their offspring to be martyrs (perhaps by instilling in them certain beliefs about the heavenly rewards that await martyrs). In this scenario the martyrdom of the offspring increases the inclusive fitness of the parents (perhaps via a boost in the social status of the family). The martyrs themselves, however, are evolutionary losers – hapless victims of the generally adaptive rule of thumb "believe, without question, whatever your grown-ups tell you" (Dawkins, 2006a, p. 174; see again Simon, 1990).

[xi] The breakage itself would be normatively normal (Normal) yet statistically abnormal. But what about the belief system as a whole? Surely it would cease *its* Normal functioning when a doxastic shear pin broke? Here we return to the overlaps encountered in section 5, and may again invoke Millikan (1993) for an alternative construal: Perhaps the belief system would be made to labour (Normally?) under external conditions not Normal for performance of its proper function.

[xii] The claim that mentally healthy individuals hold unrealistically positive beliefs is related to – but logically distinct from – the contested claim that depressed individuals exhibit *accurate* perceptions and beliefs (a phenomenon known as "depressive realism"; see Alloy & Abramson, 1988; Colvin & Block, 1994).

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.

Akins, K. (1996). Of sensory systems and the "aboutness" of mental states. *The Journal of Philosophy, 93*(7), 337-372.

Alexander, R. D. (1979). *Darwinism and human affairs*. Seattle, WA: University of Washington Press.

Alexander, R. D. (1987). *The biology of moral systems*. New York: Aldine de Gruyter.

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology, 49*, 1621–1630.

Alloy, L. B. & Abramson, L. Y. (1988). Depressive realism: Four theoretical perspectives. In L. B. Alloy (Ed.), *Cognitive processes in depression* (pp. 223-265). New York: Guilford Press.

American Psychiatric Association. (2000). *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision (DSM-IV-TR).* Washington, DC: American Psychiatric Association.

Atran, S. (2004). *In Gods we trust: The evolutionary landscape of religion*. New York: Oxford University Press.

Atran, S. & Norenzayan, A. (2004). Religion's evolutionary landscape: Counterintuition, commitment, compassion, communion. *Behavioral and Brain Sciences, 27*, 713-770.

Bandura, A. (1989). Human agency in social cognitive theory. *American Psychologist, 44*, 1175-1184.

Bargh, J. A., Chen, M. & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*, 230-244.

Barrett, J. L. (2000). Exploring the natural foundations of religion. *Trends in Cognitive Sciences, 4*(1), 29-34.

Baumeister, R. F. (1989). The optimal margin of illusion. *Journal of Social and Clinical Psychology, 8*, 176-189.

Bayes, T. R. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London, 53*, 370–418.

Bayne, T. & Fernández, J. (2009). Delusion and self-deception: Mapping the terrain. In T. Bayne & J. Fernández (Eds.), *Delusion and Self-Deception: Affective and Motivational Influences on Belief Formation* (pp. 1-21). Hove: Psychology Press.

Bayne, T. & Pacherie, E. (2005). In defence of the doxastic conception of delusions. *Mind and Language, 20*(2), 163-188.

Beja-Pereira, A., Luikart, G., England, P. R., Bradley, D. G., Jann, O. C., Bertorelle, G., et al. (2003). Gene-culture coevolution between cattle milk protein genes and human lactase genes. *Nature Genetics, 35*, 311–313.

Benabou, R. & Tirole, J. (2002). Self-confidence and personal motivation. *The Quarterly Journal of Economics, 117*(3), 871-915.

Benedetti, F., Pollo, A., Lopiano, L., Lanotte, M., Vighetti, S. & Rainero, I. (2003). Conscious expectation and unconscious conditioning in analgesic, motor, and hormonal placebo/nocebo responses. *The Journal of Neuroscience, 23*(10), 4315-4323.

Bentall, R. P. & Kaney, S. (1996). Abnormalities of self-representation and persecutory delusions: A test of a cognitive model of paranoia. *Psychological Medicine, 26*, 1231-1237.

Bering, J. M. (2002). The existential theory of mind. *Review of General Psychology, 6*, 3-24.

Bering, J. M. (2006). The folk psychology of souls. *Behavioral and Brain Sciences, 29*, 453-498.

Bering, J. M. & Johnson, D. D. P. (2005). "O Lord . . . you perceive my thoughts from afar": Recursiveness and the evolution of supernatural agency. *Journal of Cognition and Culture, 5*, 118-142.

Bering, J. M., McLeod, K. A. & Shackelford, T. K. (2005). Reasoning about dead agents reveals possible adaptive trends. *Human Nature, 16*, 360-381.

Berrios, G. E. (1991). Delusions as 'wrong beliefs': A conceptual history. *British Journal of Psychiatry, 159*, 6-13.

Binkofski, F., Buccino, G., Dohle, C., Seitz, R. J. & Freund, H. -J. (1999). Mirror agnosia and mirror ataxia constitute different parietal lobe disorders. *Annals of Neurology, 46*, 51-61.

Blackwell, L., Trzesniewski, K. & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development, 78*, 246-263.

Bloom, P. (2004). *Descartes' baby: How child development explains what makes us human*. London: Arrow Books.

Bloom, P. (2005). Is God an accident? *Atlantic Monthly, 296*, 105–112.

Bloom, P. (2007). Religion is natural. *Developmental Science, 10*(1), 147–151.

Boden, M. (1984). Animal perception from an Artificial Intelligence viewpoint. In C. Hookway (Ed.), *Minds, Machines and Evolution* (pp. 153-174). Cambridge: Cambridge University Press.

Bowles, S., Choi, J.-K. & Hopfensitz, A. (2003). The co-evolution of individual behaviours and social institutions. *Journal of Theoretical Biology, 223*(2), 135-147.

Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences, 100*(6), 3531–3535.

Boyer, P. (2001). *Religion Explained: The Evolutionary Origins of Religious Thought*. New York: Basic Books.

Boyer, P. (2003). Religious thought and behaviour as by-products of brain function. *Trends in Cognitive Sciences, 7*(3), 119-124.

Boyer, P. (2008). Religion: Bound to believe? *Nature, 455*(23), 1038-1039.

Bratman, M. E. (1992). Practical reasoning and acceptance in a context. *Mind, 101*(401), 1-15.

Breen, N., Caine, D. & Coltheart, M. (2001). Mirrored-self Misidentification: Two cases of focal onset dementia. *Neurocase, 7*, 239-254.

Breen, N., Caine, D., Coltheart, M., Hendy, J. & Roberts, C. (2000). Towards an understanding of delusions of misidentification: Four case studies. *Mind and Language, 15*(1), 74-110.

Brüne, M. (2001). De Clerambault's syndrome (erotomania) in an evolutionary perspective. *Evolution & Human Behavior, 22*(6), 409-415.

Brüne, M. (2003). Erotomanic stalking in evolutionary perspective. *Behavioral Sciences and the Law, 21*(1), 83-88.

Brüne, M. (in press). Erotomania (De Clerambault's Syndrome) revisited - Clues to its origin from evolutionary theory. In S. P. Shohov (Ed.), *Advances in Psychology Research* (Vol. 21, pp. 185-212).

Bulbulia, J. (2004). The cognitive and evolutionary psychology of religion. *Biology and Philosophy, 19*, 655-686.

Bushman, B. J., Ridge, R. D., Das, E., Key, C. W. & Busath, G. L. (2007). When God sanctions killing: Effect of scriptural violence on aggression. *Psychological Science, 18*(3), 204-207.

Buss, D. M. & Haselton, M. G. (2005). The evolution of jealousy: A response to Buller. *Trends in Cognitive Sciences, 9*(11), 506-507.

Butler, P. V. (2000). Reverse Othello syndrome subsequent to traumatic brain injury. *Psychiatry: Interpersonal & Biological Processes, 63*(1), 85-92.

Camerer, C. (2003). *Behavioral Game Theory*. Princeton: Princeton University Press.

Casperson, L. W. (1999). Head movement and vision in underwater-feeding birds of stream, lake, and seashore. *Bird Behavior, 13*, 31-46.

Chow, Y. C., Dhillon, B., Chew, P. T. & Chew, S. J. (1990). Refractive errors in Singapore medical students. *Singapore Medical Journal, 31*, 472-473.

Coltheart, M. (1996). Are dyslexics different? *Dyslexia, 2*, 79-81.

Coltheart, M. (2002). Cognitive neuropsychology. In H. Pashler & J. Wixted (Eds.) *Stevens' handbook of experimental psychology (3rd ed.), Vol. 4: Methodology in experimental psychology*. (pp. 139-174). New York: John Wiley & Sons.

Coltheart, M., Menzies, P. & Sutton, J. (forthcoming). Abductive inference and delusional belief. In R. Langdon & M. Turner (Eds.) *Delusion and confabulation: Overlapping or distinct psychopathologies of reality distortion.* Macquarie Monographs in Cognitive Science series. Series edited by Coltheart, M. Psychology Press.

Colvin, C. R. & Block, J. (1994). Do positive illusions foster mental health? An examination of the Taylor and Brown formulation. *Psychological Bulletin, 116*(1), 3–20.

Coyne, J. & Kitcher, P. (2007, November 15th). Letter to the Editor re Fodor. *London Review of Books, 29*.

Cross, P. (1977). Not can but will college teaching be improved? *New Directions for Higher Education, 17*, 1-15.

Currie, G. (2000). Imagination, delusion and hallucinations. *Mind and Language, 15*, 168-183.

Currie, G. & Jureidini, J. (2001). Delusion, rationality, empathy: Commentary on Davies et al. *Philosophy, Psychiatry, & Psychology, 8*(2-3), 159-162.

David, A. S. (1999). On the impossibility of defining delusions. *Philosophy, Psychiatry, and Psychology, 6*(1), 17-20.

David, A. S. & Halligan, P. W. (1996). Editorial. *Cognitive Neuropsychiatry, 1*, 1-3.

Davidson, D. (1994). Radical interpretation interpreted. In J. E. Tomberlin (Ed.), *Philosophical Perspectives, Vol. 8, Logic and Language.* (pp. 121-128). Atascadero, CA: Ridgeview.

Davidson, D. (2001). *Inquiries into Truth and Interpretation* (2nd ed). Oxford: Clarendon Press.

Davies, M. & Coltheart, M. (2000). Introduction: Pathologies of belief. In M. Coltheart & M. Davies (Eds.), *Pathologies of belief.* (pp. 1-46). Malden, MA: Blackwell Publishers.

Davies, M., Coltheart, M., Langdon, R. & Breen, N. (2001). Monothematic delusions: Towards a two-factor account. *Philosophy, Psychiatry, and Psychology, 8*(2-3), 133-158.

Dawkins, R. (1982). *The extended phenotype*. San Francisco: Freeman.

Dawkins, R. (1986). *The Blind Watchmaker*. New York: W. W. Norton & Company, Inc.

Dawkins, R. (2006a). *The God delusion*. London: Bantam Press.

Dawkins, R. (2006b). *The selfish gene: 30th anniversary edition*. Oxford: Oxford University Press.

Dean, C. & Surtees, P. G. (1989). Do psychological factors predict survival in breast cancer? *Journal of Psychosomatic Research, 33*(5), 561-569.

Dennett, D. C. (1971). Intentional systems. *Journal of Philosophy, LXVIII*, 87-106.

Dennett, D. C. (1978). *Brainstorms: Philosophical essays on mind and psychology*. Cambridge, MA: The MIT Press. A Bradford Book.

Dennett, D. C. (1982). Beyond belief. In A. Woodfield (Ed.), *Thought and Object*. Oxford: Clarendon Press.

Dennett, D. C. (1987). *The Intentional Stance*. Cambridge, MA: The MIT Press.

Dennett, D. C. (1990a). Attitudes about ADHD: Some Analogies and Aspects. In K. Conners & M. Kinsbourne (Eds.), *ADHD: Attention Deficit Hyperactivity Disorders*. Munich: MMV Medizin Verlag.

Dennett, D. C. (1990b). The interpretation of texts, people and other artifacts. *Philosophy and Phenomenological Research, 50, Supplement*, 177-194.

Dennett, D. C. (1995a). *Darwin's Dangerous Idea*. London: Penguin.

Dennett, D. C. (1995b). How to make mistakes. In J. Brockman & K. Matson (Eds.) *How things Are* (pp. 137-144). New York: William Morrow and Company.

Dennett, D. C. (1998). *Brainchildren - Essays on Designing Minds*: MIT Press/Bradford Books and Penguin.

Dennett, D. C. (2005, August 28th). Show me the science. *The New York Times*, p. 11.

Dennett, D. C. (2006a). *Breaking the spell: Religion as a natural phenomenon*. New York: Viking.

Dennett, D. C. (2006b, November 3rd). Thank Goodness! *Edge: The Third Culture*.

Dennett, D. C. (2007a, November 15th). Letter to the Editor re Fodor. *London Review of Books, 29*.

Dennett, D. C. (2008). Fun and games in Fantasyland. *Mind and Language, 23*(1), 25-31.

Dijksterhuis, A., Chartrand, T. L. & Aarts, H. (2007). Effects of priming and perception on social behavior and goal pursuit. In J. A. Bargh (Ed.), *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 51–131). New York: Psychology Press.

Duntley, J. & Buss, D. M. (1998). *Evolved anti-homicide modules*. Paper presented at the Human Behavior and Evolution Society Conference, Davis, CA, July, 1998.

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality and development*. Philadelphia: Psychology Press.

Easton, J. A., Schipper, L. D. & Shackelford, T. K. (2007). Morbid jealousy from an evolutionary psychological perspective. *Evolution and Human Behavior, 28*, 399–402.

Ellis, A. W. & Young, A. W. (1988). *Human Cognitive Neuropsychology*. Hove, E. Sussex: Lawrence Erlbaum Associates.

Ellis, H. D. (2003). Book review: Uncommon psychiatric syndromes. *Cognitive Neuropsychiatry, 8*(1), 77-79.

Enoch, M. D. & Ball, H. N. (2001). *Uncommon psychiatric syndromes* (4th ed.). London: Arnold.

Fehr, E. & Fischbacher, U. (2003). The nature of human altruism. *Nature, 425*, 785-791.

Fehr, E. & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences, 8*(4), 185-190.

Fehr, E., Fischbacher, U. & Gaechter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature, 13*, 1-25.

Fehr, E. & Gaechter, S. (2002). Altruistic punishment in humans. *Nature, 415*, 137-140.

Feinberg, T. E. (2001). *Altered egos: How the brain creates the self*. Oxford: Oxford University Press.

Feinberg, T. E. & Shapiro, R. M. (1989). Misidentification-reduplication and the right hemisphere. *Neuropsychiatry, Neuropsychology, & Behavioral Neurology, 2*(1), 39-48.

Feldman, M. W. & Cavalli-Sforza, L. L. (1989). On the theory of evolution under genetic and cultural transmission with application to the lactose absorption problem. In M. W. Feldman (Ed.), *Mathematical evolutionary theory* (pp. 145-173). Princeton, NJ: Princeton University Press.

Fisher, H. (2006). The drive to love: The neural mechanism for mate choice. In R. J. Sternberg & K. Weis (Eds.), *The New Psychology of Love* (2nd ed.). New Haven: Yale University Press.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fodor, J. A. (1986). Precis of the modularity of mind. Meaning and cognitive structure. *Behavioral and Brain Sciences, 8*, 1-42.

Fodor, J. A. (2007, October 18th). Why pigs don't have wings. *London Review of Books*.

Fowers, B. J., Lyons, E. M. & Montel, K. H. (1996). Positive marital illusions: Self-enhancement or relationship enhancement? *Journal of Family Psychology, 10,* 192–208.

Fowers, B. J., Lyons, E., Montel, K. H. & Shaked, N. (2001). Positive illusions about marriage among the married, engaged, and single. *Journal of Family Psychology, 15*, 95–109.

Friedrich, J. (1996). On seeing oneself as less self-serving than others: The ultimate self-serving bias? *Teaching of Psychology, 23*(2), 107-109.

Gagné, F. M. & Lydon, J. E. (2004). Bias and accuracy in close relationships: An integrative review. *Personality and Social Psychology Review, 8*(4), 322-338.

Gendler, T. (2008). Alief and belief. *Journal of Philosophy, 105*(10), 634-663.

Ghiselin, M. T. (1974). *The economy of nature and the evolution of sex*. Berkeley, CA: University of California Press.

Gigerenzer, G. & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality *Psychological Review 103*(4), 650-669.

Gigerenzer, G. & Goldstein, D. G. (1999). Betting on one good reason: The take the best heuristic. In G. Gigerenzer, P. M. Todd and the ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 75-95). New York: Oxford University Press.

Gigerenzer, G., Todd, P. M. & the ABC Research Group. (1999). *Simple heuristics that make us smart.* New York: Oxford University Press.

Goldstein, D. G. & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109*(1), 75-90.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology, 206*, 169–179.

Gintis, H. (2003). The hitchhiker's guide to altruism: Gene-culture co-evolution and the internalization of norms. *Journal of Theoretical Biology, 220,* 407–418.

Gintis, H., Bowles, S., Boyd, R. & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior, 24*(3), 153–172.

Gintis, H., Smith, E. & Bowles, S. (2001). Costly signalling and cooperation. *Journal of Theoretical Biology, 213*, 103-119.

Gould, S. J. & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London, Series B, 205*(1161), 581-598.

Gould, S. J. & Vrba, E. S. (1982). Exaptation: A missing term in the science of form. *Paleobiology, 8*(1), 4-15.

Guthrie, S. E. (1993). *Faces in the clouds: A new theory of religion*. Oxford: Oxford University Press.

Hamilton, A. (2007). Against the belief model of delusion. In M. C. Chung, K. W. M. Fulford & G. Graham (Eds.), *Reconceiving Schizophrenia* (pp. 217-234). Oxford: Oxford University Press.

Hamilton, W. D. (1964). Genetical evolution of social behavior I and II. *Journal of Theoretical Biology, 7*, 1–52.

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

Haselton, M. G. (2007). Error Management Theory. In R. F. Baumeister & K. D. Vohs (Eds.), *Encyclopedia of social psychology* (Vol. 1, pp. 311-312). Thousand Oaks, CA: Sage.

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*(1), 81-91.

Haselton, M. G. & Buss, D. M. (2003). Biases in social judgment: Design flaws or design features? In J. P. Forgas, K. D. Williams & W. von Hippel (Eds.), *Responding to the Social World: Implicit and Explicit Processes in Social Judgments and Decisions* (pp. 23-43). New York: Cambridge University Press.

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

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

Henrich, J. & Fehr, E. (2003). Is strong reciprocity a maladaptation? On the evolutionary foundations of human altruism. In P. Hammerstein (Ed.), *Genetic and cultural evolution of cooperation* (pp. 55-82). Cambridge, MA: MIT Press.

Hinde, R. A. (1999). *Why gods persist: A scientific approach to religion*. London: Routledge.

Holden, C. & Mace, R. (1997). Phylogenetic analysis of the evolution of lactose digestion in adults. *Human Biology, 69*, 605–628.

Humphrey, N. (2002). *The Mind Made Flesh*. Oxford: Oxford University Press. See www.humphrey.org.uk

Humphrey, N. (2004). The placebo effect. In R. L. Gregory (Ed.), *Oxford Companion to the Mind* (2nd ed.) (pp. 735-736). Oxford: Oxford University Press. See www.humphrey.org.uk

Huq, S. F., Garety, P. A. & Hemsley, D. R. (1988). Probabilistic judgements in deluded and non-deluded subjects. *Quarterly Journal of Experimental Psychology A, 40*(4), 801-812.

Jahoda, M. (1953). The meaning of psychological health. *Social Casework, 34*, 349-354.

Jahoda, M. (1958). *Current concepts of positive mental health*. New York: Basic Books.

Jaspers, K. (1913/1963). *General psychopathology* (J. Hoenig & M. W. Hamilton, Trans. Vol. One). Baltimore: The Johns Hopkins University Press.

Johnson, D. D. P. (2005). God's punishment and public goods: A test of the supernatural punishment hypothesis in 186 world cultures. *Human Nature, 16*(4), 410-446.

Johnson, D. D. P. & Bering, J. M. (2006). Hand of God, mind of man: Punishment and cognition in the evolution of cooperation. *Evolutionary Psychology, 4*, 219-233.

Johnson, D. D. P. & Krueger, O. (2004). Supernatural punishment and the evolution of cooperation. *Political Theology, 5*(2), 159-176.

Johnson, D. D. P., Stopka, P. & Knights, S. (2003). The puzzle of human cooperation. *Nature, 421*, 911-912.

Katzir, G., & Howland, H. C. (2003). Corneal power and underwater accommodation in great cormorants (Phalacrocorax carbo sinensis). *The Journal of Experimental Biology, 206*, 833-841.

Katzir, G., & Intrator, N. (1987). Striking of underwater prey by a reef heron, Egretta gularis schistacea. *Journal of Comparative Physiology A, 160*, 517-523.

Kelemen, D. (2004). Are children 'intuitive theists'? *Psychological Science, 15*, 295–301.

Kinderman, P. & Bentall, R. P. (1996). Self-discrepancies and persecutory delusions: Evidence for a model of paranoid ideation. *Journal of Abnormal Psychology, 105*(1), 106-113.

Kinderman, P. & Bentall, R. P. (1997). Causal attributions in paranoia and depression: Internal, personal, and situational attributions for negative events. *Journal of Abnormal Psychology, 106*(2), 341-345.

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.), *Evolutionary social psychology* (pp. 21-47). Hillsdale, NJ: Lawrence Erlbaum Associates.

Langdon, R., & Coltheart, M. (2000). The cognitive neuropsychology of delusions. *Mind and Language, 15*(1), 183-216.

Langdon, R., Cooper, S., Connaughton, E. & Martin, K. (2006). A variant of misidentification delusion in a patient with right frontal and temporal brain injury. *Abstracts of the 6th International Congress of Neuropsychiatry, Sydney, Australia. Neuropsychiatric Disease and Treatment, 2*(3 Suppl), S8.

Langdon, R., McKay, R. & Coltheart, M. (2008). The cognitive neuropsychological understanding of persecutory delusions. In D. Freeman, R. Bentall & P. Garety (Eds.), *Persecutory Delusions: Assessment, Theory, and Treatment* (pp. 221-236). New York: Oxford University Press.

Lockard, J. S. (1978). On the adaptive significance of self-deception. *Human Ethology Newsletter, 21*, 4-7.

Lockard, J. S. (1980). Speculations on the adaptive significance of self-deception. In J. S. Lockard (Ed.), *The evolution of human social behavior* (pp. 257-276). New York: Elsevier.

Lockard, J. S. & Paulhus, D. L. (Eds.). (1988). *Self-deception: An adaptive mechanism?* Englewood Cliffs, NJ: Prentice Hall.

Lotem, A., Schechtman, E. & Katzir, G. (1991). Capture of submerged prey by little egrets, Egretta garzetta garzetta: Strike depth, strike angle and the problem of light refraction. *Animal Behaviour, 42*, 341-346.

Maslow, A. H. (1950). Self-actualizing people: A study of psychological health. *Personality, Symposium No. 1*, 11-34.

McClenon, J. (2002). *Wondrous healing: Shamanism, human evolution and the origin of religion*. DeKalb: Northern Illinois University Press.

McEwen, B. S. (1998). Protective and damaging effects of stress mediators. *New England Journal of Medicine, 338*, 171–179.

McKay, R. (forthcoming). Motivated misbelief? Motivational processes in delusion and confabulation. In R. Langdon & M. Turner (Eds.) *Delusion and confabulation: Overlapping or distinct psychopathologies of reality distortion.* Macquarie Monographs in Cognitive Science series. Series edited by Coltheart, M. Psychology Press.

McKay, R. & Efferson, C. (in preparation). The wheat and the chaff: The subtleties of Error Management Theory.

McKay, R., Langdon, R., & Coltheart, M. (2007a). Models of misbelief: Integrating motivational and deficit theories of delusions. *Consciousness and Cognition, 16*, 932-941.

McKay, R., Langdon, R. & Coltheart, M. (2007b). The defensive function of persecutory delusions: An investigation using the Implicit Association Test. *Cognitive Neuropsychiatry, 12*(1), 1-24.

McKay, R., Langdon, R. & Coltheart, M. (2009). "Sleights of mind": Delusions and self-deception. In T. Bayne & J. Fernández (Eds.), *Delusion and Self-Deception: Affective and Motivational Influences on Belief Formation* (pp. 165-185). Hove: Psychology Press.

McKay, R., Novello, D. & Taylor, A. (in preparation). The adaptive value of self-deception.

McKenna, F. P., Stanier, R. A. & Lewis, C. (1991). Factors underlying illusory self-assessment of driving skill in males and females. *Accident Analysis and Prevention, 23*(1), 45-52.

Merton, R. K. (1968). *Social Theory and Social Structure*. New York: Free Press.

Millikan, R. (1984a). *Language, thought and other biological categories*. Cambridge, MA: MIT Press.

Millikan, R. (1984b). Naturalistic reflections on knowledge. *Pacific Philosophical Quarterly, 65*(4), 315-334.

Millikan, R. (1993). *White queen psychology and other essays for Alice*. Cambridge, MA: MIT Press.

Millikan, R. (2004). *Varieties of meaning: The 2002 Jean Nicod lectures*. Cambridge, MA: The MIT Press. A Bradford Book.

Moritz, S., Werner, R. & von Collani, G. (2006). The inferiority complex in paranoia readdressed: A study with the Implicit Association Test. *Cognitive Neuropsychiatry, 11*(4), 402-415.

Mowat, R. R. (1966). *Morbid jealousy and murder*. London: Tavistock.

Murdock, G. P. & White, D. R. (1969). Standard cross-cultural sample. *Ethnology, 8*, 329-369.

Murray, S. L., Holmes, J. G. & Griffin, D. W. (1996). The benefits of positive illusions: Idealization and the construction of satisfaction in close relationships. *Journal of Personality and Social Psychology, 70*, 79–98.

Myers, D. (2002). *Social psychology* (7th ed.): McGraw-Hill.

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

Noë, A. (2004). *Action in perception*. Cambridge, MA: MIT Press.

Norenzayan, A. & Shariff, A. F. (2008). The origin and evolution of religious prosociality. *Science, 322*, 58-62.

Paulhus, D. L. (1988). *Manual for the Balanced Inventory of Desirable Responding*. Toronto: Multi-Health Systems.

Peck, M. S. (1978). *The road less traveled*. New York: Simon & Schuster.

Pichon, I., Boccato, G. & Saroglou, V. (2007). Nonconscious influences of religion on prosociality: A priming study. *European Journal of Social Psychology, 37*, 1032–1045.

Pinker, S. (1997). *How the mind works*. New York: W. W. Norton & Company.

Povinelli, D. J. & Bering, J. M. (2002). The mentality of apes revisited. *Current Directions in Psychological Science, 11*, 115-119.

Premack, D. & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(4), 515-526.

Pronin, E., Gilovich, T. & Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychological Review, 3*, 781-799.

Pronin, E., Lin, D. Y. & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin, 28*(3), 369-381.

Quillian, L. & Pager, D. (2001). Black neighbors, higher crime? The role of racial stereotypes in evaluations of neighborhood crime. *American Journal of Sociology, 107*(3), 717-767.

Quine, W. V. O. (1960). *Word and Object*. Cambridge: The MIT Press.

Quine, W. V. O. & Ullian, J. S. (1978). *The Web of Belief* (2nd ed.). New York: Random House.

Ramachandran, V. S. (1994a). Phantom limbs, neglect syndromes, repressed memories, and Freudian psychology. *International Review of Neurobiology, 37*, 291-333.

Ramachandran, V. S. (1994b). Phantom limbs, somatoparaphrenic delusions, neglect syndromes, repressed memories and Freudian psychology. In O. Sporns & G. Tononi (Eds.), *Neuronal group selection*. San Diego: Academic Press.

Ramachandran, V. S. (1995). Anosognosia in parietal lobe syndrome. *Consciousness and Cognition, 4*(1), 22-51.

Ramachandran, V. S. (1996a). The evolutionary biology of self-deception, laughter, dreaming and depression: Some clues from anosognosia. *Medical Hypotheses, 47*(5), 347-362.

Ramachandran, V. S. (1996b). What neurological syndromes can tell us about human nature: some lessons from phantom limbs, Capgras syndrome, and anosognosia. *Cold Spring Harbor Symposia on Quantitative Biology, 61*, 115-134.

Ramachandran, V. S., Altschuler, E. L. & Hillyer, S. (1997). Mirror Agnosia. *Proceedings of the Royal Society B: Biological Sciences, 264*, 645-647.

Ramachandran, V. S. & Blakeslee, S. (1998). *Phantoms in the brain: Human nature and the architecture of the mind*. London: Fourth Estate.

Randolph-Seng, B. & Nielsen, M. E. (2007). Honesty: One effect of primed religious representations. *The International Journal for the Psychology of Religion, 17*(4), 303-315.

Randolph-Seng, B. & Nielsen, M. E. (2008). Is God really watching you? A response to Shariff and Norenzayan (2007). *The International Journal for the Psychology of Religion, 18*(2), 119-122.

Reed, G. M., Kemeny, M. E., Taylor, S. E. & Visscher, B. R. (1999). Negative HIV-specific expectancies and AIDS-related bereavement as predictors of symptom onset in asymptomatic HIV-positive gay men. *Health Psychology, 18*, 354–363.

Reed, G. M., Kemeny, M. E., Taylor, S. E., Wang, H.-Y. J. & Visscher, B. R. (1994). "Realistic acceptance" as a predictor of decreased survival time in gay men with AIDS. *Health Psychology, 13*, 299–307.

Roes, F. L. & Raymond, M. (2003). Belief in moralizing gods. *Evolution and Human Behavior, 24*(2), 126-135.

Rossano, M. J. (2007). Supernaturalizing social life: Religion and the evolution of human cooperation. *Human Nature, 18*, 272–294.

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

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*(6), 383-384.

Rozin, P., Millman, L. & Nemeroff, C. (1986). Operation of the laws of systematic magic in disgust and other domains. *Journal of Personality and Social Psychology, 50*(4), 703-712.

Ruffle, B. J. & Sosis, R. (2007). Does it pay to pray? Costly ritual and cooperation. *The B.E. Journal of Economic Analysis and Policy, 7*(1 (Contributions)), Article 18.

Schipper, L. D., Easton, J. A. & Shackelford, T. K. (2007). Morbid jealousy as a function of fitness-related life-cycle dimensions. *Behavioral and Brain Sciences, 29*(6), 630.

Schloss, J. P. (2006). Evolutionary theory and religious belief. In P. Clayton and Z. Simpson (Eds.), *Oxford handbook of religion and science*. New York: Oxford University Press.

Shariff, A. F. & Norenzayan, A. (2007). God is watching you: Priming God concepts increases prosocial behavior in an anonymous economic game. *Psychological Science, 18*(9), 803-809.

Silva, J. A., Ferrari, M. M., Leong, G. B. & Penny, G. (1998). The dangerousness of persons with delusional jealousy. *Journal of the American Academy of Psychiatry and the Law, 26*, 607– 623.

Simon, H. A. (1990). A mechanism for social selection and successful altruism. *Science, 250*(4988), 1665-1668.

Smith, D. L. (2006). In praise of self-deception. *Entelechy: Mind and Culture, 7*.

Sosis, R. (2004). The adaptive value of religious ritual. *American Scientist, 92*, 166-172.

Stephens, G. L. & Graham, G. (2004). Reconceiving delusion. *International Review of Psychiatry, 16*(3), 236-241.

Stich, S. (1990). *The fragmentation of reason*. Cambridge, MA: The MIT Press.

Stone, T. & Young, A. W. (1997). Delusions and brain injury: The philosophy and psychology of belief. *Mind and Language, 12*, 327-364.

Tallis, F. (2005). *Love sick.* London: Arrow Books.

Taylor, S. E. (1989). *Positive illusions: Creative self-deception and the healthy mind*. New York: Basic Books.

Taylor, S. E. & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103*, 193–210.

Taylor, S. E. & Brown, J. D. (1994). Positive illusions and well-being revisited: Separating fact from fiction. *Psychological Bulletin, 116*(1), 21-27.

Taylor, S. E., Kemeny, M. E., Bower, J. E., Gruenewald, T. L. & Reed, G. M. (2000). Psychological resources, positive illusions, and health. *American Psychologist, 55*, 99–109.

Taylor, S. E., Kemeny, M. E., Reed, G. M., Bower, J. E. & Gruenewald, T. L. (2000). Psychological resources, positive illusions, and health. *American Psychologist, 55*, 99–109.

Taylor, S. E., Lerner, J. S., Sherman, D. K., Sage, R. M. & McDowell, N. K. (2003). Are self-enhancing cognitions associated with healthy or unhealthy biological profiles? *Journal of Personality and Social Psychology, 85*(4), 605–615.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46*, 35–57.

Trivers, R. L. (1974). Parent-offspring conflict. *American Zoologist, 14*, 249-264.

Trivers, R. L. (1985). *Social evolution*. Menlo Park, CA: Benjamin-Cummings.

Trivers, R. L. (2000). The elements of a scientific theory of self-deception. In D. LeCroy & P. Moller (Eds.), *Evolutionary perspectives on human reproductive behavior. Annals of the New York Academy of Sciences* (pp. 114-131).

Trivers, R. L. (2006). Foreword to Richard Dawkins' *The Selfish Gene*. In *The selfish gene: 30th anniversary edition* (pp. xix-xx). Oxford: Oxford University Press.

Vaillant, G. (1977). *Adaptation to life*. Boston: Little, Brown.

Van Leeuwen, D. S. N. (2007). The spandrels of self-deception: Prospects for a biological theory of a mental phenomenon. *Philosophical Psychology, 20*(3), 329–348.

Vazquez, C., Diez-Alegria, C., Hernandez-Lloreda, M. J. & Moreno, M. N. (2008). Implicit and explicit self-schema in active deluded, remitted deluded, and depressed patients. *Journal of Behavior Therapy and Experimental Psychiatry, 39*, 587-599.

Voland, E. (2008). *The evolution of morality – What is conscience good for? How cooperative breeding might pave another route to altruism.* Paper presented at the The XIX Biennial Conference of the International Society for Human Ethology (ISHE08), Bologna, Italy.

Voland, E. & Voland, R. (1995). Parent-offspring conflict, the extended phenotype, and the evolution of conscience *Journal of Social and Evolutionary Systems, 18*(4), 397-412.

Voltaire, F. M. A. (1759/1962). *Candide* (T. G. Smollett, Trans.). New York: Washington Square Press.

Wallace, B. (1973). Misinformation, fitness and selection. *American Naturalist, 107*, 1-7.

Wenger, A. & Fowers, B. J. (2008). Positive illusions in parenting: Every child is above average. *Journal of Applied Social Psychology, 38*(3), 611-634.

Williams, A. F. (2003). Views of U.S. drivers about driving safety. *Journal of Safety Research, 34*(5), 491-494.

Wilson, D. S. (2002). *Darwin's cathedral: Evolution, religion and the nature of society*. Chicago: University of Chicago Press.

Wong, T. Y., Foster, P. J., Hee, J., Ng, T. P., Tielsch, J. M. & Chew S. J. et al. (2000). Prevalence and risk factors for refractive errors in an adult Chinese population in Singapore. *Investigative Ophthalmology and Visual Science, 41*, S324.

Yamagishi, T., Terai, S., Kiyonari, T., Mifune, N. & Kanazawa, S. (2007). The Social Exchange Heuristic: Managing errors in social exchange. *Rationality and Society, 19*(3), 259–291.

Young, A. W. (1999). Delusions. *The Monist, 82*(4), 571-589.

Young, A. W. (2000). Wondrous strange: The neuropsychology of abnormal beliefs. *Mind and Language, 15*(1), 47-73.

Young, A. W., Robertson, I. H., Hellawell, D. J., de Pauw, K. W. & Pentland, B. (1992). Cotard delusion after brain injury. *Psychological Medicine, 22*, 799-804.

Zahavi, A. (1995). Altruism as a handicap - the limitations of kin selection and reciprocity. *Journal of Avian Biology, 26*, 1-3.

Zolotova, J. & Brüne, M. (2006). Persecutory delusions: Reminiscence of ancestral hostile threats? *Evolution and Human Behavior*, *27*(3), 185 - 192.

Acknowledgements

The first author was supported by a research fellowship as part of a large collaborative project coordinated from the Centre for Anthropology and Mind (http://www.cam.ox.ac.uk) at the University of Oxford and funded by the European Commission's Sixth Framework Programme ("Explaining Religion"). Thanks to Tim Bayne, Fabrizio Benedetti, Max Coltheart, Zoltan Dienes, Charles Efferson, Ernst Fehr, Philip Gerrans, Nick Humphrey, Robyn Langdon, Genevieve McArthur, Fabio Paglieri, Martha Turner and Harvey Whitehouse for useful input and interesting discussions. We also thank Paul Bloom, Stephen Stich, three anonymous reviewers, Jesse Bering, Ben Bradley, Mitch Hodge, David Hugh-Jones, Josh Sadlier, Konrad Talmont-Kaminski, Neil van Leeuwen and the participants in the Lyon Institute for Cognitive Sciences virtual conference on “Adaptation and Representation” (http://www.interdisciplines.org/adaptation) for valuable comments on earlier drafts of this paper.