BEHAVIORAL AND BRAIN SCIENCES (2009) **32**, 493–561
doi:10.1017/S0140525X09990975

The evolution of misbelief

Ryan T. McKay
Institute for Empirical Research in Economics, University of Zurich, Zurich 8006, Switzerland; and Centre for Anthropology and Mind, University of Oxford, Oxford OX2 6PE, United Kingdom
ryantmckay@mac.com
http://homepage.mac.com/ryantmckay/

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[1] 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.[2] A misbelief, then, is a belief that to some degree departs from actuality – that is, 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 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

RYAN T. MCKAY is a Research Fellow at the University of Oxford, UK. He was educated at the University of Western Australia (B.Sc. Hons. in Psychology) and at Macquarie University (MClinPsych, Ph.D.) in Sydney, Australia. His research interests include cognitive neuropsychiatry, evolutionary psychology, and behavioural economics. He has held previous postdoctoral positions in Boston (Tufts University), Belfast (Queen's University), and Zürich (University of Zürich). He has also previously worked as a clinical neuropsychologist at the National Hospital for Neurology and Neurosurgery in London and as a lecturer in psychology at Charles Sturt University in Australia. Recently, he has been conducting experimental investigations in the cognitive science of religion.

DANIEL C. DENNETT is University Professor, Fletcher Professor of Philosophy, and Co-Director of the Center for Cognitive Studies at Tufts University. He is the author of *Consciousness Explained* (1991), *Darwin's Dangerous Idea* (1995) and *Breaking the Spell: Religion as a Natural Phenomenon* (2006), as well as other books and articles in philosophy of mind, cognitive science, and evolutionary theory. He is also the author or co-author of three target articles (1983, 1988, 1992) and 32 commentaries in *Behavioral and Brain Sciences* and was an Associate Editor of the journal for many years.

 0140-525X/09 $40.00

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 sect. 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?[3]

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 target article 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; others may be false but not demonstrably so; 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 10 seconds every day (contra the advertisement).
2. My cheap Schmasio watch loses 10 minutes every day (contra the advertisement).

Examples of the latter limitation include:

1. My toaster does not keep time at all.
2. My Bolex loses a second every day (within warranted limits).
3. My cheap Schmasio loses a minute every day (within warranted limits).
4. 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,[4] 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 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

https://doi.org/10.1017/S0140525X09991440 Published online by Cambridge University Press

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 10 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,[5] 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.[6] 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. (Millikan 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 instead its history.[7] Given that organ transplants have not featured

https://doi.org/10.1017/S0140525X09991440 Published online by Cambridge University Press

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 explanation below) – and its limitations are "forgivable."

Let us consider a further parallel with our proposed misbelief taxonomy, this time by examining two types of *misperception*. 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),[8] 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 "small-n" normal on statistical grounds. After all, the prevalence of myopia varies across ethnic groups, and is as high as 70–90% in some Asian populations (Chow et al. 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), 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 *neuropsychology*, 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 sect. 9), delusions are misbeliefs *par excellence* – false beliefs that are held with strong conviction regardless of counterevidence 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 et al. (2001) investigated two cases of this delusion and uncovered two apparent routes to its development. The delusion of the first patient ("F.E.") 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 ("T.H.") was intact. This patient, however, appeared to be "mirror agnosic" (Ramachandran et al. 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 et al. 2000, for transcripts of interviews with the two patients; see Feinberg 2001, Feinberg & Shapiro 1989, and Spangenberg et al. 1998, 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 et al. 1999), and non-delusional prosopagnosia is quite common. Breen et al. (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 et al., in press; Davies & Coltheart 2000; Davies et al. 2001; Langdon & Coltheart 2000; McKay et al. 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*) consists of 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, the Capgras delusion (the belief that a loved one has been replaced by an impostor) and the 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 et al., in press), 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 et al. (2009) have suggested that whereas some delusions – for example, mirrored-self misidentification – might involve the former tendency (see Langdon & Coltheart 2000; Langdon et al. 2006; Stone & Young 1997; also see Huq et al. 1988), others – for example, delusional denial of paralysis ("anosognosia") – might involve the latter (see Ramachandran 1994a; 1994b; 1995; 1996a; 1996b; 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).[9] 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).[10] 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; 2006b; Dennett 1995a; 2006a). Although such cases are interesting in their own right, the adaptive misbeliefs we pursue in this article 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 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 the Cotard delusion may occur subsequent to neuropsychological damage. For example, Young et al. (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?
— William Shakespeare (*Richard II*, Act I, scene iii, lines 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.*
— Steven 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.
— Michael. T. 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 sect. 10). Thus in Shakespeare's *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 7:45 am. . . . 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 refute 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 (1983)

Theoretical considerations converging from several different research traditions suggest that any such systematic falsehood must be unstable, yielding ephemeral instances,

https://doi.org/10.1017/S0140525X09991440 Published online by Cambridge University Press

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 the 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 2008, 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 experiments with disgust (e.g., Rozin et al. 1986) 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 (OCD)

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.*
>
> — Stephen Stich (1990, p. 62)

> *[T]he human mind shows good design, although it is design for fitness maximization, not truth preservation.*
>
> — Martie Haselton and Daniel Nettle (2006, p. 63)

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, footnote 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. (Stich 1990, pp. 61–62)

What these examples suggest is that when there are reliable "asymmetries in the costs of errors" (Bratman 1992) – that is, 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 for 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 in members of other racial and ethnic groups (Haselton & Nettle 2006; Krebs & 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 et al. 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 overinferring 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 instill 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. Some authors, however, have suggested that certain *delusions* also involve error management processes. Schipper et al. (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: "[H]ypersensitive jealousy mechanisms . . . may serve the adaptive purpose of preventing partner infidelity" (Schipper et al. 2007, p. 630; see also Easton et al. 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 et al. 1998) – other deluded individuals frequently fail to act in accordance with their delusions. Individuals with the 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 article 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. The shear pin locks the propeller to the propeller shaft and 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,[11] 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 & Kinsbourne, in press; McKay et al. 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, "B.X.," was a gifted musician who had been left a quadriplegic following a car accident. B.X. 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, B.X.'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, it comes from an isolated case study, and Butler's theorising is unavoidably post hoc. Moreover, the fact that B.X. 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 et al. 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 et al. 2007b; Moritz et al. 2006; although, see Vazquez et al. 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, that is, 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; 2003a; 2003b).

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 2004b; Dawkins 2006a; Ruffle & Sosis 2007; Sosis 2004), 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 it is also apparent in some less-contemporary thinking:

> God is, or is not. … Let us weigh up the gain and the loss by calling heads that God exists … if you win, you win everything; if you lose, you lose nothing. Wager that he exists then, without hesitating! (Pascal 1670/1995, pp. 153–54.)

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 further on), 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 & Krüger 2004; Johnson et al. 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 et al. 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 et al. 2003; Henrich & Fehr 2003).

Johnson, Bering, and colleagues (Bering & Johnson 2005; Johnson 2005; Johnson & Bering 2006; Johnson & Krüger 2004; Johnson et al. 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. Johnson, Bering, and colleagues thus argue for supernatural belief as an 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 middleman, 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 et al. 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 – that is, that 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. Johnson 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" (Johnson 2005, 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 et al. 2007). Thus, much as people walk more slowly when the concept "elderly" is primed (Bargh et al. 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 et al. 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 both the "religious prime" and the "secular prime" categories utilized by Shariff and Norenzayan 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 et al. (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 et al. (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., Dawkins 2006a; Sosis 2004; Wilson 2002), the currently dominant evolutionary perspective on religion remains a by-product perspective (Atran 2004; Atran & Norenzayan 2004; Bloom 2004; 2005; 2007; Boyer 2001; 2003; 2008b; 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 (Barrett 2000; Guthrie 1993), 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 et al. 2002; Gintis 2000) argue that the puzzle of large-scale human cooperation may be solved by invoking cultural group selection (Boyd et al. 2003; Henrich & Boyd 2001) or gene-culture coevolution (Bowles et al. 2003; also 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. (Trivers 2006, p. xx; see also Alexander 1979; 1987; Lockard 1978; 1980; Lockard & Paulhus 1988; Trivers 1985; 2000)

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 et al. (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 Prisoner's 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.*
> — Marie Jahoda (1958, p. 6)

> *[T]he healthy mind is a self-deceptive one.*
> — Shelley 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.[12] 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 (see, e.g., McKenna et al. 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 et al. 2002; Pronin 2004).

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 (1994b), 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 Benabou & Tirole 2002; Krebs & Denton 1997).

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 et al. 1996; Fowers et al. 2001; Murray et al. 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 fewer 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 et al. 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 et al. 1999) and a longer survival time (Reed et al. 1994; for a review, see Taylor et al. 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), 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,

p. 246). 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 is 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*" (Humphrey 2002, pp. 261 and 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 by the time of 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; and so forth. 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 et al. 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 that of 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 regardless of whether 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) and 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 instilling 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.
— Paul Bloom (2004, pp. 222–23)

[S]ystematic bias does not preclude a tether to reality.
— Martie Haselton and Daniel 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 subpersonal 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.

https://doi.org/10.1017/S0140525X09991440 Published online by Cambridge University Press

We began this article 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 evolution has equipped us with for forming and testing beliefs will be less than "optimal" – and that sometimes it will break. Moreover, we have seen a variety of ways in which these suboptimal 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.

ACKNOWLEDGMENTS

The first author, Ryan McKay, 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 "Adaptation and Representation"(http://www.interdisciplines.org/adaptation), for valuable comments on earlier drafts of this paper.

NOTES

1. Doxastic = of or pertaining to belief.

2. 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.

3. For ease of exposition, we tend to conflate "adaptive" and "adapted" throughout this target article. 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.

4. 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 case of DVD region codes. 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).

5. Stephen Stich, in a 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."

6. 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 designers for their intentions where artifacts are concerned. Most likely the designers 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.

7. Fodor (2007) has vigorously challenged not just Millikan's claim, but also 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; 2007; 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* known to be for flying.

8. 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 et al. (1991).

9. Nevertheless, evolved cognitive systems are remarkably supple, as researchers in Artificial Intelligence (AI) 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.

10. 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).

11. 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.

12. 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).

Open Peer Commentary

When is it good to believe bad things?

doi:10.1017/S0140525X09991142

Joshua M. Ackerman[a], Jenessa R. Shapiro[b], and Jon K. Maner[c]

[a]*Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02142;* [b]*Department of Psychology, University of California, Los Angeles, Los Angeles, CA 90095-1563;* [c]*Department of Psychology, Florida State University, Tallahassee, FL 32306-4301.*

joshack@mit.edu

http://web.mit.edu/joshack/www/

jshapiro@psych.ucla.edu maner@psy.fsu.edu

http://www.psy.fsu.edu/faculty/maner.dp.html

Abstract: Positive and negative misbeliefs both may have evolved to serve important adaptive functions. Here, we focus on the role of negative misbeliefs in promoting adaptive outcomes within the contexts of romantic relationships and intergroup interactions. Believing bad things can paradoxically encourage romantic fidelity, personal safety, competitive success, and group solidarity, among other positive outcomes.

In their article, McKay & Dennett (M&D) define evolved misbeliefs, or illusions, as those that are adaptively superior to fully accurate beliefs. The authors focus their discussion on the value of positive misbeliefs, but there are also reasons to believe that negative misbeliefs can serve adaptive functions as well. In this commentary, we consider negative misbeliefs within two important social contexts: (1) close relationships and (2) intergroup interactions.

Misbeliefs related to close relationships. The formation and maintenance of close relationships are fundamental human pursuits (Ackerman & Kenrick 2008; Kenrick et al., in press). Romantic relationships are particularly important because mating represents the sine qua non of evolutionary success. Positive misbeliefs may aid these romantic pursuits, as in M&D's example of the over-perception of positive spousal attributes. However, close relationships may also benefit from negative illusions. For example, women tend to believe that men are less interested in romantic commitment than those men actually are (Haselton & Buss 2000), especially prior to the onset of sexual activity in relationships (Ackerman et al., submitted). M&D suggest that, although the system that generates such misbeliefs is probably adaptive, the misbeliefs themselves are not (because accurate beliefs would be equally protective without suffering from false positive errors). However, underestimating male commitment could lead women to set higher thresholds for suitors to overcome, leading men to expend greater effort and investment in courtship (see Ackerman & Kenrick 2009), and ultimately boosting the romantic returns that women receive (e.g., mate quality, economic resources, actual commitment). Comparatively, accurate beliefs about potential romantic partners might facilitate accurate decision making, but would be unlikely to garner these additional benefits.

Another example pertains to misbeliefs about alternative relationship partners. People in committed relationships tend to display cognitive biases that inhibit straying from those relationships (e.g., Maner et al. 2008; 2009), such as believing that attractive relationship alternatives are less appealing than they actually are (Johnson & Rusbult 1989; Simpson et al. 1990). These negative illusions down-regulate threats posed by romantic alternatives, increasing the long-term success of one's current relationship. Long-term romantic relationships serve important functions linked to social affiliation and offspring care, as well as providing more obvious reproductive benefits, and thus negative misbeliefs about relationship alternatives can promote a range of adaptive outcomes. Accurate beliefs about attractive alternatives, however, could promote infidelity and destabilize one's relationship.

Misbeliefs related to intergroup interactions. In addition to romantic relationships, group-level relationships are also fundamental components of human evolutionary success (Kenrick et al., in press; Neuberg & Cottrell 2006). Throughout human evolutionary history, hostile outgroups have posed threats to personal safety and group resources. Many of these threats were transient, with periods of conflict interspersed with periods of relative peace (e.g., Baer & McEachron 1982). Accurate beliefs acknowledging that outgroups were not always threatening could have supported increased intergroup contact. However, the potential for threat in intergroup interactions would likely remain high, as initially peaceful or cooperative encounters between unfamiliar parties can quickly turn dangerous (e.g., through simple misunderstandings or signals of vulnerability). Negative outgroup illusions could have enhanced fitness to the extent that they led people to be wary, reducing the probability of loss or harm from a hostile outgroup member (see Ackerman et al. 2006; 2009).

In fact, negative misbeliefs can strengthen the drive to compete with other groups for status and resources (Campbell 1965; Sherif et al. 1961). For example, sports teams may perform better because of the misbeliefs they hold about the motivation and skill of their rivals. Similarly, religions may facilitate conversion by asserting the falsity and profaneness of other gods. In the political realm, nations are frequently in conflict with one another over natural and social resources, and exhibit extreme ideological and ethnocentric beliefs as a result (Campbell 1965). Governments that construe other nations as "Evil Empires" may be more motivated to economically out-produce and even attack those nations (thereby attaining resources, if they win). In contrast, accurate beliefs about opposing groups would provide no extra incentive to compete and might even de-motivate groups with relatively lower standing and abilities.

Much of the work on negative misbeliefs and intergroup threat has explored the role of race as a heuristic cue to group membership. People tend to associate particular racial groups with specific threats (e.g., Black males with physical danger; Cottrell & Neuberg 2005), and these biases become especially strong in the presence of other threat-relevant cues (e.g., angry expressions). For example, people believe that neutrally expressive outgroup men are more threatening when seen in the context of other, angry outgroup men (Shapiro et al. 2009);

frightened people believe that outgroup men are more angry than they truly are (Maner et al. 2005); and pregnant women, whose fetuses are especially vulnerable early in development, exhibit greater ethnocentric beliefs during their first trimester (Navarrete et al. 2007). Such negative illusions could promote outgroup avoidance (see also Mortensen et al., in press) which, in evolutionary contexts, could have served important self-protective functions.

Finally, misbeliefs about outgroup threat elicit not only outgroup avoidance, but also ingroup solidarity (Becker et al., submitted; Coser 1956; Tajfel & Turner 1986). This solidarity provides a number of advantages. Consider that the pursuit of economic and physical resources is often a zero-sum game, and thus groups must manage their resources by discouraging exploitation from selfish members. Cooperation is one solution to potential intragroup conflict, and negative illusions about the dangers of other groups may improve cooperation by providing a common threat and promoting intragroup unity (e.g., Hammond & Axelrod 2006; Van Vugt et al. 2007).

Conclusion. Many negative misbeliefs continue to provide adaptive benefits in modern times, and yet may also result in detrimental social outcomes such as the perpetration of problematic stereotypes and prejudices. Despite such modern troubles, there is reason to believe that, as with positive misbeliefs, negative misbeliefs evolved to meet recurrent challenges in the ancestral world.

Non-instrumental belief is largely founded on singularity[1]

doi:10.1017/S0140525X09991154

George Ainslie

151 Coatesville VA Medical Center, Coatesville, PA 19320.

george.ainslie@va.gov www.picoeconomics.org

Abstract: The radical evolutionary step that divides human decision-making from that of nonhumans is the ability to excite the reward process for its own sake, in imagination. Combined with hyperbolic over-valuation of the present, this ability is a potential threat to both the individual's long term survival and the natural selection of high intelligence. Human belief is intrinsically "unfounded" or under-founded, which may or may not be adaptive.

McKay & Dennett (M&D) depict the category of adaptive groundless beliefs as a small, albeit fascinating, exception to their "default presumption – that true beliefs are adaptive and misbeliefs maladaptive" (target article, sect. 15, para. 3). They review many kinds of examples, such as self-confirming beliefs (placebos), beliefs that by their nature cannot be tested (faiths), and beliefs that could be tested but are not (delusions). The striking feature of these cases is that they are not sharply demarcated from grounded beliefs, and thus represent not a small cabinet of curiosities but demonstrations of a basic inadequacy in the conventional understanding of belief. The authors start toward repairing this inadequacy by pointing out that in many cases, "[beliefs'] adaptiveness is independent of their truth or falsity" (sect. 14, para. 2). This implies that beliefs are ultimately selected for functionality, but it raises the question exemplified by M&D's quote from Humphrey: "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, cited in sect. 13). The authors analyze the problem in terms of adaptiveness, but really cannot do without a key intervening variable, reward.

It is certainly true that "the driving force behind natural selection is survival and reproduction, not truth" (from Bloom [2004], quoted in sect. 15). However, evolution has developed the reward process as a proxy for survival and reproduction, outcomes that are too global and usually too distant to select the behaviors of individual organisms. Although "survival is the only hard currency of natural selection" (sect. 15. last para.), it affects choice only by backing the token currency of reward; to the extent that organisms are engineered to learn at all, they are engineered to maximize prospective reward, which is the quantity that must have an "exchange rate with truth" (sect. 15, concluding para.). A further constraint is that the valuation of prospective reward seems to be fundamentally tied to the Weber-Fechner law by which most psychophysical quantities are perceived (Gibbon 1977), causing it to be discounted for delay in a hyperbolic curve rather than a "rational," exponential curve (Green & Myerson 2004; Kirby 1997). The exchange rate of reward with truth is necessarily close to parity when animals have only "aliefs," not beliefs (sect. 8), and the hyperbolic over-valuation of imminent rewards should not matter when animals' long-term interests are served by instincts that make long-term preparations such as hoarding, dam building, and migrating rewarding in the short term. However, with selection for increasing intelligence has come increasing imagination, and with imagination the unhitching of reward from adaptiveness, of short-term from long-term interests, and of belief from truth.

Imagination is governed by reward. It discovers short cuts that detach reward contingencies from the adaptive functions that originally selected for them. People have learned to mate without reproducing, fight without needing to, and commit themselves to costly hobbies that do not contribute to surviving offspring. Furthermore, we have learned to rob future welfare for present pleasure, not just with addictive substances but also with socially accepted activities ranging in excitement level from death-defying adventure to simple procrastination (Ainslie 2010). Most important for the present discussion, we have learned to make occasions for current emotional reward from events that are not currently happening, in the form of memories, fantasies – and beliefs (as opposed to aliefs). Intelligence obviously has many adaptive features, but its ongoing evolution must be limited – probably has already been limited – by the availability of constraints on the urge for diverting long-term resources to current consumption. This is the context in which we need to address the question of why people cannot get well, or get confident, or get happy, "without having to wait . . . for outside permission."

Where reward is strongly bound to survival resources – food, warmth, avoidance of injury – the cost of misbelief will be deprivation or pain, so instrumental beliefs will be constrained mostly by their predictiveness, as the authors also note. Where hard-wired sensations are not involved, or even where they are significantly delayed, the prospective benefits and costs of belief obviously depend on the reward that can be expected from imagination. Sources of this reward vary widely; for instance: sublime fantasy, puzzle-solving, vicarious adventure, or gratification of urges to obey compulsions or entertain anxiety or disgust. We learn to imagine various scenarios on various occasions based on the patterns of reward that ensue, cultivating some feelings and avoiding or resisting the urges for others. Belief might be best defined as the faculty that directs imagination so as to improve long-term outcomes, relative to the results of spontaneous immersion in the moment; but this improvement is measured in reward, which corresponds to adaptiveness only to the extent that evolution has had time to modify the proxy function of reward to keep up with increasing *Homo* intelligence. And there remains motivational pressure for belief to serve spontaneous immersion, in the form of wishful thinking.

I have described elsewhere how hyperbolic discounting of reward predicts regularities in the competition of reward-seeking processes (Ainslie 2001, pp. 48–104, 161–97; 2005). Here the important aspect is that imagination ad lib exhausts itself in premature payoffs. When one occasion for reward is as

good as another, they will replace each other randomly, and the imagining will have the quality of a daydream. Conversely, if there is a single, relatively rare occasion that stands out from the others, it will make the corresponding imagination robust. The experience of such *singularity* may be much like that of having solved a puzzle or detected a fact of nature. The occasion in question will stand out from the common ruck of imaginings just as a fact stands out from a fantasy.

Where information about the natural world is absent or ambiguous, singularity may be the best clue about how it functions – parsimony is a decent starting place for theories. But a belief that distinctly delivers good news and bad news will be productive of reward in its own right, regardless of its eventual accuracy. The emotional effectiveness of singular occasions may be experienced as a kind of factuality, more or less confounded with the factuality that comes from physical observation. In the most conspicuous cases, remembered events are experienced again on their anniversaries, especially when the anniversary is a round number; original works of art are felt to be more "real" than exact copies; and placebos (as in sect. 13) are effective in proportion to the expensiveness of the ingredients or the prestige of the healer. Even realistic beliefs get additional value by serving as occasions for emotional reward, as in the "drug effect" of money (Lea & Webley 2006). Conversely, faced with unwelcome urges such as hypochondria, phobic anxiety, or a sense of being dirty, a person searches for a favorable interpretation of the situation – whether she can feel well, or safe, or clean. This interpretation cannot be arbitrary; wishes have little impact. She must choose her belief on the basis of "facts" that she discerns in events beyond her control – a pill given by a doctor, a lucky charm or safety signal, or a "scientific" disinfectant. The belief may even become stabilized as a personal rule: in effect, "I will not give in to panic or disgust when this signal is present." The same role of singularity can be seen in many other misbeliefs. For instance, delusions (sect. 9) tend to be based on a logical deduction or a remarkable coincidence, and religious faiths (sect. 11) depend on the singularity that comes from having had long histories of consensual agreement – hence their fear of heresies. It would be fruitless to try to decide whether such hedonically based beliefs are more or less adaptive than veridicality; evolution veered away from veridicality with the apes.

NOTE

1. The author of this commentary is employed by a government agency, and as such this commentary is considered a work of the U. S. government and not subject to copyright within the United States.

False beliefs and naive beliefs: They can be good for you

doi:10.1017/S0140525X09991178

Marco Bertamini[a] and Roberto Casati[b]

[a]*School of Psychology, University of Liverpool, Liverpool L69 7ZA, United Kingdom;* [b]*CNRS Institut Nicod, Ecole Normale Supérieure, 75005 Paris, France.*

m.bertamini@liv.ac.uk **http://www.liv.ac.uk/vp/**
casati@ehess.fr **http://www.institutnicod.org**

Abstract: Naive physics beliefs can be systematically mistaken. They provide a useful test-bed because they are common, and also because their existence must rely on some adaptive advantage, within a given context. In the second part of the commentary we also ask questions about when a whole family of misbeliefs should be considered together as a single phenomenon.

If humans are biologically engineered to appraise the world accurately, how can we explain misbeliefs? After asking this question, McKay & Dennett (M&D) analyse various misbeliefs. Those resulting from a breakdown in the system, and those that are by-products, do not threaten the claim of adaptiveness of the belief system. Positive illusions are the only bona fide example of misbeliefs. We shall integrate this account by first making a case for the adaptiveness of some mistakes in the conception of the physical world, and by discussing the possibility of a general egocentric bias in generating positive illusions.

The grand aim of Naive physics (NP) is to fully describe common beliefs about the physical world. Naive physics can be traced back to Gestalt psychologists such as Köhler, and to the seminal work by Lipmann and Bogen (1923). The term is also used in artificial intelligence and robotics (Hayes 1978). Despite its grand aim, interest in NP has focused on the discovery that people make some systematic mistakes about everyday phenomena. Examples include judgements about the pendulum motion (Bozzi 1958); predictions of motion of an object in terms of direction, path of motion, and acceleration (Hecht & Bertamini 2000; McCloskey et al. 1980); and predictions about what is visible in a mirror (Bertamini & Parks 2005). In the case of the pendulum, people consider as "natural" a movement that is actually artificially contrived. We can be sure that some mistakes are not cultural whims because they match scientific theories of the past (i.e., Aristotelian mechanics). NP beliefs are not necessarily approximations or simplified representations of the physical world (Cavanagh 2005). In some cases the implied physics is complex, for instance, when subjects deem as correct cast shadows that require light to bend around corners or to be projected from physically impossible locations (Casati 2008).

Even if these mistakes are the manifestation of (implicit) mental models (McCloskey 1983), where do these models come from? Typically NP beliefs are resilient and non-revisable, thus pointing to some modular underlying mechanism. Some NP beliefs are grounded on evidence provided by the visual system. The belief that a pendulum looks unnatural when it moves, for example, originates from how people perceive motion (Bozzi 1958; Pittenger 1989). Aspects of how people reason are also important, as exemplified by the reliance on prototypes of actions (Yates et al. 1988) and heuristics (Proffitt 1999). Mistaken beliefs that originate from properties of perceptual or reasoning mechanisms could be classified as evolutionary by-products. On the other hand, one can ask the question of why these as opposed to other by-products occur. System limitations should also be considered from an evolutionary standpoint. For example, if waitresses make larger mistakes than housewives in the water-level task (the orientation of water in a tilted glass) this may be because the glass as a frame of reference is more important to them in their job than it is to other people (Hecht & Proffitt 1995). This may seem paradoxical but it suggests that attention to a local frame of reference, which is crucial for a task, makes it harder to learn about more abstract frames of reference. Context is, therefore, critical here. At least some NP beliefs, we surmise, are examples of systematically mistaken adaptive beliefs. In spite of their wrongness they provide contextually useful representations.

We are not claiming that each specific NP belief is an adaptation. Our perceptual system and our thoughts may lead us to them as a response to a situation. This brings us to the second point of our commentary.

Adaptiveness itself is hard to assess. Veridicality is not sufficient as a criterion. Just like percepts, most beliefs are prima facie veridical (they do not interfere with our interactions with the world) but compliance with logic or the laws of physics is not what they (beliefs as well as percepts) have evolved towards. An adapted organism is one that has accumulated characteristics that maximise fitness, not knowledge per se. Positive illusions are adaptive because they lead people to engage in adaptive behaviours. Whatever the mechanism, positive views of one's medical condition and of one's ability to influence it lead to increased health. Quite possibly the effects are not directly in terms of guiding deliberation and choice, rather they are ancillary

effects, such as triggering emotional adjustments and immune reactions. The evidence about biased responses concerning the self is vast, and controversial. It spans items as diverse as: self-serving biases and positive illusions (Taylor & Brown 1994b), implicit egotism (Pelham et al. 2005), narcissism (Nuttin 1985), self-enhancement (Sedikides & Gregg 2008), and self-resemblance and trust (DeBruine 2002), among others.

But are these beliefs specific adaptations or are they facets of a powerful but unspecific underlying mechanism, which we may call "looking after number one"? We think the jury is still out. If specific beliefs originate from specific adaptations, then it should be possible to find not only examples of "positive" illusions about oneself, but also of "negative" illusions about oneself that are, under different circumstances, adaptive. We would, therefore, need an example of a trait that is both generally perceived as positive (e.g., height) and yet such that people tend to see themselves as lacking because the resulting underestimation has a specific adaptive effect. If, on the contrary, we only have examples of overestimations (i.e., errors in the direction perceived as positive) then the most economical hypothesis is that they are all related, and originate from the same generic bias in favour of the self. Another problem with the idea that specific beliefs are specific adaptations is the fact that biases in favour of the self exist also for neutral or non-beneficial aspects. For instance, preferences are influenced by presence in their formulation of the first letter of the name of the person expressing the preference (Nuttin 1985); compliance with a request increases when someone is told that they share a birthday with the requester (Burger et al. 2004); and people overestimate the size of their own head (more than other people's heads) (Bianchi et al. 2008). It is unclear what the benefits are for these effects, and it seems more likely that they all originate from a generic (and adaptive) egocentric bias.

Extending the range of adaptive misbelief: Memory "distortions" as functional features

doi:10.1017/S0140525X09991397

Pascal Boyer

Department of Psychology, Washington University in St. Louis, St. Louis, MO 63130.

pboyer@artsci.wustl.edu

http://artsci.wustl.edu/~pboyer

Abstract: A large amount of research in cognitive psychology is focused on memory distortions, understood as deviations from various (largely implicit) standards. Many alleged distortions actually suggest a highly functional system that balances the cost of acquiring new information with the benefit of relevant, contextually appropriate decision-making. In this sense many memories may be examples of functionally adaptive misbelief.

Memory illusions or distortions are a major area of recent research (Brainerd & Reyna 2005; Roediger 1996; Schacter & Coyle 1995). They are very diverse, ranging from intrusions in word-list recall to therapy-influenced imaginings of previous lives or systematic abuse.

Dramatic memory distortions seem to influence belief-fixation. For instance, in the illusory truth effect, statements read several times are more likely rated as true than statements read only once. People who repeatedly imagine performing a particular action may end up believing they actually performed it (imagination inflation). Misinformation paradigms show that most people are vulnerable to memory revision when plausible information is implied by experimenters. In social contagion protocols, people tend to believe they actually saw what is in fact suggested by the confederate with whom they watched a video.

Another major type of distortion is revision of prior mental states under the influence of newly received information or changed contexts. People modify their autobiographical memories to fit implicit "theories of change." They, for instance, think that one gets better at a particular task with practice and therefore revise their memories of past performance to fit the predicted performance curve (Ross & Wilson 2003). In a similar way, in hindsight protocols people revise memories of their own prior guesses (e.g., that London has 10 million inhabitants) after receiving feedback information. Most familiar is attitude-revision, in which subjects routinely mis-remember previously held and subsequently changed attitudes.

These distortions seem to result from the normal standard operation of memory systems. Yet they result in misbelief. Why is that the case?

Distortion is a normative notion, so what is the standard against which memory systems are failing? Surprisingly, this is generally left implicit in memory research. In contrast to, say, decision-making, in which human "biases" are described as deviations from normative models, there are no explicit standards in memory research. That is because an explicit standard for memory performance would require a description of memory functions, and traditionally memory researchers have not been overly preoccupied by functional considerations, with a few exceptions (Anderson & Schooler 2000; Nairne et al. 2008).

As a consequence, memory performance is evaluated against generally tacit, apparently self-evident commonsense assumptions – we can infer those assumptions from the very fact that some memory processes are treated as "distortions." As mentioned above, it seems that they constitute deviations from a tacit and largely implausible view of memory systems. One assumption seems to be that memory as storage of information is not subject to the same cost-benefit constraints as the rest of cognition, so that information acquired should be stored rather than transformed, *pace* Bartlett (1932). Another assumption is that memory retrieval has its own function, independent from decision-making, so that one should, for instance, expect people to recall attitudes that did not lead to particular decisions.

But both assumptions are biologically odd. It makes obvious sense to consider memory retrieval as a biological function that comes at a cost and is therefore designed to maximize return on that cost (Dukas 1999). Also, it makes evolutionary sense to keep in mind that organisms do not develop cognitive abilities (e.g., retrieval of past experience) for abstract epistemic benefits (knowing what used to be the case). They retrieve information inasmuch as it helps fitness-enhancing decision-making in the present (Suddendorf & Corballis 2007).

Seen in this perspective, many cases of "distortion" appear highly functional. Consider misinformation and other situations in which memories are influenced by confederates' suggestions. The possibility and need of acquiring vast information from conspecifics also creates the possibility of error and deception. For each item of information, memory and decision-making systems must, implicitly or explicitly, assess the costs and benefits of including information in a belief-box or, alternatively, of keeping track of the information's "source-tag." It is certainly plausible that, *in some circumstances*, it is too costly to keep the source-tags for many items of information if they are all used to build a coherent, usable account of one's own experience. In the same way, repetition effects show that internal judgments of familiarity and fluency play an important role in decision-making. Intuitive epistemics here uses the external world regularity that *in some circumstances* true information is more frequent than false information. What matters for adaptive design is that the circumstances in question be such that this sort of decision-making does not lead to *excessive* vulnerability.

Now turn to attitude revision. In a functional perspective, accurate memory of past attitudes would be an odd proposition for a well-designed memory system. To preserve traces of past, now-irrelevant attitudes without compromising its computations, the system would need to quarantine them from on-line motivation and decision-making (Cosmides & Tooby 2000). The

extra cost of such computational "cordoning off" of memories may not be offset by the advantages, if any, of maintaining a record of past attitudes. In the same way, schema-based biased reconstruction of autobiographical memories, as occurs when people hold a particular, often implicit "theory of change" for a particular domain, may also contribute to efficient here-and-now decision-making by saving costs on specific but irrelevant episodic traces (Klein et al. 2002). Finally, a hindsight bias may constitute the most efficient way of making updated information more accessible than wrong information (Hoffrage et al. 2000). In such a perspective, the study of memory "distortions" could be part of a functional account of the systems involved, as is the case for perceptual illusions (Roediger 1996).

Is all this adaptive? An evolutionary perspective on memory cannot maintain the assumption of a frictionless, cost-free recording of experience that seems to be the implicit standard in memory research. Memory need be only as "good" as the advantage in decision-making it affords, measured against the cost of its operation (Nairne et al. 2008). This is why we go around assuming that we always knew what we now know, and believed the same beliefs; and we often construe as direct experience what we only know from others' reports – but all this is part and parcel of having a highly efficient memory system. If that is the case, it may well be that a great number of our memories, as beliefs about past occurrences, are instances of adaptive misbeliefs.

Positive illusions and positive collusions: How social life abets self-enhancing beliefs

doi:10.1017/S0140525X0999118X

Jonathon D. Brown

Department of Psychology, University of Washington, Seattle, WA 98195-1525.

jdb@uw.edu

http://faculty.washington.edu/jdb

Abstract: Most people hold overly (though not excessively) positive self-views of themselves, their ability to shape environmental events, and their future. These positive illusions are generally (though not always) beneficial, promoting achievement, psychological adjustment, and physical well-being. Social processes conspire to produce these illusions, suggesting that affiliation patterns may have evolved to nurture and sustain them.

In a classic scene from the Woody Allen movie, *Everything You Wanted to Know About Sex (But Were Afraid To Ask)*, sperm congregate in a holding area awaiting ejaculation. One sperm, played by Allen, is gripped by existential doubt as he contemplates his impending odyssey into the great unknown. As humorous as Allen's dilemma is, imagine his character's distress if he had paused to consider his odds of successfully fertilizing an egg (roughly 1 in 40,000,000, assuming ovulation). These odds would surely shake the confidence of even the most Panglossian spermatozoon, let alone Allen's anxiety-ridden schlemiel.

Of course, sperm do not calculate probabilities. But had nature endowed them with the ability to do so, she would have needed to similarly endow them with the ability to inflate their own likelihood of success. Otherwise, they, like Allen, would be paralyzed by the reality they were about to confront.

Extrapolating from Hollywood cinema is obviously hazardous, and spermatozoa are not people (although both function to pass their genetic material to the next generation), but Allen's scene touches on two important questions: Are people positively biased in their beliefs, and are these beliefs ultimately beneficial? In 1988, Shelley Taylor and I examined research relevant to these questions and offered two conclusions. First, when it comes to self-relevant beliefs and appraisals (e.g., "How kind am I?" "How capable am I?" "How bright is my future likely to be?"), people are positively biased (Taylor & Brown 1988; see also Brown 1986; 1991; 2007; Taylor 1989; Taylor & Brown 1994a; 1994b). On virtually all positively valued attributes, most people view themselves in unrealistically (though not excessively) positive terms. Second, we argued that these positive illusions are ordinarily beneficial. Under normal circumstances, people who entertain moderately (though not excessively) positive self-beliefs fare better on measures of achievement, adjustment, and physical well-being than those who are less positively biased. Certainly, there are limits to the benefits positive illusions provide (Baumeister 1989; Dunning et al. 2004), and we never claimed that the more biased one is, the better off one is going to be (see also, Marshall & Brown 2007). Instead, our claim was simply that (a) most people view themselves in overly positive terms and (b) under many – if not most – conditions, these beliefs are beneficial.

In their target article, McKay & Dennett (M&D) echo these arguments, concluding that positive illusions provide the firmest evidence for evolved misbelief. From this perspective, natural selection favored those whose self-perceptions were positively biased. In sympathy with this conclusion, my colleagues and I have found consistent evidence that positively biased self-perceptions are a pervasive, cross-cultural phenomena (Brown 2003; Brown et al. 2009; Brown & Kobayashi 2002; 2003; Cai et al. 2007; 2009; Kobayashi & Brown 2003).

At the same time, I think the target article would have benefited by taking a broader view of positive illusions. An exclusive focus on people's self-enhancing beliefs (e.g., "My commentary is more insightful than most other commentaries") ignores the myriad processes that conspire to produce and sustain them (Brown 1991). In most instances, positive illusions are the downstream product of an extensive system of information-processing biases and selective affiliation patterns. Insofar as these biases and patterns generate and perpetuate adaptive illusions, they may also be products of natural selection.

Numerous cognitive processes, such as self-serving attributions, idiosyncratic trait definitions, and biased judgments of a trait's importance sustain positive illusions (for a review, see Brown 1998), but interpersonal processes ordinarily produce them. For the most part, people believe positive things about themselves because they receive mostly positive feedback from the people they spend most of their lives with (Murray et al. 1996). In this sense, positive collusions produce positive illusions.

Positive collusions rely on two interrelated processes. First, people's self-enhancing biases include aspects of what William James (1890) called the "extracorporeal material self." The extracorporeal material self refers to everyone and everything we call "mine" or "my." With respect to positive illusions, we exaggerate not only our own virtues, but also those of our friends, neighbors, colleagues, family members, and loved ones (Brown 1986; 1991; Brown & Kobayashi 2002). Positive collusions begin the moment we are born. Most (though certainly not all) parents view their infants in overly positive terms, believing their offspring are cuter, smarter, and more socially advanced than are most other infants. As children grow, they internalize these biased evaluations, producing the well-known "better than average" effect (Alicke 1985; Brown 1986).

It is hardly surprising that parents view their infants through rose-colored glasses; what is surprising, however, is just how tenuous the self-other connection can be in order for this positivity bias to emerge. Research on in-group favoritism in the minimal group paradigm (Billig & Tajfel 1973; Tajfel et al. 1971) makes this point most graphically. In these studies, people are arbitrarily divided into groups on some patently trivial basis (e.g., they drew a blue marble from a bag instead of a red one). Despite the meaninglessness and obviously arbitrary nature of this designation, people view their fellow in-group members in more positive terms than out-group

members (Brewer 1979). In short, anything or anyone that is part of "me" is viewed in more positive terms than anything or anyone that is "not me."

How do these effects sustain people's beliefs in their own capacities? Once we forge an association with someone (e.g., make a friend; join a club; select a mate), we become part of that person's extracorporeal self and reap the self-enhancing benefits the association provides (i.e., we receive feedback that we are more likable, capable, and charming than are most other people). In this fashion, mutual admiration begets mutual benefits.

Ideology as cooperative affordance

doi:10.1017/S0140525X09991403

Joseph Bulbulia[a] and Richard Sosis[b]

[a]*Faculty of Humanities and Social Sciences, Victoria University, Wellington, New Zealand;* [b]*Department of Anthropology, U-2176, University of Connecticut, Storrs, CT 06269-2176.*

joseph.bulbulia@vuw.ac.nz

www.victoria.ac.nz/religion/staff/joseph_bulbulia/

richard.sosis@uconn.edu

www.anth.uconn.edu/faculty/sosis/

Abstract: McKay & Dennett (M&D) observe that beliefs need not be true in order to evolve. We connect this insight with Schelling's work on cooperative commitment to suggest that some beliefs – ideologies – are best approached as social goals. We explain why a social-interactive perspective is important to explaining the dynamics of belief formation and revision among situated partners.

Legend holds that on arriving at Veracruz, Cortés burned his ships so that his armies could not retreat. His men became predictably committed to fighting. Similarly, our contracts, emotions, affiliations, markings, gifts, punishments, and other costly acts anticipate our future responses. These factors transform partner options, enabling reliable forecasting of cooperative behaviors. Such predictability enhances cooperation's prospects for success. Schelling called these expressions "commitment devices" (Schelling 1960). His concept helps to explain otherwise perplexing behavior, but can it help explain belief? To think so might seem strange. Cortés allegedly burned his ships to motivate action, irrespective of belief. To generalize: If beliefs represent environments, the faculties that generate belief appear poorly equipped for predicting social commitment. Environments constantly change. Yet, a commitment device must anchor cooperative futures against these sea tides.

Nevertheless, certain beliefs – that the ship is burning, for example – proximately motivate social responses. The effect is well illustrated by religious commitment. Peter believes his God abides. From this conviction, Peter receives strong motivations, for example, to stand this holy ground, come what may. Like a boat on fire, his belief in God narrows Peter's strategic options, by overdetermining one. Where religious beliefs are shared, a universe of possible interactions strongly contracts, affording cooperation's success. Where religious commitment motives actions by sacred rewards, religious partners will suffer fewer distractions from personal risks. Cortés' sabotage does not promote cooperation through intrinsic reward; rather, it sets a trap. As such, it remains a poor instrument by which to disable anxiety, as slings and arrows rain down. Furthermore, where religious beliefs can be reliably recognized, fellow believers may find a common inspiration that they *know* to be common. The affective and symbolic cues of religious culture give what Schelling calls "salience" for otherwise risky coordination points. Notice, religious culture supports coordinated action for collective problems whose nature cannot be anticipated. At best, Cortés' act is only useful for the fight. Finally, religious beliefs can be evoked and assessed by ordeals that appear "crazy" without such beliefs (Irons 2008). Where opportunists threaten religious cooperation, evidence for commitment can be discerned from our deeds. To generalize: While actions are important to social commitment, beliefs intricately interact with actions and motivations to support effective social prediction (Bulbulia 2009). Such prediction requires shared epistemic habits that maintain common social goals as the world changes. We call the products of these habits "ideologies."

Ideologies function as commitment devices, though they function differently to burning boats. Indeed, commitment devices function best when we are unaware of their existence. In the Cortés legend, commitment arises through explicit means – removing the antisocial option: *Run away*! However, because motivations are affected by confidence, commitment theory predicts tendencies to strongly deny ideology's social causes. To think that ideology is believed for commitment, rather than as simple truth, enables one to second-guess one's ideology, and with it, the social commitments ideology inspries. This second-guessing may impair the social prediction so fundamental to cooperation's success. In their discussion of "alief," McKay & Dennett (M&D) observe how discrepancies sometimes arise between explicit knowledge (the bridge is safe) and implicit response (vertigo) (also explored in McKay & Cipolotti 2007; Dennett 1991). Commitment theory predicts the opposite relationship will hold too: consciousness will obscure motivations arising from collective goals (epistemic boat burning). For again, it is belief *as true* that motivates. We notice, however, that incorrigible persistence in believing, come what may, is unlikely to afford cooperative outcomes. Commitment theory predicts that ideologies will instead shift to meet strategic demands: Beliefs are subtle beasts.

There is much evidence for such subtlety. For example, Festinger et al. describe a UFO cult dealing with the pathos occasioned by the failure of a predicted doomsday (Festinger et al. 1956). While some cult members packed up and left, most remained, updating their beliefs to explain the persistence of life as the effect of the group's piety and prayer. Such intellectual *leger de main*, however striking, is not restricted to UFO-spotters. The dissonance literature shows that we often revise peripheral beliefs to meet our goals, not Bayesian demands. Such results are important to commitment models because they reveal that motivations shape our conscious beliefs, and so, that the link between belief and motivation is a two-way street. Moreover, commitment theory enriches dissonance models by focusing to the dynamics of goal maintenance for interactions whose success depends on reliable social prediction.

Organizations of the environments in which we interact (developmental and local) powerfully affect our cooperative commitments; their functional elaboration is critical to the explanation of ideology. While our understanding of these mind/world systems remains obscure, initial results reveal a fascinatingly strong capacity for sacred traditions (core elements of which have been conserved for centuries) to promote cooperative behaviors in large social worlds (Bulbulia, in press; Sosis 2000). For example, the neuroscience of charismatic authority suggests that neural circuits supporting ideological commitments are similar to those recruited during hypnotic suggestion (Deeley et al. 2003; Schjødt et al., submitted; Taves 2009). Charismatic authority appears to work like a trance. Other research shows that impersonal elements of culture – its music, symbolic displays, and large-scale ritual events – dramatically affect social sensibility and emotions, suggesting that charismatic enchantment extends to impersonal culture and its instruments (Alcorta & Sosis 2005; Baumgartner et al. 2006; Bulbulia, in press). Among these instruments, synchronous body practices appear especially effective at evoking and maintaining cooperative orbits (Hove & Risen, in press; Wiltermuth & Heath

2008). In other works we suggest that ritual, music, and symbolic practices are fundamental to establishing the informational and motivational settings that maintaining cooperative behaviors at small and large scales (Bulbulia 2004a; Bulbulia & Mahoney 2008; Sosis 2003; 2005).

To summarize, commitment theory is important to naturalistic study of belief because it reveals that a core subset of positive illusions are better approached as social goals, masquerading as beliefs. These ideologies interact with our social and cultural circumstances to promote accuracy, not in representing the world as it is, but rather in forecasting what we will do next.

Adaptive diversity and misbelief[1]

doi:10.1017/S0140525X09991415

Edward T. Cokely[a] and Adam Feltz[b]

[a]*Max Planck Institute for Human Development, Center for Adaptive Behavior and Cognition, 14195 Berlin, Germany;* [b]*Departments of Philosophy and Interdisciplinary Studies, Schreiner University, CMB 6208, Kerrville, TX 78028.*

cokely@mpib-berlin.mpg.de ADFeltz@schreiner.edu

http://faculty.schreiner.edu/adfeltz/Lab/adam_feltz.html

Abstract: Although it makes some progress, McKay & Dennett's (M&D's) proposal is limited because (1) the argument for adaptive misbelief is not new, (2) arguments overextend the evidence provided, and (3) the alleged sufficient conditions are not as prohibitive as suggested. We offer alternative perspectives and evidence, including individual differences research, indicating that adaptive misbeliefs are likely much more widespread than implied.

Evolutionary perspectives on adaptive misbelief are not new (Byrne & Kurland 2001; Haselton & Buss 2000; Trivers 1985; 2000; see also, Gigerenzer & Brighton 2009; Gigerenzer et al. 1999). What is new, however, is the precise analysis of the conditions of adaptive misbelief presented in the target article. Unfortunately, the target article's impact is limited by its reliance on controversial "better than average" effects and the relatively non-restrictive nature of the proposed sufficient conditions. Here, we briefly document these concerns and discuss some relevant phenomena in individual differences research. Ultimately, we argue that adaptive misbeliefs are likely much more widespread than is implied.

McKay & Dennett (M&D) suggest that adaptive misbeliefs are reflected in better-than-average and similar overconfidence type effects. However, there are concerns about the stability, universality, and reality of such illusions (Gigerenzer et al. 2008; Larrick et al. 2007; Moore & Healy 2008; see also, Juslin & Olsson 1997; Juslin et al. 2000). To illustrate, when most people report that they are better than average drivers they are not wrong or biased. Instead, data indicate that only a very small number of people are responsible for the vast majority of motor vehicle accidents. Thus, driving ability is not normally distributed and so most people are technically correct when they believe they are better than average drivers. This kind of example is not uncommon. Better-than-average and overconfidence type effects are often complicated by statistical artifacts and non-ecological task contexts (Gigerenzer et al. 1999; Krueger & Mueller 2002).

More problematic than the quality of the proposed evidence, however, are the following set of alleged sufficient conditions offered for systematic adaptive misbelief: (a) the belief is the result of "design" (where design is appropriately defined); (b) the belief misrepresents information to the possessor of the belief; (c) the misrepresentation of information is beneficial to the possessor of the belief (sect. 2, para. 5); and (d) the misbelief is systematic (sect. 4). If these conditions are only sufficient, then in contrast to what is implied, M&D have not captured a unique way in which misbelief can be adaptive. Rather, they have only pointed out one of many possible ways. This worry results in interpretative issues with M&D's general argument.

Assuming that many beliefs could satisfy (a), it is unclear what degree of misrepresentation or benefit is sufficient for a belief to satisfy conditions (b) and (c). According to M&D, a misbelief is one that is "false," or "to some degree departs from actuality," or "to some extent wide of the mark" (sect. 1, para. 1). These comments indicate that any belief that departs from reality in any way satisfies condition (b). The only way a belief could fail to satisfy (b) is if the content of the belief does not even in part misrepresent reality. If that is correct, then it is likely many (if not most) of our beliefs satisfy condition (b) (something M&D realize, sect. 1). It is also unclear how and in what ways the misbelief must be beneficial in order to satisfy condition (c). We can grant that positive illusions may be adaptively beneficial to the possessors of those beliefs in a number of profoundly interesting ways. But again, it is a very modest and easily satisfied condition if the misbelief only needs to provide *some* adaptive benefit to the possessor.

Condition (d) also is satisfiable in a number of ways. M&D appear to endorse a "one size fits all" model of misbelief that would be adaptive for whoever holds such misbeliefs (condition [d]). But there is more than one way that misbeliefs can be systematic. For instance, there can be misbeliefs that are systematically related to stable individual differences among groups of people. There is evidence that personality traits (e.g., the Big Five) are related to individual differences in beliefs about the nature of the world (Langston & Sykes 1997) and to fundamental philosophical beliefs regarding moral objectivism, compatibilism, and intentional action (Cokely & Feltz 2009a; 2009b; Feltz & Cokely 2008; 2009).

To take just one example, those who are neurotic are likely to think that the world is dangerous. Those who are not neurotic tend not to have this belief (particularly so for extraverts and those who are agreeable) (Langston & Sykes 1997, p. 154). On the face of it, these are contrary beliefs. So, either neurotic individuals have a misbelief or non-neurotic individuals have a misbelief – and perhaps both have misbeliefs. Evidence also indicates that some personality types are related to beneficial life outcomes and that personality traits are partially genetic in origin (Bouchard 1994). Hence, it appears that at least some systematic individual differences in beliefs are likely to be excellent candidates to satisfy (a)–(d).

Given that it is likely that quite a few of our beliefs satisfy (a)–(d), M&D underestimate the number of misbeliefs that are adaptive. Moreover, it may be that individual differences in misbeliefs are adaptive for both the specific misbelieving actor and for other non-misbelieving members of their group. That is, differences in belief might enable more effective allocation of limited resources in groups, benefiting both accurate and misbelievers alike (Wolf et al. 2007). In summary, we argue that although the proposed parameters offered by M&D do provide substantive increases to theoretical specification, they do not support bold claims such as "the exchange rate with truth is likely to be fair in most circumstances" (sect. 15, final para.). It is possible that adaptive misbeliefs are in the minority; however, this has yet to be adequately evaluated and does not follow from the evidence or argument provided. In contrast, we suspect that there are many relatively unexplored opportunities for theoretical and translational progress at these frontiers (e.g., the modeling of decisions and design of better choice environments; Johnson & Goldstein 2003; Todd & Gigerenzer 2007; Weber & Johnson 2009).

NOTE

1. Authorship of this commentary is equal.

Delusions and misbeliefs

doi:10.1017/S0140525X09991191

Max Coltheart

Macquarie Centre for Cognitive Science, Macquarie University, Sydney, NSW 2109, Australia.

max@maccs.mq.edu.au

www.maccs.mq.edu.au/~max

Abstract: Beliefs may be true or false, and grounded or ungrounded. McKay & Dennett (M&D) treat these properties of belief as independent. What, then, do they mean by *misbelief*? They state that misbeliefs are "simply false beliefs." So would they consider a very well-grounded belief that is false a misbelief? And why can't beliefs that are very poorly grounded be considered delusions, even when they are true?

Suppose a man goes to see his psychiatrist complaining of anxiety and depression and, when asked what was making him anxious and depressed, replies that it was because his wife was having an affair with a man in the office where she worked. Suppose there follows a discussion between clinician and patient about the patient's reasons for believing that his wife was being unfaithful to him. Suppose the reasons the patient offers include such things as "She wears a different dress every day, and she always puts on makeup very carefully each morning" and "Sometimes she phones me to say she has a deadline to meet at work and has to stay behind for an hour, and then she does come home an hour later than usual."

These don't seem to be very convincing reasons, and the other reasons the patient proffers are no more convincing, so the clinician begins to doubt the reasonableness of the patient's belief. Hence the clinician follows this up by asking the patient whether he has taken any steps to verify his belief. The patient says that he has; that on various occasions he has hidden outside his wife's place of work to see whether she ever leaves in the company of a man. Asked whether he had ever seen her do this, he says "No, she has always left by herself," but then volunteers the comment that he must have been unlucky in his choice of days; indeed, he mentions, he once performed this stakeout every day for a week, with negative results, which, he adds, must have meant that the male coworker concerned had been away from work that week. The patient also mentions that he has confided his worries about his wife's infidelity to his children, who pointed out to him that his reasons for the infidelity belief are flimsy in the extreme, and urged him to abandon the belief; but this has made no difference to the strength of his belief.

Given that this man cannot produce a single piece of evidence that plausibly supports his belief, and given that, even though the results of his investigations have been uniformly negative, this has not shaken him in the belief, does it not seem natural to regard this belief as a delusion? Similarly, might we not expect the clinician to conclude that this patient needs treatment? If we answer both of these questions in the affirmative, what is our reason for this? The answer is obvious: it's because this man has a belief that is held (a) with strong conviction regardless of the counterevidence and (b) despite the efforts of others to dissuade him.

Now suppose that, some years later, the clinician discovers that the man's belief was true after all: His wife *had* been having an affair at that time, and indeed it was with that particular male coworker. Does that mean that it had been a mistake to consider the patient's belief as a delusion? If the essence of the concept of delusional beliefs is that they are beliefs that are strongly and incorrigibly held in the absence of adequate grounds for doing so, then no mistake was made. It would have been very strange if at that time the clinician had mentally noted: "Before I decide whether this man needs treatment, I will have to find out whether or not his wife really is having an affair."

I consider that this example shows that, when one is classifying a particular belief as a delusion or not a delusion, whether the belief is *true* is irrelevant. What is relevant is whether the *grounds for the belief* are good enough. They weren't good enough in the case of our delusionally jealous patient (even though his belief, as it happened, was true).

What are the implications of this conception of delusion for the target article? First of all, the infidelity scenario is a specific example of a general possibility accepted by McKay & Dennett (M&D): that "ungrounded beliefs [can] be serendipitously true" (sect 1, para. 2; though the positive connotation of "serendipitously" is not quite right here; something like "accidentally" is needed). But, importantly, they note that they will not consider such ungrounded beliefs as misbeliefs. So our patient's belief about his wife's unfaithfulness does not count as a misbelief for M&D. If so, it isn't clear what they mean by "misbelief." The first sentence of their article says, "A misbelief is simply a false belief." Later they acknowledge that false beliefs can be well-grounded and true beliefs can be held with no grounds, so that truth and groundedness are independent. Which of these two is critical for characterizing their concept of misbelief? Is a belief that is very well-grounded but false a misbelief?

Later on in the introduction of their article the authors offer a tentative taxonomy of misbelief: "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" (sect. 1, para. 4). But in neither case does the method via which the belief is generated guarantee that the belief is false; it might be true – in which case it doesn't count as a misbelief.

What, then, is the difference between misbelief and false belief? If all misbeliefs are false beliefs, and if "misbelief" and "false belief" are not synonymous, then there must be false beliefs that are not misbeliefs. What criterion classifies false beliefs into those that are misbeliefs and those that are not?

At the beginning of section 4, "Doxastic dysfunction," the authors write: "delusions are misbeliefs *par excellence* – false beliefs that are held with strong conviction regardless of counterevidence and despite the efforts of others to dissuade the deluded individual" (sect. 4, para. 2). The example with which I began this commentary was of a belief held with strong conviction regardless of the counterevidence and despite the efforts of others to dissuade the individual holding this belief. Do M&D want to say that, in the scenario that I outlined, the patient's belief about his wife's fidelity doesn't count as a delusion – just because it happened to be true? The requirement always to establish the objective falsity of a belief before offering a diagnosis of delusion would wreak havoc in the profession of psychiatry.

Misbelief and the neglect of environmental context

doi:10.1017/S0140525X09991208

David Dunning

Department of Psychology, Cornell University, Ithaca, NY 14853.

dad6@cornell.edu

http://cornellpsych.org/sasi/index.php

Abstract: Focusing on the individual's internal cognitive architecture, McKay & Dennett (M&D) provide an incomplete analysis because they neglect the crucial role played by the external environment in producing misbeliefs and determining whether those misbeliefs are adaptive. In some environments, positive illusions are not adaptive. Further, misbeliefs often arise because the environment commonly fails to provide crucial information needed to form accurate judgments.

The thoughtful and stimulating analysis provided by McKay & Dennett (M&D) on human misbelief is incomplete. Assuming that misbeliefs are products of faulty design features internal to the human organism, M&D have unduly ignored the important

role played by the external environment in shaping human action and outcome. This neglect of environmental context holds two crucial implications for their analysis.

First, M&D conclude that positive illusions are adaptive, and thus, the best candidates to be misbeliefs engineered by evolution. In particular, they point to the established, albeit contentious, literature suggesting that positive illusions aid people in their resilience against some of the most extreme challenges of life, such as terminal illness or the aftermath of civil war (cf. Armor & Taylor 1998).

Those studies, however, exist in only one constrained environmental context. Elsewhere, the literature is filled with numerous counterexamples, strewn across business, education, and policy worlds, in which positive illusions prove costly or even disastrous (for reviews, see Dunning 2005; Dunning et al. 2004). Some of them even come from the health domain. People treat their own high blood pressure based on mistaken ideas about their competence to do so, setting aside their doctor's orders (Meyer et al. 1985). They smoke, at least in part, because of mistaken beliefs about their ability to avoid serious illness (Dillard et al. 2006). Teenage girls who rate their knowledge about birth control highly, independent of actual knowledge, are more likely to get pregnant within a year relative to their less self-flattering peers (Jaccard et al. 2005). Just as a thought experiment, it is easy to come up with numerous contexts in which positive illusions might be the opposite of adaptive. Answer yes or no to the following thought question: When flying, I prefer my pilot to have an overconfident view of his or her ability to handle rough weather.

The point here is that the extant evidence connecting positive illusions to adaptive outcomes is mixed, at best, and depends crucially on the specific environmental context under study. This is not to dismiss those important areas where a positive connection exists; but much more work is necessary to see just how the environmental context, systematically, turns on and off the connection between positive illusion and adaptive outcomes. It may turn out that positive illusions, in the end, bring more sorrow than pleasure. At least it is worthwhile discerning more precisely the circumstances in which that is so.

Such an analysis of environmental context is crucial also to assess M&D's tentative assertion that positive illusions are specifically a product of human evolution. If they are, then they should be more consistently evident in tasks with evolutionary significance (e.g., getting a full belly, achieving reproductive success) than those without. But, to date, that careful analysis across environmental contexts has not been done.

Second, M&D take misbeliefs to be direct evidence of faulty design features in the human organism. That may be the case, but there is an equally compelling case emerging in the psychological literature that it is the environment, not human flaw, that makes these biases unavoidable. Even a perfectly rational human organism could come to hold the types of misbeliefs that M&D discuss, because the environment much more frequently provides people with incomplete or misleading data than M&D anticipate.

In my own work, I have discussed how people might come to hold overly inflated self-views because the environment fails to furnish all the data they need to form accurate self-impressions. In the course of their lives, for example, people decide on actions that they believe are the most reasonable among the choices available. However, when they choose unwisely, they do so because the environment fails to provide the data that would inform them of just how ill-advised their choices are (Dunning 2005). Give them that data and they snap quickly to a more accurate view of themselves (Caputo & Dunning 2005; Kruger & Dunning 1999).

As another example, take the observation that people tend to view others with suspicion, anticipating much more harm from others than actually is the case (e.g., Duntley & Buss 1998). M&D speculate that this bias evolved because it protected people from injury, whether physical or psychic. Recent work, however, suggests that the real potential culprit producing this bias is the environment, not a design feature of the human organism.

For example, people tend to be overly cynical about how trustworthy other people are (Fetchenhauer & Dunning 2009). We have recently demonstrated that this cynicism is produced by environmental factors, in that the environment furnishes people with incomplete feedback about their decisions to trust others. When people trust others, their trust is occasionally violated, and people quite rationally move toward a more cynical view of human nature. However, when they mistakenly decide to withhold trust from a person who actually would have honored that trust, they receive no equivalent corrective feedback. Thus, they are left with a unduly wary view of the other individual and of humanity in general. We have shown how furnishing people with complete feedback, including letting people know when their withheld trust would have been honored, rids them quickly of their cynical misbeliefs – and leads them to make trust decisions that provide greater tangible benefits (Fetchenhauer & Dunning, in press; see also Denrell 2005; Smith & Collins 2009).

In sum, M&D have taken first intelligent and careful steps toward an evolutionary treatment of human misbelief, but they need to consider the crucial role played by environmental context before their evolutionary analysis potentially veers into misbelief itself. Social psychologists often chide laypeople in their everyday lives for neglecting the impact of environmental forces on human behavior and outcomes (Nisbett & Ross 1980). Thus, as theorists, we should commit to giving environmental forces their due in our own thinking about the human condition.

Why we don't need built-in misbeliefs

doi:10.1017/S0140525X09991427

Carol S. Dweck

Department of Psychology, Stanford University, Stanford, CA 94305.
dweck@stanford.edu

Abstract: In this commentary, I question the idea that positive illusions are evolved misbeliefs on the grounds that positive illusions are often maladaptive, are not universal, and may be by-products of existing mechanisms. Further, because different beliefs are adaptive in different situations and cultures, it makes sense to build in a readiness to form beliefs rather than the beliefs themselves.

McKay & Dennett (M&D), in their fascinating and thought-provoking article, conclude that positive illusions meet the criteria for evolved misbeliefs. I propose that the case is not made for the status of positive illusions or indeed for the idea of evolved misbeliefs.

The target article suggests that positive illusions are clearly adaptive, are universal, and are not by-products of other beliefs. Each of these suggestions can be questioned. First, a closer look at the psychological literature shows the pitfalls of positive illusions – how an inability to see their own weaknesses can prevent people from reaching important goals and can endanger their health and safety (Dunning et al. 2004). Looking back, one can easily see how hunters who overestimated their abilities vis-à-vis predators might not have survived to reproduce; people who had overoptimistic views about food for the winter might have starved; and parents who overestimated their children's skills might have put them in jeopardy.

Second, a closer look at the literature in cultural psychology casts doubt on the universality of positive illusions. Positive illusions are found to be a feature of Western societies, which focus on individuals and their personal prowess, but these illusions are absent or considerably weaker in Eastern cultures that focus on self-criticism, self-improvement, and adjusting to others (Heine & Lehman 1995; Kitayama et al. 1997).

Third, the authors argue that religion is not a candidate for an evolved misbelief in part because religious beliefs are by-products of other more fundamental mechanisms, such as heightened perceptions of agency. However, they do not seem to hold positive illusions to the same rigorous standard. Are not positive illusions the very embodiment of a heightened sense of one's own agency? Thus, the case that positive illusions are singular candidates for evolved misbeliefs is still open.

However, an even more fundamental suggestion in the target article is that shared, adaptive misbeliefs need to be built in. Is this so? The psychological literature is replete with evidence for innate or very early core knowledge (e.g., knowledge about objects and number; Spelke & Kinzler 2007), as well as for such things as (a) attentional biases (e.g., to human voices: Vouloumanos & Werker 2007; to top-heavy forms like faces: Cassia et al. 2004; to negative affect: Vaish et al. 2008), (b) sensitivity to contingencies and transitional probabilities (Saffran et al. 1996; Watson 1985; see also, Johnson et al. 2007), and (c) considerable inferential capabilities (Woodward & Needham 2009).

These very basic infant attributes – core knowledge, attentional biases, sensitivities to the statistical properties of input, inferential abilities – all set infants up to learn about their worlds. Now, it makes sense to build in knowledge about things like object and number that are invariant across centuries and cultures, but, after that, it makes sense to equip infants with the apparatus to learn from their input. Indeed, it may be imperative for them to remain open to different misbeliefs, since particular misbeliefs may vary greatly in their adaptiveness across situations and cultures. In this way, Western babies can develop positive illusions, but Eastern babies can develop more self-critical and cautious stances.

In my decades of research, I have been struck by one thing more than any other: the rapidity with which children and adults alike key into the rules, beliefs, and values in a new environment. In a series of studies (Kamins & Dweck 1999; Mueller & Dweck 1998; see also Cimpian et al. 2007), we have shown how children are affected in dramatically different ways when, after a successful performance, they are praised *once* for their intelligence as opposed to their effort. After praise for intelligence, they adopt a belief in fixed intelligence and act in accordance with it. For example, they choose to work on tasks that will validate their intelligence and, after a failure, will make negative inferences about their intelligence, resulting in impaired performance. After praise for effort, children adopt a belief that ability can be increased through effort and act in accordance with that belief. They choose to work on challenging tasks that will increase their ability, and after a failure, will continue to apply effort, resulting in increased performance. We have repeated this study or variants of it eight times, with the same results.

Other recent research shows how readily people can adopt prevailing beliefs, without worrying whether they are true or false (Murphy & Dweck 2009). In one study, we had people read minutes from a meeting of an organization, with the idea that they would later apply to work at that organization. The minutes implied that members of the organization believed either that intelligence was fixed or that intelligence could be developed. Before people applied to the organization, however, they went to a different room with a different experimenter to engage in a completely different task. Here, they completed a self-concept questionnaire that listed personal characteristics and asked them to rate "how much each characteristic is at the core of who you are." What happened was striking. People who had simply read about the fixed-intelligence organization said that being brilliant was more central to who they were, but those who had read about the malleable-intelligence organization said that being passionate [about learning] was more central to who they were. They had internalized the beliefs and values of the organization and this was true even when people did not like the organization they had read about!

In fact, the science of social psychology can be seen as the science of how small changes in situations can lead to large changes in beliefs and behavior (Ross & Nisbett 1991). Humans are social animals. We need to feel out and respond flexibly to new situations and this includes inferring or absorbing the (mis)beliefs that go with the new situations. If anything, it is our readiness to adopt prevalent beliefs or misbeliefs that is built into us, rather than the beliefs or misbeliefs themselves.

"Can do" attitudes: Some positive illusions are not misbeliefs

doi:10.1017/S0140525X09991439

Owen Flanagan

Department of Philosophy, Duke University, Durham, NC 27708-0743.
ojf@duke.edu http://www.duke.edu/∼ ojf

Abstract: McKay & Dennett (M&D) argue that positive illusions are a plausible candidate for a class of evolutionarily "selected for" misbeliefs. I argue (Flanagan 1991; 2007) that the class of alleged positive illusions is a hodge-podge, and that some of its members are best understood as positive attitudes, hopes, and the like, not as beliefs at all.

Since positive illusions are the one example McKay & Dennett (M&D) find of a bona fide contender for an adaptive evolutionary favored epistemic disability, my view (Flanagan 1991; 2007) that positive illusions may not be a class of well-behaved misbeliefs at all should matter.

Start with this conditional: *If* there are positive illusions and *if* they are, as the psychologists say they are, (a) common, plus correlated with (b) moral decency and (c) happiness, positive affect, optimism, as well as with (d) the capacity to engage in profitable, creative, productive work, then there is a problem with the view that flourishing requires, or demands, overcoming the tendency to harbor false beliefs. The reason is simple: The normative claim that contemporary people flourish truly only if they live in the light of the true, is in competition with the psychologist's claim that the capacities to love, work, and be happy are enhanced by false belief.

The conditional that causes the latter problem – competition between the ends of flourishing and fitness – and that also, albeit independently, warrants M&D's ingenious explanation for why there are positive illusions, involves accepting that there are positive illusions. But we can challenge sensibly the antecedent of the conditional.

Accept that "positive illusions" are states of mind that benefit the consumer, but reject the claim that they (all, most, many) are best interpreted as involving false beliefs, as opposed to having positive expectations and hopes – in other words, a positive attitude. Hopes and a "can do" attitude need not require false belief. "Exaggerated" and "unrealistic" are adjectives used to describe the whole set of allegedly questionable epistemic states of mind in Taylor and Brown's famous meta-analysis (Taylor & Brown 1988), which include unrealistic positive evaluations, exaggerated perceptions of control and mastery, and unrealistic optimism. One ought to worry about inferring false beliefs from (even) correct ascriptions of lack of realism or exaggerated views about one's powers and abilities.

When Muhammad Ali famously remarked before his final fight with Smokin' Joe Frazier – the "Thrilla in Manila" – that "It will be a killa . . . and a chilla . . . and a thrilla . . . when I get the gorilla in Manila," did he believe that he would kick Smokin' Joe's ass? Or is he best understood as doing something, performing an action that was, in effect, part of the fight before the first bell sounded? Both boxers presumably believed that they could win and hoped that they would win. So far there is no epistemic mistake regardless of outcome. "Can" does not entail "will." The epistemic standards

governing hopes, desires, and the like, are different from those that govern beliefs. It would be very odd to say that the losers in zero-sum games always have false beliefs. In fact, Ali might have believed that he could not win unless he made Smokin' Joe worry that he might know how to beat him. Furthermore, Smokin' Joe need not have believed he would lose after Ali's provocation. The effect might work this way: Ali knows how to do things with words. He speaks with the intention of undermining Smokin' Joe's confidence, and does so. In this case, the mechanisms at work do not operate via beliefs at all, although they might commonly be assimilated to that class of states specified as *folk* psychologically. The point is that many things we do with words and thoughts can be viewed as strategic – engendering self-confidence, undermining the competitor's abilities – and not as straightforwardly epistemic.

Consider this from Aristotle:

> We ought not to follow the proverb-writers, and "think human, since you are human." Or "think mortal, since you are mortal." Rather, as far as we can go, we ought to be pro-immortal, and go to all lengths to live a life that expresses our supreme element; for however this element may lack in bulk, by much more it surpasses everything in power and value. (Aristotle 1985, *Nichomachean Ethics*, X: 13.37)

Interpreted one way Aristotle can be read as encouraging two false beliefs; interpreted another way he can be read as encouraging an attitude that one can achieve something excellent if one sets one's eyes on the goal. Coaches often speak this way to their charges. A professional tennis match always produces one winner and one loser. Both players, if they are any good, go into the match believing that they can win, indeed that they will win. Believing one can win is a true belief. Hoping that one will win is a sensible expectation. In neither case is there a mistake.

A sensible counterfactual test for whether a person in fact holds a belief or is in some associated epistemic state in a strong and objectionable way would be: Does the state-in-question yield, and if so how quickly, easily, and so on, when there is strong countervailing evidence? If I get prostate cancer, or divorced, or in a motorcycle accident despite saying that I think I won't, I will quickly yield my initial thought or claim that these calamities will not befall me. Taylor and Brown (1988, p. 197) write that "the extreme optimism individuals display [about such probabilities] appears to be illusory." This is not obvious. Optimism can be unrealistic, perhaps – illusory is a different matter.

The overall point is that the positive illusion literature conflates and assimilates systematically such states as hopes, expectations, and positive attitudes with states of false belief, when the charitable analysis need not involve attributing any belief at all, let alone a false one. One final point: There is reason to believe that one class of alleged "positive illusions," self-serving ones, is not common outside of the West (Flanagan 1991; 2007; Heine et al. 1999). If so, this might cause trouble for M&D since there is no common phenotypic trait to explain it as an adaptation.

Adaptive misbelief or judicious pragmatic acceptance?

doi:10.1017/S0140525X0999121X

Keith Frankish

Department of Philosophy, The Open University, Milton Keynes, Buckinghamshire MK7 6AA, United Kingdom.

k.frankish@open.ac.uk

http://www.open.ac.uk/Arts/philos/frankish.htm

Abstract: This commentary highlights the distinction between belief and pragmatic acceptance, and asks whether the positive illusions discussed in section 13 of the target article may be judicious pragmatic acceptances rather than adaptive misbeliefs. I discuss the characteristics of pragmatic acceptance and make suggestions about how to determine whether positive illusions are attitudes of this type.

McKay & Dennett (M&D) ask if there are adaptive misbeliefs. The positive illusions they discuss in section 13 of the target article are plausible candidates for propositional attitudes that are adaptive irrespective of their truth, but I want to question whether they are really beliefs. The authors adopt a broad definition of belief as a functional state that "implements or embodies" (sect. 1, para. 1) the endorsement of a state of affairs as actual. However, this glosses over a distinction often made between *belief* and *acceptance* (e.g., Bratman 1992; Cohen 1989; 1992; Engel 1998; 2000; Frankish 2004; Stalnaker 1984), and it may be that some putative adaptive misbeliefs are better classified as judicious pragmatic acceptances.

The belief/acceptance distinction is drawn in slightly different ways by different writers, but a central claim is that belief is involuntary and acceptance voluntary. To accept a proposition is to adopt a policy of treating it as true (taking it as a premise) for the purposes of reasoning and decision making. Now acceptance can be motivated by epistemic reasons, and when it is it can be regarded as a form of belief. (I would argue that all-or-nothing belief, mentioned by the authors in note 2, is a truth-directed form of acceptance; see Frankish 2004; 2009.) However, we can also accept things for non-epistemic reasons – ethical, professional, religious, and so on. For example, loyalty may require a person to accept that their friend is telling the truth, and professional ethics may oblige a lawyer to accept that their client is innocent, even if they do not believe these things (Cohen 1992). Acceptance can also be prudential, designed to simplify complex deliberations or handle error-management considerations of the sort discussed in section 9 of the target article (Bratman 1992). (M&D suggest that such considerations need not motivate belief, but only a cautious action policy. The present suggestion, however, is that they may prompt the formation of a *deliberative* policy, which constitutes a type of propositional attitude.) I shall refer to acceptance that is motivated by non-epistemic concerns as *pragmatic acceptance*.

With the notion of pragmatic acceptance in place, what should we say about the unrealistically positive self-appraisals identified by Taylor and her colleagues (e.g., Taylor 1989; Taylor & Brown 1988)? Are these genuine misbeliefs or pragmatic acceptances, motivated perhaps by a sense of their therapeutic value or a desire to maintain a comforting self-image? The distinction between belief and acceptance is often overlooked, so it is not enough to note that these attitudes are typically classified as beliefs. Nor would it be sufficient to detect the influence of pragmatic motives in their formation, since these may be operative in both cases – illicitly in one case, legitimately in the other. In short, how can we tell the difference between beneficial misbelief and judicious pragmatic acceptance?

One way is by considering subjects' attitudes to their self-appraisals. In particular, do they feel they have control over these judgements and do they think it is legitimate to allow non-epistemic factors to influence them? If so, this would suggest that their attitude is one of pragmatic acceptance rather than belief. There is some evidence that this is the case. Everyday wisdom says it is beneficial to adopt a positive outlook – to think positively, be optimistic, and have confidence in oneself – and we often strive to take this advice to heart. Moreover, we do so without feeling that we are thereby violating epistemic norms, even if we have no evidence to support the views adopted. This is not decisive, however. Our control here may be only indirect, and some self-deception may be involved.

A second consideration is the deliberative context in which our positive illusions are active. Pragmatic acceptance, unlike belief, is context-dependent. More specifically, our beliefs (including truth-directed acceptances) guide us in an open-ended range of deliberations, including, crucially, ones where we want to be guided only by the truth – *truth-critical* deliberations (Frankish 2004). Our beliefs are our best bets at truth, and they are what we

rely on when we want to rely on the truth. Pragmatic acceptances, on the other hand, are operative only in contexts where non-epistemic values, such as loyalty or professional ethics, matter to us more than truth, or where we prefer to err on the side of caution. (Note that this means that pragmatic acceptance requires the ability to classify deliberations as truth-critical or not, and hence requires metacognitive abilities.)

The question, then, is whether we rely on our optimistic self-appraisals in truth-critical contexts. Is there evidence for this? It might be replied that people report their self-appraisals with apparent sincerity, and that sincere reports are the product of truth-critical deliberation. This is too swift, however. If positive self-appraisals have considerable therapeutic value, then that would be a reason for us to treat deliberations about ourselves as *not* truth-critical, unless there is a lot at stake. Moreover, this may go for what we tell ourselves as much as for what we tell others; there need not be any conscious insincerity involved.

Experiment should help here. For example, we might ask subjects to form assessments of their own abilities and attributes, offering varying rewards for accuracy. (It would not matter if accuracy could not easily be determined, provided subjects thought it could.) If a subject revised or abandoned an assessment as the rewards – and thus the truth-criticality of the context – increased, this would suggest it was an object of pragmatic acceptance rather than belief.[1] Some pragmatic acceptances, however, will be hard to detect. Truth-criticality is determined by the subject's priorities; and, in general, the stronger a person's pragmatic reasons for accepting a certain claim, the harder it will be to create conditions under which they will treat deliberations involving it as truth-critical. Indeed, at the extreme, they may treat none as such, rendering their attitude functionally equivalent to belief.

Despite these practical difficulties in applying the distinction between belief and pragmatic acceptance, it is important to keep the distinction in mind when theorizing about adaptive misrepresentations. For one thing, it suggests there are distinct routes to the formation of such attitudes – one involving the overriding of barriers to the influence of motivational processes on belief formation (the breaking of what M&D call "doxastic shear pins"), the other involving mechanisms of pragmatic acceptance in which such barriers are not present. There may also be differences in the psychological and physiological effects of optimistic self-appraisals depending on whether they are pragmatically accepted or genuinely believed. This again may be a matter for experiment.

NOTE

1. Thanks to Ryan McKay for this suggestion.

On the adaptive advantage of always being right (even when one is not)

doi:10.1017/S0140525X09991221

Nathalia L. Gjersoe and Bruce M. Hood

Department of Experimental Psychology, University of Bristol, Bristol BS8 1TU, United Kingdom.

N.L.Gjersoe@bristol.ac.uk **B.M.Hood@bristol.ac.uk**

http://psychology.psy.bris.ac.uk/people/nathaliagjersoe.html

http://brucemhood.wordpress.com

Abstract: We propose another positive illusion – overconfidence in the generalisability of one's theory – that fits with McKay & Dennett's (M&D's) criteria for adaptive misbeliefs. This illusion is pervasive in adult reasoning but we focus on its prevalence in children's developing theories. It is a strongly held conviction arising from normal functioning of the doxastic system that confers adaptive advantage on the individual.

McKay & Dennett (M&D) address a wide variety of misbeliefs and whittle the range down to conclude that one type of misbelief – positive illusions – remains as viable evidence for the existence of adaptive misbeliefs. We concur with M&D's line of reasoning and propose an additional form of positive illusion that may serve as an example of an adaptive misbelief – overconfidence in the veracity and generalisability of one's theories. This positive illusion is common across a range of domains in adult reasoning (e.g., Klahr & Dunbar 1988; Rozinblit & Keil 2002), but we will focus on its prevalence in theory formation in the developing mind as an incidence that is naturally occurring, pervasive, and seemingly robust. M&D are clear about the qualities that a possible candidate for adaptive misbelief must have: It must be a belief, it must arise in the normal course of the doxastic system's proper functioning and, most important, it must confer adaptive advantage on the individual. We address each of these points in turn.

Karmiloff-Smith (1992), in her outline of the representational redescription model, suggests that all children go through three phases in theory formation. Briefly, in Phase 1 children collect data from the world, treating each experience as an independent event with little or no generalisation between occurrences. In Phase 2 they consolidate independent representations into a unified theory, rejecting contrary external evidence while the theory is strengthened. And in Phase 3 they test the theory on a range of external examples, adjusting and broadening it to account for a variety of anomalies. The theoretical entrenchment exhibited in Phase 2 can result in errors and inflexibilities not evident in Phases 1 and 3 that lead to the characteristic U-shaped curve of behavioural success on a variety of tasks (e.g., Karmiloff-Smith 1986; Newport 1981). The classic demonstration of this developmental pattern is Karmiloff-Smith and Inhelder's (1974) block-balancing task. When asked to balance a series of blocks, some of which had been covertly weighted such that their balancing point was off-centre, 4- and 8-year-olds consistently passed while 6-year-olds consistently failed. Karmiloff-Smith suggests that the 4-year-olds succeed by treating each block as an independent task but the 6-year-olds have a theory that blocks balance in their symmetric centre and they overgeneralise this to apply to all blocks, even when this strategy consistently fails. Eight-year-olds, by contrast, hold the same theory but are flexible enough to also take into account the extra dimension, asymmetric weight, and adapt their strategy.

To what extent can children's experience of theoretical entrenchment in Phase 2 be referred to as a "belief"? M&D dismiss aliefs and judicious psychological biases as candidates for adaptive misbelief because, they argue, the consumers don't really *believe* the bias, they just respond *as though* they were in danger or *in case* danger might be lurking. Conversely, in many ways children's overconfident belief in their theory is akin to M&D's description of delusions as examples of misbeliefs *par excellence* in that they are "held with strong conviction regardless of counterevidence and despite the efforts of others to dissuade the deluded individual" (sect. 4, para. 2). To test whether 3-year-olds had a perseverative theory that all objects fall straight down, Hood (1995) presented an array in which objects fell down a curved tube to a displaced location. All 3-year-olds searched directly below the dropping point even if there was only one tube (and thus no physical connection between the dropping point and the favoured target), repeatedly in the face of counter-evidence (up to 20 consecutive trials) and persistently regardless of how many times the experimenter explained the role of the tubes to them. This persistence implies that children really *believe* that their theory is correct. M&D dismiss delusions as examples of adaptive misbelief because they arise from an improperly functioning doxastic system. Conversely, Phase 2 theoretical perseveration occurs in all children across a range of domains and seems to be a built-in feature of a properly functioning theory-formation mechanism. Indeed, children often make up observables in support of their theory when the perceptual

experience lets them down (e.g., Baker et al. 2009; Massey & Gelman 1988).

Last, it is necessary that a proposed adaptive misbelief should convey adaptive advantage to the individual. Overconfidence in one's theories conveys adaptive advantage insofar as it enables them to creatively simplify a problem by ignoring some of the complicating factors. "[I]t seems possible for the child to experience surprise and question his theory only if the prediction he makes emanates from an already powerful theory expressed in action" (Karmiloff-Smith & Inhelder 1974, p. 209). Thus, Phase 2 enables children to unify representations into coherent (but overgeneralised) theories that in turn lead to new, broader theories and greater behavioural mastery. Second, overconfidence in one's theories sustains and enhances health in an everyday sense by decreasing exposure to cognitive dissonance, which has been shown to lead to feelings of anxiety and stress (Aronson 1969), which in turn result in negative physiological effects. Consequently, overconfidence in one's theories may also result in exaggerated feeling of control, a positive illusion that M&D list as adaptive in its own right.

Thus, overconfidence in the veracity and generalisability of one's theory fits the criteria laid out by M&D as necessary to be considered as an adaptive misbelief. Children certainly *believe* that they are right; this belief is systematic and misinforms the organism as a whole, occurs for all children across a range of microdomains, and persists into adulthood. Therefore, it can be considered a naturally occurring feature of a properly functioning doxastic system. It can also be construed as adaptive in leading the individual to undertake adaptive actions and by enhancing health and fitness. In children, this tendency is evident not only in subjective self-evaluation, but also in objective theories about how the world works that, in turn, guide their behaviour. A phase in which this is especially prominent occurs across a variety of microdomains and may be a fundamental and important feature of properly functioning theory-building doxastic systems.

Error management theory and the evolution of misbeliefs

doi:10.1017/S0140525X09991440

Martie G. Haselton[a] and David M. Buss[b]

[a]*Departments of Communication and Psychology, University of California, Los Angeles, Los Angeles, CA 90095;* [b]*Department of Psychology, University of Texas, Austin, TX 78712.*

haselton@ucla.edu **dbuss@psy.utexas.edu**

http://www.sscnet.ucla.edu/comm/haselton/home.html

http://www.davidbuss.com

Abstract: We argue that many evolved biases produced through selective forces described by error management theory are likely to entail misbeliefs. We illustrate our argument with the male sexual overperception bias. A misbelief could create motivational impetus for courtship, overcome the inhibiting effects of anxiety about rejection, and in some cases transform an initially sexually uninterested woman into an interested one.

McKay & Dennett (M&D) provide a useful analysis of the evolution of misbelief, making a number of important distinctions, including one between misbeliefs that are tolerable byproducts of evolved psychological adaptations and those that would have been adaptive in and of themselves. A reasonable primary hypothesis is that selection has shaped the human mind to form true beliefs about the world. The ultimate criterion of evolutionary selection, as M&D rightly point out, is reproductive success, not the accurate detection or preservation of truth. We, and others, have argued that selection has favored psychological adaptations that do not always maximize truthful beliefs; these adaptations instead can result in *misbeliefs* (e.g., Haselton & Buss 2000; Haselton & Nettle 2006).

Humans appear to possess cognitive biases which lead to systematic misbeliefs and require scientific explanation. These include the positive illusions that compel us to have a rosy outlook on the future (Taylor & Brown 1988), sex-linked biases such as men's tendency to overestimate women's sexual interest (e.g., Abbey 1982), and perceptual biases such as auditory looming, the tendency to overestimate the proximity to self of approaching objects compared to receding objects that are in fact equally distant (Neuhoff 2001). We articulated error management theory (EMT; Haselton & Buss 2000) as a theory to explain how evolution could lead to adaptive biases, some of which entail misbeliefs. Many problems of judgment under conditions of uncertainty can be framed as having two possible errors – false positive and false negative errors. According to EMT, in forming judgments under uncertainty, if there were recurrent asymmetries in the costs of these errors over evolutionary history, selection should produce a system that errs in the less costly direction. For example, for men estimating a woman's sexual interest, we hypothesized that the reproductively more costly error would have been to underestimate her interest and miss a reproductive opportunity. Thus, EMT predicts that men possess an adaptive bias toward overestimating women's sexual interest.

M&D affirm the logic of EMT, but argue that selection can solve adaptive problems of the sort explained by EMT in ways other than creating misbeliefs. They argue that humans do not need to possess *biased beliefs* if *biased actions* can accomplish the same ends while preserving true beliefs. We agree entirely with this point. It is possible, for example, that selection could design an adaptation in which men *acted* as if a larger number of women were sexually interested in them than actually were, in order for them not to miss a potential sexual opportunity, while not truly believing that those women are sexually interested. Similarly, it might be possible for selection to fashion an adaptation in which people act as though more people harbor homicidal intent than they actually do, in order to avoid the costly cases in which people actually do harbor such thoughts, without actually believing that those individual do harbor homicidal intent.

Just because selection *can* solve these adaptive problems without misbelief does not mean that selection *has* solved these problems without misbelief. The argument that selection could craft an adaptation for thermoregulation other than sweat glands (e.g., dogs thermoregulate through evaporation from a protruding tongue) is not an argument that selection has not fashioned sweat glands in humans.

Ultimately, the question of whether misbeliefs are part of the design of EMT biases is an open issue that must be decided on a case-by-case basis with empirical research. However, we suggest that there are no compelling reasons to discount the possibility that misbeliefs, including functional misbeliefs, are part of the evolved design of EMT biases. Consider the male sexual overperception bias. A misbelief that a woman is sexually interested could facilitate access to sexual opportunities in at least three ways. First, it could provide the motivational impetus for courtship behavior. Second, it could allay a man's anxiety about being rejected, eliminating a common cognitive barrier to initiating courtship (Kugeares 2002). If it turns out that his belief was indeed incorrect, it is not terribly costly for him to revise his beliefs about a particular woman after being rebuffed (e.g., "I thought she was sending me sexual signals, but it turns out I was wrong"). Third, a man's misbelief, by motivating attraction tactics or elevating confidence, could transform a woman who is initially sexually uninterested in him into one who is sexually interested – an outcome showing that the initial misbelief itself can sometimes provide functional benefits. Hence, the EMT-generated misbelief can, in principle, solve the adaptive

problem of maximizing sexual opportunities more effectively than an adaptation lacking the misbelief design feature.

Although we advanced the theory to explain cognitive biases, the core logic of EMT is neutral in predicting where in the perception-belief-action chain selection will shape a bias. All that is required is that, ultimately, humans behave so that they minimize the more costly of the two errors in question, even if this cost minimization ends up producing a larger number of overall errors. To discover where in this chain a bias exists must be empirically adjudicated. On the basis of the existing empirical evidence, however, we suggest that biasing action is unlikely to be the sole outcome of selection in which there has been recurrent cost asymmetries associated with errors.

M&D's analysis will stimulate empirical research about particular EMT biases. Some biases may be instances of biased action without involving misbelief. Others may entail misbeliefs. A subset of these may be cases in which the misbelief is not simply a tolerable byproduct of an adaptively biased cognitive system but is itself adaptive. M&D make a compelling argument that positive illusions qualify as adaptive misbeliefs because they positively affect an individual's fitness by motivating striving for favorable outcomes. We suggest that that some EMT biases, such as the male sexual overperception bias, also can motivate adaptive action through misbeliefs by providing motivational impetus for action, overcoming inhibitions associated with action, and transforming the psychological states of others in ways beneficial to the holder of misbeliefs.

God would be a costly accident: Supernatural beliefs as adaptive

doi:10.1017/S0140525X09991245

Dominic D. P. Johnson

Department of Politics and International Relations, University of Edinburgh, Edinburgh, EH8 9LD Scotland, United Kingdom.
dominic.johnson@ed.ac.uk http://dominicdpjohnson.com/

Abstract: I take up the challenge of why *false* beliefs are better than "cautious action *policies*" (target article, sect. 9) in navigating adaptive problems with asymmetric errors. I then suggest that there are *interactions* between supernatural beliefs, self-deception, and positive illusions, rendering elements of all such misbeliefs adaptive. Finally, I argue that supernatural beliefs cannot be rejected as adaptive simply because recent experiments are inconclusive. The great costs of religion betray its even greater adaptive benefits – we just have not yet nailed down exactly what they are.

The greatest challenge to McKay & Dennett's (M&D's) argument is why *false* beliefs are necessary to achieve adaptive behavior – why not (as M&D note in sect. 9, para. 2) just have "cautious action *policies*" instead? I don't believe this problem was completely resolved in the target article, so I tackle it with reference to the "supernatural punishment hypothesis" (Johnson 2009; Johnson & Bering 2006; Johnson & Krüger 2004), since the same problem haunts that hypothesis as well.

The argument is that the costs of selfishness increased when humans evolved language and Theory of Mind (ToM), because social transgressions became much more likely to be detected and punished. Supernatural punishment offered a cautionary mind-guard to reduce selfishness and avoid real-world costs. But why bring God into it? A Darwinian perspective suggests that atheists could simply develop a "cautious action policy" – becoming more prudent about when to be selfish. A first line of defense comes from M&D's categories of evolutionary limitations: (1) economics – a fear of supernatural agency may have been biologically cheaper or more efficient; (2) history – a capacity for supernatural beliefs may have been more readily available, given the prior evolution of ToM; (3) adaptive landscape – fear of detection and punishment by *supernatural* agents may have been a small step up the local fitness peak from fear of detection and punishment by *human* agents.

A stronger line of defense is that, while a cautious action policy might work in principle, the whole point of error management theory is that it pays to *overestimate* the probability of detection, not to get it right or to weigh up the costs and benefits "rationally" (Haselton & Buss 2000; Haselton & Nettle 2006; Nettle 2004). Believing (irrationally) that *supernatural* agents are watching is a good way to ensure systematic overestimation of the *actual* risk of detection and punishment (by other human beings; Johnson 2009). The power of religion appears to stem precisely from its irrational and non-falsifiable features (Rappaport 1999), and empirical data suggest that religious beliefs are more effective at promoting group survival than similar but *non*-religious beliefs (Sosis & Bressler 2003). Cautious action policies might work in reducing selfishness, but they may not be as effective as God.

My next concern is that supernatural agency, self-deception, and positive illusions are treated as independent phenomena, with only positive illusions making the cut for an adaptive misbelief. However, there are important *interactions* between these three phenomena that make elements of all of them adaptive.

First, self-deception is essential to many supernatural beliefs. If supernatural punishment is to affect people's behavior, they must believe in it – despite lacking any direct evidence whatsoever and despite having to ignore counter-evidence. This is classic self-deception (Trivers 2000). Interestingly, this self-deception can be reinforced by the belief itself – in many religions, it is common for someone's misfortune to be treated as *evidence* of wrongdoing, since gods or spirits "evidently" punished the victim (Bering & Johnson 2005).

Second, self-deception is essential to many positive illusions. For example, positive illusions have been suggested to be adaptive in conflict, bluffing superior power or skill to deter opponents (Johnson 2004; Trivers 2000; Wrangham 1999). Self-deception is essential here to avoid "behavioural leakage" that would otherwise give the game away (nervous Nellies are less convincing bluffers than cool-hand Lukes). This may be why, as Daniel Kahneman notes, "all the biases in judgment that have been identified in the last 15 years tend to bias decision-making toward the hawkish side" (quoted in Shea 2004). Positive illusions appear to be advantageous enough that numerous psychological biases converge to promote them despite the evidence.

Third, supernatural beliefs may be an *example* of positive illusions. As M&D note, people often cite God as giving them "the strength to go on." If health or fitness advantages derive from such beliefs, then religious beliefs are adaptive according to M&D's own criteria. Religious beliefs may involve all three types of positive illusions: positive self-evaluations (God chose me/us), illusions of control (God will help me/us in difficult times), and optimism about the future (God has a plan; Heaven awaits). Similar beliefs are common among the world's numerous religions.

My final concern is M&D's rejection of supernatural beliefs as adaptive, which hinges on a perceived lack of empirical evidence. This is problematic for three reasons. First, in the literature M&D focus on, researchers tend to use religious primes derived from Western Judeo-Christian traditions (e.g., "divine," "God," and "prophet" in Shariff & Norenzayan 2007), whereas the relevant supernatural concepts in our evolutionary history could be anything from dead ancestors, spirits, ghosts, witches, inanimate objects, and so forth. Similarly, modern *religious* agents are only one possible type of supernatural agency, whereas subjects' behavior may also be influenced by other *sources* such as superstition, folklore, karma, Just World beliefs (the belief that victims of tragedy somehow deserved it), or everyday "comeuppence" and "just deserts." Given this diversity of possible supernatural agents and sources, personal religious

affiliations and devoutness among experimental subjects may be somewhat independent of how supernatural beliefs – in general – influence people's behavior (M&D predict an interaction of personal religious devoutness and behavior). Current experiments may not, therefore, be able to differentiate the behavior of "believers" and "non-believers" – Joe Bloggs may be an avowed atheist who, on his way to Las Vegas, is nevertheless very concerned about seeing a black cat or wearing his lucky jacket or what his grandmother would have said.

Second, even if we had incontrovertible evidence that supernatural cues (e.g., via experimental primes) promoted higher donations in economic games, this is far from evidence that religious beliefs are biologically *adaptive.* On the contrary, it could be evidence that religious primes turn people into suckers who give away precious resources. Such behavior, on its own, would not survive natural selection – without additional field experiments measuring fitness consequences, evidence for altruism is hardly evidence of an adaptive trait. Therefore, the (excellent) current laboratory experiments that M&D focus on cannot yet be used as deal-breakers as to whether (mis)belief is adaptive or not.

Third, having rejected supernatural beliefs as adaptive, M&D's null hypothesis is that religious beliefs are a non-adaptive byproduct of cognitive mechanisms adapted for other purposes – evolutionary accidents, in other words. However, if religious beliefs are accidental byproducts, we might expect natural selection to have eradicated them because (as M&D note) they impose significant fitness costs in terms of time, effort, and resources (Sosis & Alcorta 2003). So why do they persist?

Even if some religious beliefs persist as "sticky" cultural parasites, it does not preclude them from also promoting individual or group fitness at certain times or contexts (in which case they may not be "parasites"). The universality and power of religious beliefs of some form or other – despite their costs – to billions of people around the world, every culture in history, and every hunter-gatherer society, strongly suggests that religion confers adaptive fitness benefits, for individuals and/or groups (at least in some contexts, for some people, and for some periods of human history). Of course, universality need not imply adaptation: other non-adaptive traits such as chins and male nipples are also globally and historically universal. However, they do not impose significant costs. Religion does.

The only theories that solve this paradox are religion-as-adaptive hypotheses that propose how costly (mis)beliefs beget even greater benefits for individuals and/or groups (Johnson 2008; Norenzayan & Shariff 2008; Sosis & Alcorta 2003; Wilson 2002), or are outweighed by the costs of non-belief (Cronk 1994; Johnson 2009; Johnson & Bering 2006). Byproduct theories of religion offer no solution to its greatest puzzle, for God would be a costly accident.

A positive illusion about "positive illusions"?

doi:10.1017/S0140525X09991257

Vladimir J. Konečni

Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109.

vkonecni@ucsd.edu

http://psychology.ucsd.edu/people/faculty/vkonecni.php

Abstract: Rather than being a genuine adaptation, "positive illusions" are examples of doxastically uncommitted policies implemented at both the individual and societal levels. Even when they are genuine misbeliefs, most positive illusions are not evolved but ephemeral – a phenomenon limited to a particular social and economic moment. They are essentially a consumer response to messages from the pop-psychology industry in the recently terminated era of easy credit.

The article by McKay & Dennett (M&D) presents a thoughtful classification and analysis of the evolutionary issues involved in misbelief. The notion that certain misbeliefs may arise through the normal functioning of the belief-formation system (as opposed to its breakdown), by virtue of relying on incomplete or inaccurate information, is clearly acceptable (and not new). What is debatable, however, is the authors' key proposition, that a subclass of such misbeliefs has been systematically adaptive in the evolutionary past. The usefulness of this suggestion depends to a large extent on finding an example that meets the authors' commendably sound and strict criteria, yet the sole example of adaptive, evolved misbelief that is proposed by M&D, "positive illusions," is not convincing.

One should first note that the concept of positive illusions, as well as the term itself and psychologists' (mis?)beliefs about the positive consequences of self-serving distortions of reality, are all of quite recent vintage (e.g., Taylor 1989). M&D helpfully contrast this view of mental health with Jahoda's (1958) earlier and more measured one. Whether or not one wishes to engage in culture-theorizing about the contrast between the fear of social ridicule and the self-restraint evident in the Eisenhower era (demonstrated, for example, in Asch's [1956] "social conformity" experiments), on one hand, and the less-disguised greed and self-promotion of the more recent I-want-it-all-now generations, on the other, the fact is that the content of many positive illusions is a quite recent phenomenon and that the results of many of the studies are likely to be ephemeral and support Gergen's (1973) "social psychology as history" view. It is therefore risky (if not unwarranted) to be talking about the creation and implementation of misbelief as adaptive – let alone adapted – selection-driven behaviors (see the authors' Note 3).

There are also questions about the empirical evidence marshaled by M&D (all of it dating from after about 1985). A number of studies purporting that "most people ... see themselves as better than most others on a range of dimensions" (target article, sect. 13, para. 2) appear to be methodologically unsound. M&D should have more closely examined the presence of problems and alternative explanations related to the framing of questions, the differential social desirability of various response alternatives, and the Pygmalion effect before implying that positive illusions were present and favored in the ancestral environment. In addition, quotes from psychologists firmly committed to the environmentalist position – such as the social-learning theory, with its (mis?)beliefs about the teachability and ready amelioration of just about every personal shortcoming – cannot be considered an entirely unbiased source.

Other studies have tended to ignore the participants' referential framework and may not have dealt with *mis*beliefs. For example, people who claim that their current partner is better than most are likely to be referring to their past partners' failings and the undesirable traits and behaviors of people in all those failed marriages that they know and read about. Even with regard to an inflated opinion of one's children, studies have presumably not polled the opinions of the parents (including potential ones) who terminated pregnancies – or who committed infanticide, physical and/or sexual abuse, and the more common acts of neglect. If even such parents, as is possible and even likely, were to have an inflated idea of the merits of their offspring and potential offspring, this would raise interesting questions about the meaningfulness of using the questionnaire-retrospective research approach to probe matters relevant to evolutionary adaptation.

To the extent that positive illusions can, in fact, be adequately documented (regardless of whether or not they are evolved, adaptive misbeliefs), it is of interest to try to place them in a broader contemporary context. If people's positive illusions about their personal worth and ability are translated into behavior evident to others, all sorts of negative consequences are likely to ensue, from mild ridicule to severe ostracism. Unless, that is, the unbridled expression of positive illusions has been

proclaimed a desirable social norm. It is clear that there is no shortage, perhaps especially in the United States, of change agents, and socio-cultural, economic, and even legal factors, involved in the encouragement of positive illusions: the generally prevailing environmentalist ("nurture") bias in the educational system and mass culture; the politico-legal doctrine of universal entitlement and reduced personal responsibility (including exaggerated emotivist explanations of both legal and illegal behavior); and the broad societal push toward spending on credit, embodied in the "optimistic" (something for nothing, "no money down") consumption-based policies for one's alleged betterment and advancement. The empirical findings, to the extent that they are reliable and valid in the first place, document what is essentially a consumer response to the ubiquitous messages from the pop-psychology and advertising industries, which make wildly unrealistic promises and encourage an assertive expression of self-worth. Most of these are American homegrown products, but they have been distributed widely, especially in the Western world.

However, the present financial crisis may have already provided a corrective to positive illusions at both the personal and societal levels. The crisis has certainly led to a dramatic drop in the previously inflated average self-image, for example, by people in countries as different as Iceland and Latvia. Predictably, Western politicians and bankers will resist this trend. Quite recently, in the *Financial Times*, the executive chairman of the giant international banking concern HSBC declared: "About 80 per cent of this country [United Kingdom] considers itself middle class. I doubt that was true then [a generation ago]" (Barber 2009). Yet, on the same date and in the same news source, it was reported that McDonald's is the largest private employer in *France* (Morrison 2009).

One is left with the conclusion that rather than being a genuine adaptation, most positive illusions are examples of doxastically uncommitted action policies implemented at both the individual and societal levels; and even when they are doxastically relevant, genuine misbeliefs, they are unlikely to be evolved and adaptive – and are instead an ephemeral phenomenon limited to the present social and economic moment. Or perhaps limited to the recent past, for there are already signs of a reduction of positive illusions as a function of the current financial crisis.

Benign folie à deux: The social construction of positive illusions

doi:10.1017/S0140525X09991269

Dennis L. Krebs[a] and Kathy Denton[b]

[a]*Department of Psychology, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada;* [b]*Psychology Department, Douglas College, New Westminster, BC, V3L 5B2, Canada.*

krebs@sfu.ca dentonk@douglas.bc.ca

http://www.sfu.ca/psyc/faculty/krebs/publications.htm

Abstract: McKay & Dennett (M&D) have done an admirable job of distinguishing among various forms of misbelief and evaluating the idea that they stem from evolved mental mechanisms. We argue that a complete account of misbeliefs must attend to the role that others play in creating and maintaining positive illusions.

In their analysis of the sources of misbeliefs, McKay & Dennett (M&D) focus on how belief-producing mental mechanisms are designed. Although people may develop beliefs on their own and cherish them in private, they acquire many of their beliefs from others, and they use other people to evaluate them. Often, these beliefs pertain to ephemeral phenomena for which there are no objective criteria, such as whether one is likable or attractive. In contexts in which individuals stand to benefit from accurate representations of reality, they may solicit reality checks from others and correct their beliefs accordingly. However, when individuals stand to benefit from misrepresentations of reality, they may manipulate others into validating them, which in turn may help the manipulators believe that the misrepresentations are true.

Evolutionary theory leads us to expect people to be disposed to seek the truth when truth-seeking is the most adaptive strategy. However, truth-seeking is not always the most adaptive strategy, and the evidence clearly establishes that people are not naturally inclined to process all social information in objective or impartial ways. As expressed by Haidt:

> Research on social cognition . . . indicates that people often behave like "intuitive lawyers" rather than like "intuitive scientists". . . . Directional goals (motivations to reach a preordained conclusion) work primarily by causing a biased search in memory for supporting evidence only. . . . Self-serving motives bias each stage of the hypothesis-testing sequence, including the selection of initial hypotheses, the generation of inferences, the search for evidence, the evaluation of evidence, and the amount of evidence needed before one is willing to make an inference. (Haidt 2001, p. 821)

Strategies of social belief validation. People invoke several strategies to maximize the probability that others will validate their misbeliefs about themselves and others. First, they express their misbeliefs selectively to those they consider most likely to validate them – usually people who have a vested interest in the misbeliefs or in pleasing the misbelief-holder. For example, in conversations with in-group members, people express beliefs that favor their in-groups and demean their out-groups, and they express beliefs about their worth to their friends and relatives. M&D give examples of positive illusions that increase people's chances of surviving by improving their health. Because such beliefs also may benefit those whose fitness is linked to sick people's welfare, these people may have a vested interest in adopting and supporting them. For example, believing that a friend, mate, or relative will recover from an illness may induce one to behave in ways that increase the probability of him or her recovering, which in turn may enhance one's welfare.

Second, people buttress the misbeliefs they voice to others with a biased sample of evidence. For example, people may brag about their successes and hide their failures. And finally, people turn to others to support their misbeliefs. For example, Denton and Zarbatany (1996) found that when people made mistakes, they made excuses to their friends, who in turn supported them. An interesting dynamic often occurs when people express a biased sample of evidence to their friends in support of their misbeliefs: Their friends end up forming more extreme misbeliefs than the people seeking validation are comfortable accepting. For example, Krebs and Laird (1998) found that participants' friends made more exculpating judgments for the transgressions that the participants committed than the participants made themselves.

Social conspiracies. Friends and relatives tend to engage in a subtle form of reciprocity with respect to positive illusions about one another – "You support my illusions, and I will support yours" – which gives rise to benign folie à deux: "You are wonderful." "So are you." In some cases, this initiates a self-fulfilling prophesy. If each of us thinks that the other is socially attractive, funny, beautiful, of high worth, then our beliefs are at least partially validated.

Adaptive functions of illusions about one's worth. In an earlier paper, we asserted that illusions about one's own worth are adaptive because they help people deceive others about their worth (see Krebs & Denton 1997). M&D questioned this assertion, because they questioned whether "others *are* deceived about the worth of self-deceptive individuals" (target article, sect.

12, para. 4, emphasis theirs), and because the results of a study revealed that people tend not to trust *highly self-deceptive people* (emphasis added). However, our assertion pertained to an adaptive function of species-specific illusion-producing mechanisms, not to the adaptive function of high levels of self-deception. Extreme degrees of self-deception (high and low) are probably maladaptive. Our claims can be understood in the context of the social model of misbeliefs we have outlined in this commentary. Everyone is self-deceptive in the sense that everyone harbors positive illusions about their worth. Everyone is disposed to propagate and reinforce these positive illusions by deceiving others about their worth, and this process pays off in a variety of ways.

Although people's (mis)beliefs about their worth are bound to affect others' judgments, there are two important constraints on the extent to which observers are fooled by invalidly high estimates of people's worth. First, observers inevitably evaluate these beliefs on other criteria, such as their predictive ability and the extent to which they are shared by others. We would not expect high levels of self-deception that give rise to huge discrepancies between evaluations of one's self-worth and more objective evidence to be persuasive to others.

Second, observers are sensitive to the costs and benefits of misreading signals of others' worth. In general, observers are evolved to be wary of deception in contexts in which it is more adaptive to form accurate estimates of others' worth than it is to form inaccurate estimates. In the absence of other information, such as when people are making a first impression, people's conceptions of their worth may be relatively persuasive within an optimal range. However, we would not expect others to support highly exaggerated illusions except in rare cases in which they collaborate in the illusions, such as pathological cases of folie à deux. If you are the King, I am the Queen.

(Not so) positive illusions

doi:10.1017/S0140525X09991270

Justin Kruger[a], Steven Chan[b], and Neal Roese[c]

[a]*Stern School of Business, New York University, New York, NY 10012;* [b]*Department of Marketing, Stern School of Business, New York University, New York, NY 10012;* [c]*Kellogg School of Management, Northwestern University, Evanston, IL 60208.*

jkruger@stern.nyu.edu
http://w4.stern.nyu.edu/faculty/facultyindex.cgi?id=370
schan@stern.nyu.edu
http://w4.stern.nyu.edu/marketing/research.cfm?doc_id=872
n-roese@kellogg.northwestern.edu
http://www.kellogg.northwestern.edu/Faculty/Directory/Roese_Neal.aspx

Abstract: We question a central premise upon which the target article is based. Namely, we point out that the evidence for "positive illusions" is in fact quite mixed. As such, the question of whether positive illusions are adaptive from an evolutionary standpoint may be premature in light of the fact that their very existence may be an illusion.

> *It ain't so much the things we don't know that get us into trouble. It's the things we know that just ain't so.*
>
> — Artemus Ward (as cited in Gilovich 1991)

When the *Journal of Personality and Social Psychology* (*JPSP*) compiled a list of the papers most often cited in its pages, the salutatorian was not a *JPSP* article (nor was the valedictorian, but that is another story). Nor was it even a social psychology article – at least not exclusively (Quinones-Vidal et al. 2004). It was Taylor and Brown's (1988) review of "Positive Illusions," the series of biases that McKay & Dennett (M&D) nominate as one of the "best candidates for *evolved misbelief*" (target article, sect. 13).

And for good reason (both the citation count and nomination). Taylor and Brown (1988) not only advanced a compelling argument as to why some judgmental biases might actually be healthy (an argument equally compellingly taken to its logical extreme by the target article), but also provided an elegant summary of the various ways in which people see themselves and the world around them in an unrealistically favorable light. People overestimate how their traits and abilities stack up against those of others. They believe that their futures are rosier than those of the average person. And they believe that they control what they do not. Collectively, these "positive illusions" instantly rang true for both psychologists and laypeople alike, corroborating both common wisdom and an idea as old as psychology itself (Freud 1920; Roese & Olson 2007). Indeed, so intuitive is the insight that it is hard for most of us even to imagine a time in which it could have been controversial, but indeed it once was (Festinger 1954). As we now know, "most people view themselves as better than average on almost any dimension that is both subjective and socially desirable" (sect. 13, para. 2).

The problem, to paraphrase the opening quote, is that sometimes what we know just ain't so.

Consider what is perhaps the best-known illustration of the "above-average effect," from the College Board Survey (1976–77). When 828,516 high-school students taking the SAT were asked about their leadership ability, 70% claimed to be "above-average" and only 2% "below-average." When asked about their "ability to get along with others," 89% reported that they were above-average, fewer than 0.5% admitted to being below-average, and a full 25% reported being in the top first percentile.

One could point out that the results are misleading, and they are most certainly that. For one, the students may have felt pressured to present themselves in a favorable light. The questions, after all, were being asked by the same people determining their suitability for college. For another, the colloquial definition of "average" is considerably gloomier than the dictionary definition. Most of the students, therefore, could very well have been above-average – at least by their construal of the term. Even by the mathematical definition the sample may have been above-average. The sample, although large, was comprised exclusively of high-school students taking the SAT, the majority of whom could have very well been better leaders and more of the "get-along-with-others" type than the average high-school student (this is even possible with a representative sample if the distribution of performance is sufficiently skewed, but never mind that).

All of these are reasons to question the results of the College Board Survey and the many others like it. They are not the reasons to doubt the existence of positive illusions such as the "above-average effect," however, because the effect replicates even when the above factors are accounted for (e.g., Kruger & Burrus 2004; Dunning et al. 1989). Nor even is the reason for doubt the fact that the above-average effect and unrealistic optimism often fail to replicate in "collectivist" cultures such as those found in East Asia, South America, and the Middle East (Heine et al. 1999; Henrich et al., in press).

Instead, the reasons for doubt are the many instances in which healthy populations exhibit, not positive illusions, but systematic negative ones. In the case of the above-average effect, although people overestimate their ability relative to others on easy tasks, such as using a computer mouse or riding a bicycle, they tend to underestimate their ability on difficult tasks, such as programming a computer or telling a really good joke (Kruger 1999; Moore 2007). Similarly, although college students believe that they are more likely than the average student to experience common desirable events such as getting a job with a starting salary over $25,000 or living past 70, they believe that they are less likely than the average person to land a salary over $250,000 or live past 100 – despite the fact that the latter are at least as desirable as the former (Kruger & Burrus 2004).[1]

Indeed, for virtually every social comparison bias one might label "positive illusion" there appears to be a complementary "negative illusion." People overestimate their likelihood of beating a competitor when the contest is simple (such as a trivia contest involving easy categories), but underestimate those odds when it is difficult (such as a trivia contest involving difficult categories; Moore & Kim 2003; Windschitl et al. 2003). Roommates overestimate their relative contribution to tasks involving frequent contributions like cleaning the dishes, but underestimate their relative contribution to tasks involving infrequent contributions like cleaning the oven (Kruger & Savitsky 2009). And preliminary work suggests that although people overestimate their degree of control over that which can be controlled easily, they underestimate their degree of control of what cannot (Kruger, unpublished data).

That said, it would be misleading to suggest that these results imply that, ecologically speaking, negative illusions outnumber positive ones. The above research does not speak to this question, nor is it an easy – or perhaps even possible – question to answer. What the research does suggest, however, is that there is reason to question the assumption that positive illusions are the norm for healthy individuals. It may be one of those "things we know that just ain't so" (Artemus Ward, as cited in Gilovich 1991). It is perhaps worth noting that it was likely not Artemus Ward who said those words, but instead the late humorist Josh Billings.[2]

NOTES

1. As the reader may have noticed, the examples are related. In each case, people overestimate relative standing when absolute standing is high and underestimate relative standing when absolute standing is low (see Chambers & Windschitl 2004; Moore & Healy 2008).

2. Thomas Gilovich, personal communication.

Pathological and non-pathological factors in delusional misbelief

doi:10.1017/S0140525X09991282

Robyn Langdon

Macquarie Centre for Cognitive Science, Macquarie University, Sydney, NSW 2109, Australia.

rlangdon@maccs.mq.edu.au

http://www.maccs.mq.edu.au/members/profile.html?memberID=60

Abstract: In their pursuit of adaptively biased misbelief-making systems, McKay & Dennett (M&D) describe a putative doxastic shear-pin system which enables misbeliefs to form in situations of extreme psychological stress. Rather than discussing their argument, I consider how this shear-pin system might combine with both pathological belief-making ("culpable" breakdowns caused by neuropathy) and normal belief-making to explain a spectrum of delusions.

Textbook definitions of delusions (as, e.g., false beliefs) are inadequate to capture the nature of pathologies that cause delusions (see David 1999; see also McKay & Dennett's [M&D's] comment that delusions might sometimes be "serendipitously true" [sect. 1, para. 2]). The three signs that clinicians use to diagnose delusions (incorrigibility, subjective certainty, and incomprehensibility) are better pointers to the nature of these pathologies than are textbook definitions. Incorrigibility refers to the rigid persistence in the face of rational counter-argument. Subjective certainty concerns the quality of self-evident truth with which delusions are espoused. Incomprehensibility can be "sheer" or "contextual" (Langdon & Bayne, in press). The sheer content of some delusions is sufficient as to render them incomprehensible; the belief that one is dead (Cotard delusion) is like this. Mundane delusions, like delusional jealousy, are contextually incomprehensible; patients with these delusions lack the evidence, or lack the irrefutable evidence that would warrant the subjective certainty with which the delusion is espoused (e.g., a patient once vehemently justified her persecutory delusion about a neighbour by referring to the provocative way in which the neighbour had intentionally jingled her keys when walking ahead of the patient).

In our original "two-deficit" model, Langdon and Coltheart (2000), we proposed two distinct, concurrent pathologies to explain the incomprehensibility and the incorrigible, subjective certainty of bizarre monothematic delusions. The first breakdown explains why a delusional person generates a fantastic thought with content so beyond the bounds of normal experience. This first break also varies from patient to patient to explain the variable delusional themes. First breaks to sensory/somatosensory mechanisms were our primary focus; these distort perceptual experience and so explain the sheer incomprehensibility of bizarre delusions. Since many non-delusional people experience distorted perceptions (e.g., phantom-limb sufferers), a second break was proposed to explain the failure to reject the bizarre thought as implausible: the incorrigibility. Current conceptions of this second break vary; see, for example, M&D's discussion of excessive biases towards either observational adequacy or doxastic conservatism. Langdon and Bayne (in press) propose that the second break in bizarre monothematic delusions is an inability to inhibit a default setting to upload and maintain the content of perceptual experience into belief (see also, Davies et al. 2001). For example, the patient with a mirrored-self misidentification delusion misperceives a stranger in the mirror, misbelieves that there is a stranger in the mirror, and cannot reason as if there only seems to be a stranger in the mirror. It is this inability to inhibit the misperceived reality that explains the self-evident, subjective certainty.

In Langdon and Coltheart (2000), although we focused on pathology, we also considered normal belief-making. Normal processes, we suggested, might nuance specific elaborations of a bizarre delusion: One Cotard patient with an internalising attributional bias might believe that God is punishing her for her evil ways, while another Cotard patient with an externalizing bias might believe that evil doctors have stolen her "life essence." Normal processes might also feature in the generation of mundane delusions by way of expectation-fuelled attentional biases concerning, for example, a straying partner (delusional jealousy) or health concerns (hypochondrial delusions). Even a mundane grandiose delusion, like believing that one is a gifted pianist despite poor playing skills and vocal audience criticism, might begin with the positive illusions that interest M&D.

Combinations of pathological and normal belief-making explain: normal beliefs about normal experiences (with no pathological breaks present); normal beliefs about bizarre experiences (with only a first break present, as occurs, e.g., when an aberrant signal of familiarity causes insightful déjà vu); mundane delusions (with only a second break present); and bizarre delusions (with two breaks present). Expanding the framework in this way prompts two questions, though. The first concerns the adoption of the belief. The adoption of a bizarre delusion which has been triggered by some disturbance of sensory/somatosensory processing is relatively straightforward to explain by way of a default to believe our senses, at least initially. The adoption of a mundane delusion is less straightforward to explain; when and why does a worrying or a fantasizing become a convicted believing? The second question concerns the nature of the second break; is it the same in mundane and bizarre delusions? This seems unlikely if the second break in bizarre delusions is an inability to inhibit a default setting to believe what our senses tell us.

In a recent review of persecutory delusions, which are often mundane, Langdon et al. (2008) found no compelling evidence for the involvement of right-frontal brain damage; we suspect that brain damage of this type underpins the second break in bizarre delusions. More recently, therefore, we have shifted to a more general, "two-factor" approach (e.g., Coltheart 2007) to ask two questions about each delusion: (1) What generates the

delusional content in the first place? (2) Having once entertained a particular thought, why does a deluded patient cling to it rather than reject it? Sometimes the answers will be neuropsychological and sometimes they will be motivational.

But, even if the aetiology of the second break is sometimes motivational and sometimes neuropsychological, the second break might still be similar at an *on-line* cognitive level. M&D's idea of "doxastic shear-pins" is relevant here. If belief-making components shear in situations of extreme psychological stress to permit beliefs that would ordinarily be rejected, I assume that the shearing is localized and constrained by the context. I also assume that the shearing involves some on-line neural/cognitive "short-circuit," as opposed to a stable neuropsychological impairment. If so, then perhaps we might describe the second break in all delusions, bizarre or mundane, as a "doxastic inhibitory failure": a failure to "demote" a belief so as to reason about it as if it might not be true. In bizarre monothematic delusions, this failure might only manifest via an inability to inhibit a default tendency to upload and maintain (distorted) perceptual experience into (mis)belief; in mundane motivated delusions this failure might only manifest when the psychological cost of demoting the belief into a "maybe-it's-not-true" mental space is too great; and in dementing patients with widespread bizarre and/or mundane delusions this failure might reflect more general inhibitory compromise.

ACKNOWLEDGMENTS

Thanks to Neralie Wise and Emily Connaughton for comments on an earlier draft.

Are beliefs the proper targets of adaptationist analyses?

doi:10.1017/S0140525X09991294

James R. Liddle and Todd K. Shackelford

Department of Psychology, Florida Atlantic University, Davie, FL 33314.

jliddle1@fau.edu

http://www.jamesrliddle.com tshackel@fau.edu

http://www.toddkshackelford.com

Abstract: McKay & Dennett's (M&D's) description of beliefs, and misbeliefs in particular, is a commendable contribution to the literature; but we argue that referring to beliefs as adaptive or maladaptive can cause conceptual confusion. "Adaptive" is inconsistently defined in the article, which adds to confusion and renders it difficult to evaluate the claims, particularly the possibility of "adaptive misbelief."

McKay & Dennett (M&D) open their article by presenting what they consider the "prevailing assumption" (sect. 1, para. 2) of modern evolutionary analyses of belief, namely, that true beliefs are adaptive and misbeliefs maladaptive. However, M&D also present and appear to endorse the content of several quotations (e.g., from Bloom 2004; Ghiselin 1974; Haselton & Nettle 2006; Stich 1990) that showcase an alternative evolutionary perspective: that beliefs are not relevant to natural selection unless they contributed recurrently to differential reproduction, and, furthermore, that there is no reason to assume that only the true beliefs of our ancestors met this criteria. These quotes suggest that what M&D refer to as the "prevailing assumption" of evolutionary analyses of belief is in fact *not* the prevailing assumption; but this is a relatively minor issue that we do not explore further in this commentary. Instead, we address a more pressing concern: M&D's analyses are not based on a coherent definition of "adaptive."

The aim of the target article is to evaluate the assumption that misbeliefs *themselves* are maladaptive, and to examine candidates for *adaptive misbelief*. Considering this aim, we find it surprising that M&D do not provide coherent definitions of the relevant phenomena. Adaptive misbeliefs are loosely defined at various places in the article as beliefs that are "normal" (sect. 5, last para.), "beneficial" (sect. 6, first para.), that "aid survival" (sect. 6, first para.), "maximize fitness" (sect. 9, para. 5), "facilitate the negotiation of overwhelming circumstances" (sect. 10, para. 2), "facilitate the successful negotiation of social exchange" (sect. 12, first para.), promote "mental health" (sect. 13, first para.), and "sustain and enhance *physical* health" (sect. 13, para. 3, emphasis in original). M&D do not explicitly define "adaptive misbelief." Rather, they pose several questions throughout the article in the process of evaluating the plausibility of adaptive misbelief, and these questions imply the sundry definitions we noted. Because it does not provide a specific definition of "adaptive," the article lacks a consistent framework for evaluating the candidates for adaptive misbelief.

Although M&D acknowledge in a note (Note 3) that they conflate conceptually "adaptive" and "adapted" throughout the article, this acknowledgement does not diminish any confusion, as the reader is left without a specific definition for either term. M&D also highlight the distinction between psychological adaptation and biological adaptation. These terms are loosely defined with reference to a distinction between "human happiness and genetic fitness" (sect. 10, para. 5) – with genetic fitness loosely defined as "having more surviving grandoffspring" (sect. 10, para. 5). The latter definition misses many of the conceptual nuances associated with the concept of fitness from an evolutionary perspective (see Dawkins 1982). Such an oversimplification is particularly problematic for an article whose arguments hinge on whether beliefs have had an effect on fitness throughout our evolutionary history, which would ultimately determine the status of beliefs as adaptations in and of themselves.

M&D do avoid a potential confusion in their article by making a clear distinction between beliefs themselves and the information-processing mechanisms that generate beliefs (sect. 5, last para.). M&D clearly state that they are interested in the subset of misbeliefs that are generated by properly functioning cognitive mechanisms, and that these are the candidates for adaptive misbelief. However, M&D do not justify focusing on beliefs themselves as opposed to the mechanisms that generate beliefs, even though a proper adaptationist perspective (Tooby & Cosmides 1992) focuses not on the output of adaptations (e.g., beliefs), but on the design features of the adaptations (e.g., the information-processing mechanisms that generate beliefs). If the information-processing mechanisms of the mind are sensitive to context (Buss et al. 1998), then it is plausible that a belief-generating mechanism can generate true beliefs in one environment and false beliefs in a different environment. Our understanding of why specific beliefs are formed requires an understanding of the mechanisms that generate the beliefs, and referring to beliefs themselves as adaptations obfuscates the importance of the actual adaptations (i.e., the underlying mechanisms).

Despite some conceptual confusion, M&D present several thought-provoking concepts in the target article. For example, their categorization of misbeliefs in terms of the functioning (or malfunctioning) of the belief formation systems provides an important distinction, although we were surprised to see no reference to Wakefield's (1992) strikingly similar and pioneering evolutionary analyses of dysfunction. We also appreciate the concept of "doxastic shear pins" (sect. 10), which may offer a solid foundation for future empirical and theoretical work on belief formation in extraordinary, psychologically stressful situations. Finally, M&D's analysis of beliefs suggests an alternative to the proper adaptationist perspective by referring to the *output* of psychological mechanisms as adaptations. However, the merit of this alternative is difficult to determine, due to the target article's many conceptual confusions.

10,000 Just so stories can't all be wrong

doi:10.1017/S0140525X09991452

Gary F. Marcus

Department of Psychology, New York University, New York, NY 10012.
gary.marcus@nyu.edu http://www.psych.nyu.edu/gary/

Abstract: The mere fact that a particular aspect of mind could offer an adaptive advantage is not enough to show that that property was in fact shaped by that adaptive advantage. Although it is possible that the tendency towards positive illusion is an evolved misbelief, it it also possible that positive illusions could be a by-product of a broader, flawed cognitive mechanism that itself was shaped by accidents of evolutionary inertia.

In arguments about innateness, one often finds a bias towards empiricist perspectives. It is widely, though erroneously, believed that if some aspect of cognition *could* be learned, that aspect of cognition must be the product of learning; evidence for the possibility of learning is often taken as evidence against possibility of innateness. Of course, in reality, some aspects of cognition could be innate, even if they were in principle learnable. Humans might, in principle, be able to *learn* how to walk, much as they can acquire other new motor skills (e.g., juggling or skiing), but the fundamental alternating stepping reflex that underlies walking appears to exist at birth, prior to any experience of actual walking.

In a similar way, in discussions about adaptive advantage, it sometimes seems as if there is a bias towards adaptationist accounts relative to by-product accounts, such that any putative adaptive advantage apparently automatically trumps the possibility of non-adaptive accounts. If something could have been shaped by adaptive pressure, it is often assumed to have done so; but it is again a logical error to assume that simply because something *could* be explained as an adaptation then it could not also be explained in another fashion.

McKay & Dennett (M&D) make a reasonably strong case for the possibility that positive illusions, such as non-veridical beliefs about one's health, could have a history borne of direct adaptive advantage; is there any reason to consider alternative accounts? Quite possibly: Although positive illusions might inhere in some sort of domain-specific cognitive substrate that could have been specifically shaped via natural selection, I believe it is at least equally plausible that positive illusions are simply one manifestation among many of two considerably more general phenomena that are pervasive throughout human cognition (yet curiously absent from the target article): (1) confirmation bias (e.g., Nickerson 1998), in which people often tend to cling to prior beliefs even in the light of contradictory evidence; and (2) motivated reasoning, a tendency of people to subject beliefs that are potentially ego-dystonic to greater levels of scrutiny that are less likely to be ego-dystonic (Kunda 1990).

Cigarette-smokers, for example, tend to dismiss research on the dangers of smoking not because they believe themselves to be healthy, but because they work harder to deflate potentially damaging beliefs. This could be seen as an instance of a cognitive mechanism that was dedicated towards yields positive self-illusion. But much the same tendency towards motivated reasoning can be seen in undergraduates' evaluations of arguments about capital punishment: As Lord et al. (1979) showed, students tend to work harder to undermine arguments that conflict with their prior beliefs. Dozens of subsequent studies, reviewed in Kunda (1990), point in the same direction: We work harder to dismiss arguments that we don't like, whether or not those arguments pertain to our own personal well-being. Positive illusion might in this way be seen as an instantiation of motivated reasoning, rather than as the product of a dedicated mechanism with a unique adaptive history.

A species-general tendency towards confirmation bias might also subserve positive illusions, even absent specific machinery dedicated to positive illusion. Once one stumbles on a belief in one's own virtues (e.g., through the praise of one's parent), potentially disconfirming evidence may be ignored, underweighted, or simply harder to retrieve; but again, there may be nothing special about personal self-interest. Confirmation bias is as apparent in people's inferences about arbitrary rules in concept learning tasks as it is in beliefs about self (again, see Nickerson 1998, for a review).

Intriguingly, confirmation bias itself may be a byproduct of the organization of human memory (Marcus 2008; 2009). Human memory, like that of all vertebrate creatures, is organized via context rather than location. In a machine with location-addressable memory, it as easy to search for matches to be a particular criterion (data that fits some theory) as data that does not match said criteria (potentially disconfirming evidence). By contrast, in a creature that can search only by contextual matches, there is no clean way in which to access disconfirming evidence. Confirmation bias itself may thus stem from inherited properties of our memory mechanisms, rather than specific adaptive advantage. Indeed, as M&D themselves argue, an across-the-board bias towards misbelief per se is unlikely to be adaptive.

At the present time, we simply lack the tools to directly infer evolutionary history. Some adaptationist theories are likely to be correct – parental investment theory, for example, is supported by a vast range of supporting evidence; others, such as the theory that depression serves to keep its bearers from getting into trouble, seem rather more dubious. Advances in understanding how cognitive machinery is instantiated in underlying brain matter may help, as may advances in relating genetic material to neural structure; for now, we are mostly just guessing. Are positive illusions in fact evolved misbeliefs or are they merely by-products of more general mechanisms? We really can't say. My point is simply that we should be reluctant to take either option at face value. Specific properties of cognitive machinery may sometimes turn out to be by-products even when it superficially appears in principle that there is an adaptive pressure that could explain them.

It is likely misbelief never has a function

doi:10.1017/S0140525X09991300

Ruth Garrett Millikan

Philosophy Department U-54, University of Connecticut, Storrs, CT 06269-2054.
ruth.millikan@uconn.edu

Abstract: I highlight and amplify three central points that McKay & Dennett (M&D) make about the origin of failures to perform biologically proper functions. I question whether even positive illusions meet criteria for evolved misbelief.

According to Stephen Stich, "natural selection does not care about truth; it cares only about reproductive success" (Stich 1990, p. 62; quoted in the target article, sect. 9, epigram). Similarly, I suppose, natural selection "does not care about" digesting food, pumping blood, supplying oxygen to the blood, walking, talking, attracting mates, and so forth. For each of these activities can either be (biologically purposefully) set aside (the vomiting reflex, holding one's breath under water, sleeping) or simply fails to occur in many living things. Nonetheless, surely the main function for which the stomach was selected was the digestion of food, the lungs for supplying oxygen, and so forth, and a main function for which our cognitive systems were selected was the acquisition and use of knowledge – that is, true belief. As McKay & Dennett (M&D) observe, confusions about this arise from failing to take into account any of three fundamental facts about biological function, on each of which I would like, very briefly, to expand.

The first is that recurrent failure to perform the functions for which they were selected is completely normal for many biological structures and activities. The barnacle waves its little fan foot through the water once, twice, ten times, a hundred and ten times, and the hundred and eleventh time it picks up a microscopic lunch. One hundred ten failures for one success. (Bad statistics characterize almost all hunting and fleeing behaviors of animals.) A job of the gazelle's strong leg muscles is to allow it to outrun the lion, a job of the protective eye blink reflex is to keep sand out of the eye, and a job of various body membranes is to keep pathogens from entering; but, of course, none of these jobs always gets done. Our knowledge-making systems performed their functions less and less reliably, up to very recent times, I suppose, roughly as the objects of belief got further and further from in front of our noses. But, of course, it is getting straight about what is in front of our noses that is the first order of importance for us. Getting straight about things further away can be very helpful too, when we can manage it; and when we don't manage, false belief about distant things may, until recently, have been pretty much equivalent, in the simple randomness of the results, to having no belief at all about such matters. As with hunting, clearly it is better to try, even though one fairly often fails, than not to try at all. This is very different, of course, from saying that failures are helpful or that the organism has been designed to produce them. Failures are by-products of design for success. (On the inconstancy of Normal supporting conditions for proper operation of the cognitive systems, see Millikan 1998.)

The second observation is that biological systems may sometimes be designed to suspend or override the functions of certain of their parts, so as to avoid damage when Normal conditions for successful operation – operative conditions that helped account for past successes hence for selection of these parts – are conspicuously absent. Numerous animal species "play dead," perhaps actually going unconscious, in circumstances where any manifestation of normal life will only raise their chances of injury or death. The fuse blows, the shear pin breaks, by design. Again, this does not imply that failures to function properly are helpful, but only that in some circumstances it is best not to attempt to function at all.

The third observation is that structures kept in place by natural selection primarily for one purpose are sometimes also utilized by piggyback mechanisms to help serve different functions, perhaps even interfering with their original functions on occasion. Certain kinds of beliefs might be useful to us for purposes other than their normal cognitive use regardless of truth or falsity (not, however, *because* of their falsity, as M&D emphasize). This is theoretically possible, but I do not think convincing evidence for it has been offered. If certain kinds of errors are common and also systematically useful, it does not follow that they are common because they are useful (compare our first observation above). It is also very hard to tell, given not our own current concerns but the concerns of natural selection itself, whether or not an error is useful. Those more hopeful of continuing life than is justified by the evidence may live a few months longer. But Dawkins has claimed that "[a]s soon as a runt becomes so small and weak that his expectation of life is reduced to the point where benefit to him due to parental investment is less than half the benefit that the same investment could potentially confer on other babies, the runt should die gracefully and willingly. He can benefit his genes most by doing so" (Dawkins 1989, p.130). Something like this may also have been true, most places and times, for terminally ill adults being cared for by kin. People who are more confident that they can perform a certain task than justified by the evidence succeed more frequently as a result. Externally administered steroids – steroids above what the normal body usually manufactures – have a similar effect, but they are not good for you. Perhaps these people should be turning their attention to other activities just as rewarding but for them with a higher rate of success. The hypothesis that we have systems that (purposefully) override the normal belief-forming systems to create false beliefs that will motivate us more and make us more successful implies that our normal motivational systems are, for some reason, inadequately designed, hence need to be compensated for. Surely we should wonder what got in the way of better design for our motivational systems in the first place?

Are delusions biologically adaptive? Salvaging the doxastic shear pin

doi:10.1017/S0140525X09991464

Aaron L. Mishara[a,b] and Phil Corlett[a,c]

[a]*Department of Psychiatry, Brain Mapping Unit and Behavioural and Clinical Neurosciences Institute, University of Cambridge, School of Clinical Medicine, and Addenbrooke's Hospital, Cambridge CB2 2QQ, United Kingdom;* [b]*Department of Psychiatry, Clinical Neuroscience Research Unit, Yale University School of Medicine, Connecticut Mental Health Center, New Haven, CT 06519;* [c]*Department of Psychiatry, Abraham Ribicoff Research Facilities, Yale University School of Medicine, Connecticut Mental Health Center, New Haven, CT 06519.*

aaron.mishara@yale.edu
philip.corlett@yale.edu

Abstract: In their target article, McKay & Dennett (M&D) conclude that only "positive illusions" are adaptive misbeliefs. Relying on overly strict conceptual schisms (deficit vs. motivational, functional vs. organic, perception vs. belief), they prematurely discount delusions as *biologically* adaptive. In contrast to their view that "motivation" plays a psychological but not a biological function in a two-factor model of the forming and maintenance of delusions, we propose a *single* impairment in prediction-error–driven (i.e., motivational) learning in three stages in which delusions play a biologically adaptive role.

By concluding that only "positive illusions" are adaptive misbeliefs, McKay & Dennett (M&D) prematurely discount delusions as *biologically* adaptive. They do not pursue their argument that delusions may resemble the doxastic shear pin, which breaks down to preserve other "more expensive parts of the system" (sect. 10, para. 1). Therefore, they overlook the possibility that delusions may be tied to survival, and thus, by M&D's own criteria, natural selection.

In remaining alert to nourishment, danger, and reproductive opportunities, humans maintain relatively steady contact with their environment in an ongoing perception-action cycle (Fuster 2006; von Weizsäcker 1950) and attendant action-outcome learning (Dickinson & Shanks 1995). The neurobiological changes in early schizophrenia, however, disrupt the "binding" processes of perception, agency, and self (Haggard et al. 2002), reflected in the patient's experience (Mishara 2007b). How then are delusions adaptive by contributing to the patient's survival?

In making our case, we refer to an analysis of instrumental learning that cleaves the processing of actions into two distinct systems, one goal-directed, the other habitual (Daw et al. 2005). The goal-directed system involves learning flexible relationships between actions and outcomes instantiated in the more computationally intensive prefrontal cortices. On the other hand, habits involve more inflexible representations of the relations between environmental stimuli and behavioral actions; when a particular cue is perceived, a specific action is elicited irrespective of the consequences. These two systems compete to control behavior (Hitchcott et al. 2007). Delusions are subserved by the striatal habit system because the computationally intensive goal-directed system is impaired – an argument that we feel has consilience with the authors' view that misbeliefs qualify as biologically adaptive to the extent that they maintain the person's functioning while preventing further

damage. (We do not comment on the interesting debate whether delusions are beliefs. For one account, see Mishara, in press a).

M&D argue that brain dysfunction underlying delusions involves "breakage" in the belief evaluation system which is "adventitious, not designed" (sect. 10, para. 4). They advocate a two-factor model: (1) a perceptual insult which engenders odd experiences; (2) a deficit in belief evaluation which enables the entertainment and maintenance of bizarre and unlikely explanations for the experience (Davies & Coltheart 2000). Distinguishing deficit (organic-neural) versus motivational (psychological/psychodynamic-defense) approaches to delusions (McKay et al. 2007a), M&D conclude that "motivation" plays a psychological but not a biological role in the two-factor model. In contrast, we propose a single impairment in prediction-error–driven (i.e., motivational) learning in three stages: (1) delusional mood; (2) delusion as Aha-Erlebnis; (3) reconsolidation. Our model indicates how delusions may be adaptive as a shear pin function by enabling the patient to remain in *vital* connection with his/her environment:

1. Prior to delusions, a prodromal *delusional mood* may last for days, months, or even years (Conrad 1958; Jaspers 1946/1963). The patient experiences increasingly oppressive tension, *a feeling of non-finality* or expectation. Conrad (1958) calls this *Trema* (stage-fright) as the patient has the feeling that something very important is about to happen. Attention is drawn toward irrelevant stimuli, thoughts, and associative connections which are distressing and unpredictable (Kapur 2003; McGhie & Chapman 1961; Uhlhaas & Mishara 2007). This reflects an impairment in the brain's predictive learning mechanisms, such that unexpected events, prediction errors, are registered inappropriately (Corlett et al. 2007).

2. The delusions appear as an *Aha-Erlebnis*, or "revelation" (Conrad 1958), concerning what had been perplexing during delusional mood. In delusions of reference, harmless or accidental occurrences in the environment are taken as referring to the self. Conrad (1958) calls this a reflexive turning back on the self in which the universe is experienced as "revolving" around the self as middle-point. The delusions are not primarily a defensive reaction to protect the self, but involve a "reorganization" of the patient's experience to maintain behavioral interaction with the environment despite the underlying disruption to perceptual binding processes (Conrad 1958; Mishara 2010). At the Aha-moment, the "shear pin" breaks, or as Conrad puts it, the patient is unable to shift "reference-frame" to consider the experience from another perspective. The delusion disables flexible, controlled conscious processing from continuing to monitor the mounting distress of the wanton prediction error during delusional mood and thus deters cascading toxicity. At the same time, automatic habitual responses are preserved, possibly even enhanced (Corlett et al. 2009b).

3. *Reconsolidation*. Forming the delusion is associated with insight relief which stamps the delusion into memory (Miller 2008; Tsuang et al. 1988). Each time delusions are deployed, they are reinforced further, through a process of recall, reactivation, and reconsolidation, which strengthens them, conferring resistance to contradiction rather like the formation of motor-habits with overtraining (Adams & Dickinson 1981). When subsequent prediction errors occur, they are explicable in terms of the delusion and serve to reinforce it (Corlett et al. 2009b; Eisenhardt & Menzel 2007). Hence the paradoxical observation that challenging subjects delusions can actually strengthen their conviction (Milton et al. 1978). In each rehearsal of the delusion in the present instance, there is a "monotonous" spreading of the delusion to new experience (Binswanger 1965; Conrad 1958; Mishara, in press b) and, as such, it is both fixed and elastic (Corlett et al. 2009b). For example, we interviewed a middle-aged schizophrenia patient with the intractable erotomanic delusion that a college acquaintance had fallen in love with her and now controls parts of her life. Whenever she thinks of him, she hears a "car beep" or "trips while walking," i.e., signals intended to inform her that he knows she is thinking about him.

Neurobiologically, this reconsolidation-based-strengthening shifts control of behavior toward the striatal habit system. However, the ceding of behavior from effortful, conscious control is associated with a "mechanization" of experience. Schizophrenia patients delusionally refer to themselves in inhuman terms, for example as "machine," "computer," or "registering apparatus" (Binswanger 1965; Kraus 1997; Mishara 2007a), as if the delusion reflects its own disabling function of flexible conscious processing. Losing the experience as consistent intentional agent (Wegner 2004), the patients nevertheless continue to respond reflexively to the environmental cues incumbent upon them, necessary for continued survival. As complement to such delusions of alien control, however, the healthy individual has the converse "everyday delusion": She thinks that it is "I" who moves her own limbs. She calls the movement *mine* although it has its own momentum, automaticity, and finds its own way. That is, the healthy individual "overlooks" the impersonal-mechanical side of her movements in a "counter-delusion" to the patient who is unable to access the personal contribution (von Weizsäcker 1956). We are no more free from the necessity of "delusions" in our everyday functioning and its intermittent ceding to automatic processes than is the patient with schizophrenia.

Finally, the authors outline Bayesian mechanisms of rational belief formation. We propose that delusions form via the same Bayesian learning mechanisms but we challenge the strict separation between perception and belief upon which two-factor accounts are predicated (Corlett et al. 2009a; Fletcher & Frith 2009; Hemsley & Garety 1986; Uhlhaas & Mishara 2007). In our account, delusions also depend on aberrations of perception which occur when neuronal noise induces mismatches between expectancy (Bayesian priors) and experience (sensory inputs/evidence), but in terms of the single factor, prediction error.

ACKNOWLEDGMENTS

Both authors are recipients of the NARSAD Young Investigator Award. Phil Corlett is supported by the University of Cambridge Parke-Davis Exchange Fellowship in Biomedical Sciences.

The evolution of religious misbelief

doi:10.1017/S0140525X09991312

Ara Norenzayan, Azim F. Shariff, and Will M. Gervais

Department of Psychology, University of British Columbia, Vancouver, BC V6T 1Z4, Canada.

ara@psych.ubc.ca **www.psych.ubc.ca/~ara**
azim@psych.ubc.ca **www.psych.ubc.ca/~azim**
will.gervais@gmail.com

Abstract: Inducing religious thoughts increases prosocial behavior among strangers in anonymous contexts. These effects can be explained both by behavioral priming processes as well as by reputational mechanisms. We examine whether belief in moralizing supernatural agents supplies a case for what McKay & Dennett (M&D) call evolved misbelief, concluding that they might be more persuasively seen as an example of *culturally* evolved misbelief.

Is belief in supernatural agency an example of evolved "misbelief"? McKay & Dennett (M&D) consider recent psychological experiments that have investigated whether religious beliefs cause prosocial behavior such as generosity and honesty (for reviews, see Norenzayan & Shariff 2008; Shariff et al. 2010). In M&D's philosophical analysis, whether or not religion supplies a case of evolved misbelief turns out to depend on the psychological mechanism that best accounts for these effects. We therefore revisit the experimental evidence and discuss in some depth the ideomotor and supernatural watcher accounts for these effects.

M&D cite Randolph-Seng and Nielsen (2008), who critiqued Shariff and Norenzayan (2007), questioning the plausibility of the supernatural watcher hypothesis because the data could not conclusively distinguish between the ideomotor and supernatural watcher explanations. These two mechanisms gain plausibility given two distinct but well-supported empirical literatures. There is considerable evidence showing that prosocial behavior can be facilitated both by activating nonconscious altruistic thoughts (e.g., Bargh et al. 2001), and by heightened reputational concerns (e.g., Fehr & Fischbacher 2003). These two mechanisms are not mutually exclusive, however, and may even reinforce each other in everyday life.

The interesting question therefore is: What kind of laboratory evidence can provide support for the supernatural watcher account above and beyond behavioral-priming processes? First, if the priming effects of God concepts are weaker or nonexistent for non-believers, then the effect could not be solely due to ideomotor processes, which are typically impervious to prior explicit beliefs or attitudes. Second, if God primes make religious participants attribute actions to an external source of agency, these effects could not be explained by ideomotor processes, as such manipulations disambiguate the felt presence of supernatural watchers from their alleged prosocial consequences. Finally, if the supernatural watcher explanation is at play, religious primes should arouse social evaluation of the self. Moreover, such reputational awareness should moderate the magnitude of the prime's effect on prosocial behavior.

As M&D note, evidence on the first point is currently mixed. However, close examination of the findings betrays a revealing pattern. All but one of these priming studies recruited student samples, which can be problematic since beliefs, attitudes, and social identity among students can be unstable, raising questions about the reliability of chronic individual difference measures of religious belief and identity measures for students who are still in transition to adulthood (Sears 1986; Henrich et al., in press). Thus, student atheists might be at best "soft atheists." In the only religious priming experiment we are aware of that recruited a non-student adult sample (Shariff & Norenzayan 2007, Study 2), the effect of the prime emerged again for theists, but disappeared for these "hard" atheists (see Fig. 1). In addition, Henrich et al. (2009) found that across 14 small-scale societies of varying group size, where there is variability in whether supernatural agents are morally concerned, belief in the moralizing Abrahamic God (along with degree of market integration) predicted larger offers in the dictator and ultimatum games. These initial findings speak against an exclusively ideomotor account of the results, and suggest that belief – not just alief – is involved in religious prosociality.

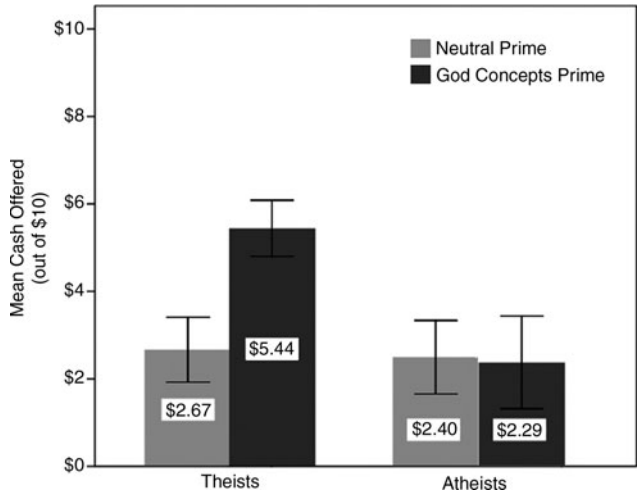

Figure 1 (Norenzayan et al.). Results from the dictator game in Shariff and Norenzayan (2007, Study 2) indicate that priming God concepts increased generosity for religious believers but not for atheists. Error bars represent standard error of the mean.

Regarding the second question, one experiment clearly separates the felt presence of a supernatural agent from prosocial outcomes. Dijksterhuis et al. (2008) found that after being subliminally primed with the word "God," believers (but not atheists) were more likely to ascribe an outcome to an external source of agency, rather than their own actions. In addition, religious belief positively correlates with greater concern with social evaluation of the self (Trimble 1997), and recent experimental evidence points to this being a causal relationship. Gervais and Norenzayan (2009) found that priming God concepts (using the same sentence unscrambling task of Shariff and Norenzayan [2007]) increased public self-awareness (Govern & Marsch 2001) – a measure that taps into feelings of being the target of social evaluation. In contrast, and as predicted, the prime had no effect on private self-awareness. Ongoing research is examining whether prosocial effects of religious primes are moderated by measures of evaluative concern, a key prediction of the supernatural watcher hypothesis, which would be incompatible with a purely ideomotor account. Thus, although M&D are right that more research is needed to reach firm conclusions, the evidence regarding the supernatural watcher hypothesis is more compelling than M&D's cautious approach suggests. But does that mean that belief in supernatural agents is an example of adaptive misbelief?

M&D briefly mention both by-product theories of religion and cultural evolutionary explanations for cooperation. We have argued elsewhere (Norenzayan & Shariff 2008; Norenzayan, in press; Shariff et al. 2010) that integrating these two frameworks yields a more cogent explanation for the rise and persistence of religious beliefs than theories which invoke a more direct genetic evolutionary argument (e.g., Bering et al. 2005; Johnson & Bering 2006). Once belief in supernatural agency emerged as a by-product of mundane cognitive processes, cultural evolution favored the spread of a special type of supernatural agent – moralizing high Gods. Growing evidence is converging on the conclusion that sincere belief in these omniscient supernatural watchers facilitated cooperation and trust among strangers (Norenzayan & Shariff 2008). Not surprisingly, this cultural spread coincided with the expansion of human cooperation into ever larger groups over the last 15 millenia (Cauvin 2000). This evolutionary scenario has the virtue of explaining an otherwise puzzling feature of religious prosociality – namely, the systematic cultural variability in the prevalence of moralizing Gods across societies that correlates with group size (e.g., Roes & Raymond 2003). Contrary to a genetic adaptation account, the deities of most small-scale societies, which more closely approximate ancestral conditions, are neither fully omniscient nor morally concerned. It is the evolutionarily recent anonymous social groups, facing the breakdown of reputational and kin selection mechanisms for cooperation, which most strongly espouse belief in such Gods. Thus, beliefs in moralizing supernatural agents may not qualify as genetically evolved misbeliefs. But they could instead be seen as examples of culturally evolved ones that played a key historical (although not irreplaceable) role in the rise and stability of large cooperative communities.

The (mis)management of agency: Conscious belief and nonconscious self-control

doi:10.1017/S0140525X09991336

Brandon Randolph-Seng

Texas Tech University, Rawls College of Business, Area of Management, Lubbock, TX 79409-2101.

b.randolph-seng@ttu.edu www.webpages.ttu.edu/brandolp

Abstract: McKay & Dennett (M&D) identify positive illusions as fulfilling the criteria for an adaptive misbelief, but could there be other

types of beliefs that may qualify as adaptive misbeliefs? My commentary addresses this and other questions through identifying belief in free will as a potential candidate as an adaptive misbelief.

To say that beliefs are untrue, or are "misbeliefs," is to say that the belief in question can be objectively verified; however, beliefs are by definition subjective. Despite the subjective nature of human beliefs, researchers have learned to use the refining process of the scientific method in order to partly uncover "reality." Such is the plight of the social scientist and the beauty of scientific method. At what point, however, do the systematic replications and validations go beyond an understanding of the underlying factors giving rise to beliefs? If science, for example, reliably demonstrates that the sun does not in reality move across the sky, does this change one's belief in how the sun is moving throughout the day? It may do so (based on a faith in science), but such change in belief does nothing to change perception. One's subjective perception of the sun's movement is what is real and is the only view that matters to normal functioning. Therefore, there are times when it may be more adaptive and functional to misperceive than correctly believe.

McKay & Dennet (M&D) suggest that in order to identify a systematically adaptive misbelief, such belief may result from processing "biases" in the sensory system itself. Such an assertion assumes that perception gives rise to (mis)belief. Although this assertion may at times be true, at other times perception and belief may be disconnected (e.g., see above), or (mis)belief may actually give rise to perception (e.g., New Look: Balcetis & Dunning 2006; Bruner 1957). In fairness, disconnects between perception and belief are discussed in M&D's discussion of "alief" and error management theory; nevertheless, their proposed connection between perception and misbelief remains unclear.

A similar point of confusion is found in the distinction M&D make between psychological and biological adaption. It has long been recognized that social psychological factors and biological factors are closely related (for a review see Cacioppo et al. 2000). Because of the complementary nature of social psychological and biological factors in human behavior, making distinctions between the two becomes meaningless without specifying how the two may be connected for a given outcome (e.g., Gailliot et al. 2007).

The lack of clarity in distinguishing between perception and belief, and psychological and biological adaption, becomes apparent as M&D investigate whether religious belief might be a good candidate for adaptive misbelief. If inferring the presence of agents is adaptive, for example, is such adaptability psychological or biological? If such agents cannot be seen and are not real, then what role does perception play? In fact, most of the empirical research in the area cited by M&D (i.e., Pichon et al. 2007; Randolph-Seng & Nielsen 2007; Shariff & Norenzayan 2007) suggest that *alief*, rather than belief, is involved, and that perceptual priming of agency rather than religion per se may be at the root of the behavioral effects found. M&D do acknowledge these possibilities, but fail to extend these possibilities to the search for adaptive misbelief (a point I return to later).

Where M&D's search does take them is to conscious self-deception. Considering their previously implied condition of adaptive misbelief arising from nonconscious perceptual biases, it is unclear why conscious self-deception is even considered, but it does provide a nice segue into positive illusions (i.e., the unrealistic optimism-type), fulfilling the stated requirements for adaptive misbelief. But why stop there? Are there other types of beliefs that could be considered adaptive misbeliefs (e.g., antecedent misbeliefs giving rise to the positive illusions described by M&D)?

One potential candidate may be gleaned from M&D's discussion of religious beliefs, namely a belief in personal agency. Recent social psychological research suggests that one's belief in free will is (or can be) an illusion (for reviews, see Bargh 2008; Wegner 2005); however, other research suggests that the more one believes in personal agency, the more prosocial and hardworking one tends to be (Baumeister et al. 2009; Stillman et al., in press; Vohs & Schooler 2008). Insofar as a belief in free will is a positive illusion (see Bargh & Earp 2009), it may be considered a pre-existing belief for the types of positive illusions discussed by M&D. For example, if one did not believe that control over one's actions was real, then there may be less reason to believe that one has what it takes to survive a life-threatening disease.

Proposing a belief in free will as a candidate for adaptive misbelief does bring to the forefront the previously discussed question of the connection between perception and belief. Insofar as the choices one makes are the result of nonconscious thinking instigated by the environment and resulting from our evolutionary past (Bargh 2008), belief in free will and nonconscious perception do not line up. Nevertheless, rational choice and conscious self-regulation are also thought to be intricately linked to the evolution of human cognition (Baumeister 2008), in which case belief may be able to feed back into perception. Such a possibility would help explain how humans, despite being mostly unaware of the various messages presented to them from the environment, can successfully navigate through their environment in order to accomplish their personally activated goals. In fact, recent research has found that people can go beyond nonconsciously regulating their responses to *consciously* perceived stimuli, to preconsciously controlling the impact of *nonconsciously* perceived stimuli on their responses (i.e., being differently influenced by subliminal primes depending on current nonconscious motivations; Randolph-Seng 2009). In this way, an adaptive misbelief, such as a belief in free will, may actually become true as human cognition evolves.

You can't always get what you want: Evolution and true beliefs

doi:10.1017/S0140525X09991476

Jeffrey P. Schloss[a] and Michael J. Murray[b]

[a]*Department of Biology, Westmont College, Santa Barbara, CA 93108;* [b]*Department of Philosophy, Franklin and Marshall College, Lancaster, PA 17604.*

schloss@westmont.edu mmurray@fandm.edu

http://www.fandm.edu/x11310?id=204

Abstract: McKay & Dennett (M&D) convincingly argue against many proposals for adaptively functioning misbelief, but the conclusion that true beliefs are generally adaptive does not follow. Adaptive misbeliefs may be few in kind but many in number; maladaptive misbeliefs may routinely elude selective pruning; reproductively neutral misbeliefs may abound; and adaptively grounded beliefs may reliably covary with but not truthfully represent reality.

In critiquing proposed examples of adaptive misbelief, McKay & Dennett (M&D) aim to confirm the assumption that, via evolution, humans "have been biologically engineered to form true beliefs" (sect. 1, para. 2), and conclude that the exchange rate between fitness and truth "is likely to be fair in most circumstances" (sect. 15, last para). We agree with their critiques (Murray & Moore 2009; Schloss 2007), but the conclusion does not follow.

First, even if many proposals for adaptive misbelief fail, this does not tell us whether adaptive misbeliefs spawned in situations M&D acknowledge as credible are common or, as they claim, limited to "certain rarefied contexts" (sect. 15, last para). For instance, the promiscuous attribution of agency and teleology, or the manifold positive illusions that may be accounted for within error management theory (Johnson 2009) are extraordinarily plentiful, persistent, and influential. Other kinds of positive illusions, from the placebo effect to magnifying the virtues of beloved people, places, nations, and traditions – consistent with proposals

for resource commitment strategies – may be even more plentiful and powerful. And contrary to M&D's tempting suggestion, such positive illusions are not restricted to subjective beliefs that are "not likely to be rudely contradicted by experience" (sect. 14, para. 3). Many of these widespread beliefs entail almost delusional denials of repeated experience. Notions that Eros lasts forever, this time it's real, and (as the sappy song says) "When we're hungry, love will keep us alive" are effective and virtually ubiquitous catalysts for reproductive pairbonding. But by non-reproductive periods of the human life cycle, those Romeos whose romantic illusions have not killed them, have oft' yielded to the wisdom of Friar Lawrence: "These violent delights have violent ends, and in their triumph die...Therefore love moderatedly: long love doth so. Too swift arrives as tardy as too slow." Yet another category altogether, unexamined by M&D, is selection for cognitive extravagance independent of problem-solving utility (Miller 2000; 2001). But even granting M&D's conclusion that there are just a few families of adaptive misbelief, we don't yet know enough about their natural history to determine how many species there are or what their carrying capacities and competitive coefficients are relative to true beliefs.

Second, even if reproductively beneficent misbeliefs are rare and most misbeliefs have costs, this does nothing to tell us how well evolution ultimately avoids such costs. Indeed, M&D elegantly acknowledge that functional normativity does not entail statistical normality: In evolution, forgivable malfunctions may be common and achieving proper function may be "positively rare" (sect. 3, para. 5). Thus, even if truth is the evolutionary target as M&D maintain, design constraints, by-product associations, and historical contingencies may make it one that cognition has a low probability of hitting.

Third, many kinds of beliefs – from debates over quantum theory to discussions of metaphysics – have no clear reproductive relevance at all. How, and whether, such beliefs are related to cognitive mechanisms that have been selected for veracity is uncertain (Cromer 1993; Wolpert 2000). What does not seem uncertain is that manifold beliefs do not influence behaviors or the behaviors they do influence are not reproductively salient. Belief-forming mechanisms generate variety that, analogous to neutral polymorphisms (Kimura 1991), may be unpruned by the adaptive consequences of their truth or falsity. Indeed, the capacity for some degree of cognitive licentiousness may itself be an adaptation to the "uncertain futures problem" (Plotkin 1997; Wagner 2005).

Finally, M&D's conclusion requires the falsity not only of the above ways in which selection fails to exclude misbelief, but also of the more global but controversial thesis that nothing at all about the process of natural selection serves to favor truth-conducive cognitive tools (Churchland 1987; Plantinga 2002; Stich 1990).

On selectionist accounts of the origin of mind, beliefs and belief forming mechanisms are selected by virtue of their capacity to support adaptive behavior or internal states. Thus, belief forming mechanisms will be selected when they yield (i) a representational model that orients organisms towards adaptive behaviors, and/or (ii) a correlational source of arousal or inhibition that serves to motivate adaptive (or inhibit maladaptive) behavior. The question then becomes: Are models that are true better at orienting organisms towards adaptive behaviors, or are true beliefs better at arousing effective desires for adaptive behaviors? From what we know about the action of natural selection, the most prudent answer may be: "There is no reason to think so."

Why is there no reason to think so? Because (in science, and in belief generally) models need only to "save appearances" in order to be successful. Consider the task of designing "thinking" robots for a competition in which the winners were duplicated (with minor program variations) for future competitions. While one would surely seek to program competing robots to form beliefs that provided an isomorphic "map" of the external environment, would one further seek to program beliefs about that environment that were true? Not obviously. Indeed, there are numerous ways of programming the robot to "conceptualize" its environment that, while representationally biased or even radically false, are nonetheless (a) appropriately isomorphic and (b) reliably adaptive behavior-inducing. Such programs would be adaptive.

What is true of programmed learning robots is true of selection-designed cognition. Dennett has aptly commented, "Lying behind, and distinct from, our reasons are evolutionary reasons, free-floating rationales that have been endorsed by natural selection" (Dennett 2006a, p. 93). Our reasons (in better moments) are truth-seeking; natural selection's are fitness seeking. We cannot know if, in achieving its reasons, selection allows us also to achieve ours.

Of course, one might respond that just because our belief-forming mechanisms are liable to error in these domains does not mean that they are routinely or irremediably unreliable (after all, we often discover our errors, like the cognitive biases mentioned above). But this offers little reassurance, since the seeming discovery of error relies on comparing beliefs to other beliefs which, for all we know, are comparably unreliable, though perhaps for different reasons.

Richard Dawkins has commented that "however many ways there are of being alive, it is certain that there are vastly more ways of being dead" (Dawkins 1996, p. 9). The same is true of being right and wrong. Natural selection is immensely effective at weeding out ways of not being alive. It is unclear how well it fares in culling ways of not believing truly.

Culturally transmitted misbeliefs

doi:10.1017/S0140525X09991348

Dan Sperber

Institut Jean Nicod, ENS, 75005 Paris, France.

dan.sperber@gmail.com www.dan.sperber.fr

Abstract: Most human beliefs are acquired through communication, and so are most misbeliefs. Just like the misbeliefs discussed by McKay & Dennett (M&D), culturally transmitted misbeliefs tend to result from limitations rather than malfunctions of the mechanisms that produce them, and few if any can be argued to be adaptations. However, the mechanisms involved, the contents, and the hypothetical adaptive value tend to be specific to the cultural case.

Most of humans' beliefs, or at least most of their general beliefs, are acquired through communication. I owe my beliefs that I was born in Cagnes-sur-mer, that Washington is the capital of the US, that mercury is a metal, that dodos are extinct, that stagflation is bad, and so on ad indefinitum, not to my own perceptions and inferences on those matters, but to the words of others. Are these beliefs "grounded" in McKay & Dennett's (M&D's) sense, that is, "appropriately founded on evidence and existing beliefs" (target article, sect. 1, para. 2)? Not on relevant evidence and beliefs available to me. I hold these beliefs because I trust their sources (or, anyhow, trusted them at the time I formed the beliefs). My trusting of sources may itself be founded on appropriate evidence of their trustworthiness, but quite often it is founded rather on my trust of yet other sources that have vouched for them; for instance, I trusted the textbooks I read because I trusted the teachers who vouched for them, and I trusted the teachers because I trusted my parents who vouched for them. Needless to say, the authors of the textbooks themselves were just reporting information from yet other sources.

Of course, however long the transmission chain, communicated beliefs may be vicariously grounded in appropriate

evidence and background beliefs that had been available to the initial communicators. Nevertheless, long chains of transmission carry serious epistemic risks of two kinds. First, judgments of trustworthiness are less than 100% reliable, so that, generally speaking, the longer the chain, the lesser its compounded reliability (and this even if, serendipitously, the initial source of the transmitted belief happens to be have been trustworthy). Second, information is typically transformed in the process of transmission. As a result, a belief at the end of the chain is quite often different in content from the one at the beginning and therefore cannot vicariously benefit from initial grounding. This is particularly true of orally transmitted cultural beliefs, notably religious beliefs of the kind studied by anthropologists. One generation's religious beliefs may undergo changes in its lifetime and anyhow is a transformation of the beliefs of the previous generation. There is no initial religious belief at the dawn of time, but rather, an increasing – and sometimes decreasing – religious tenor in a variety of beliefs; later beliefs are not copies of earlier ones.

The absence of appropriate grounding not just of religious beliefs, but of so many others cultural beliefs concerning, for example, food, health, or the moral traits of ethnic groups, means that human population are inhabited by a host of poorly grounded or ungrounded beliefs. Most of these are, in the terms of M&D, misbeliefs. In fact, most of our misbeliefs are culturally transmitted misbeliefs rather than individual mistakes, distortions, or delusions.

Does this mean that the social and cognitive mechanisms through which we come to hold cultural misbeliefs are malfunctioning? Are humans irrationally gullible? No, the prevalence of cultural misbeliefs is compatible with the view that the mental mechanisms involved in epistemic trust (Origgi 2004) and epistemic vigilance (Mascaro & Sperber 2009; Sperber et al., forthcoming) are calibrated to filter information in interpersonal communication, if not optimally, at least reasonably well. They do, however, create a susceptibility to misinformation that originated not in one's direct interlocutors but long before in extended chains of transmission. This vulnerability is enhanced when it is well beyond the individual's competence to assess the truth or at least the plausibility of the contents transmitted. This is particularly the case when the contents in questions are too obscure to be open to epistemic assessment.

In the process of cultural transmission and transformation, beliefs may lose not only their empirical grounding but also their epistemic evaluability. For a belief to be evaluable, it must have a propositional content, that is, be true-or-false. One may relax the criterion so as to take into account the fact that many, possibly most, of our beliefs are not sharply propositional and may, in a range of limiting cases, lack a truth value. Still, for beliefs to be informative and guide action, they had better, in most ordinary situations, be such that their relevant consequences, practical consequences in particular, can be inferred. Many culturally transmitted beliefs do not satisfy this criterion. Their content is not just vague; it is mysterious to the believers themselves and open to an endless variety of exegeses. These are what I have called semi-propositional or half-understood beliefs (Sperber 1982; 1997). The paradigmatic example of a semi-propositional belief is the dogma of the Holy Trinity, which the believers themselves insist is mysterious. Of course, philosophers who define a belief as an attitude *towards a proposition* may dispute that "semi-propositional beliefs" are beliefs at all. But from a cognitive and social science point of view, a definition of *belief* that excludes most religious beliefs renders itself irrelevant. In particular, it disposes by definitional fiat of a wide class of cultural beliefs of which it can be disputed whether they are false or lack truth value, but that are definitely not true and hence are misbeliefs (even religious believers would accept this of religious beliefs other than their own, i.e., of the vast majority of religious beliefs).

I have long argued that cultural misbeliefs occur and propagate as a by-product, a side-effect of our cognitive and communicative dispositions (Sperber 1985; 1990). Still, it could be that some of these misbeliefs or some classes of them contribute to the reproductive success of their carriers in a manner that indirectly contributes to their own propagation. One possible class of such adaptive cultural misbeliefs would be beliefs the expression of which contributes to group identities and solidarities that enhance the individual's fitness. Unlike the positive individual illusions discussed by M&D, the adaptiveness of such beliefs does not come from the manner in which their content guides the believers' actions. It is not the content of the beliefs that matters; it is who you share them with. Yet not just any content is equally appropriate to serve such an adaptive role. In particular, a content unproblematically open to epistemic evaluation might either raise objections within the relevant social group, or, on the contrary, be too easily shared beyond that group. So, semi-propositional contents are ceteris paribus better contents for beliefs the adaptive value of which has to do with cultural sharedness, not because these contents contribute to this adaptive value by guiding action, but because they do not stand in the way of acceptance by the relevant group. Their content may also have features that contribute positively to their cultural success, for instance by rendering them more memorable, but this is another story (see, e.g., Atran & Norenzayan 2004; Boyer 1994; Sperber 1985).

Adaptive misbeliefs and false memories

doi:10.1017/S0140525X09991488

John Sutton

Macquarie Centre for Cognitive Science, Macquarie University, Sydney, NSW 2109, Australia.

jsutton@maccs.mq.edu.au

http://www.phil.mq.edu.au/staff/jsutton

Abstract: McKay & Dennett (M&D) suggest that some positive illusions are adaptive. But there is a bidirectional link between memory and positive illusions: Biased autobiographical memories filter incoming information, and self-enhancing information is preferentially attended and used to update memory. Extending M&D's approach, I ask if certain false memories might be adaptive, defending a broad view of the psychosocial functions of remembering.

Positive illusions, including those that "propel adaptive actions" (target article, sect. 13, para. 6) are maintained over time even (within limits) in the face of recalcitrant evidence. So they require sophisticated intertemporal accounting: Memory and associated forms of mental time travel must be enlisted if positive illusions are to be stable enough to enhance fitness, to be "pervasive, enduring, and systematic" rather than mere temporary errors (Taylor & Brown 1988, p. 194). So if McKay & Dennett (M&D) are right that certain kinds of ungrounded belief are adaptive, theories of memory are directly implicated. This link extends M&D's account of adaptive misbeliefs, suggesting new questions for memory research.

The sparse literature on functional analyses of remembering addresses the adaptive nature of forgetting and the puzzling luxury of autobiographical memory (Bjork & Bjork 1988; Boyer 2008a; 2009; Glenberg 1997; Nairne 2005; Nairne et al. 2007; Schacter 2001). But the possibility that false memories (or ungrounded memories, which often contingently turn out false) could themselves be adaptive is surprising. False memories are usually seen as unfortunate outcomes of the constructive nature of remembering (Bernstein & Loftus 2009, p. 373), just as the manipulability of general belief-fixation is seen as epistemological trouble. But this standard line of thought is too quick, on two counts: reconstruction is not itself always distortion

(Barnier et al. 2008; Sutton 2009), and, as M&D suggest, falsity need not always be maladaptive.

Biased contextual and autobiographical memories filter incoming information, and, in turn, self-enhancing (or otherwise illusory) information is preferentially attended and used to update memory. This bidirectional link between memory and positive illusions lies at the heart, for example, of temporal self-appraisal theory (Wilson & Ross 2003). Memory directly supports positive illusions: When unrealistic inflation of *current* self-image is difficult, people still "selectively recall and reconstruct evidence from the past that makes them feel good about their current selves" (Wilson & Ross 2001, p. 582). Conversely, the motivation to maintain positive self-regard in the present motivates misplaced or false memories, for example, as we subjectively move favourable past events forwards in time and unfavourable past events backward (Ross & Wilson 2002, p. 800). These loops between autobiography and the control of action, the self remembered and the working self, exhibit considerable variation (Conway 2005). Some theorists stress strong reflexive feedback from self-representation into behavior, with ongoing integration lived out between actions and self-ascribed character, emotions, memories, and plans (Velleman 2006). Others note gulfs between the story and the life, seeing narrative self-descriptions more like public relations reports that float free of the causal processes behind the government's or organization's behavior (Clark 1994; Dennett 1991): This need not be morally or psychologically suspect, due to deliberate suppression or self-deceit, for the narrative and memory capacities – just like PR spokespersons – often don't have nor need the knowledge, or the contacts, or the access, either to get the back-story right or to directly feed in to future choices and actions.

Could false memories, as M&D might suggest, in some circumstances enhance affective, cognitive, and social well-being? A broad positivity bias in autobiographical memory is linked to enhanced emotion regulation: the tendency to accentuate the positive in recalling experiences from the personal past drives improvements in mood (Mather & Carstensen 2005). Such memory biases are most securely demonstrated in older adults, whose positive affect is increased when reminiscing about positive past experiences (Pasupathi & Carstensen 2003); but younger adults get the same emotionally enhancing effects of biased autobiographical remembering when they remember while focussing specifically on their emotional state rather than on accuracy (Kennedy et al. 2004). That this positivity bias drives false memories, not just general selectivity in recall, is suggested by a recent study in which older adults recall more false *positive* memories than false negative memories across a range of stimuli (Fernandes et al. 2008).

Narcissistic biases are thus not found only in subpersonal sensory systems. Just as the "narcissistic encoding" of sensory information is driven by our "sensory-motor projects" rather than by any ontological project of correctly cataloguing the world (Akins 1996, p. 370), so the range of past experiences to which we are maximally sensitive in remembering are those relevant to embedded action, or salient for current and future decision-making. The gulf between design for fitness maximization and design for truth preservation (target article, sect. 9) may be widest in relation to truths about the past, especially about the distant past and about particular past events. As Boyer notes, the human "capacity to store unique episodes" is strange when "organisms learn about the past mostly to the extent that they can extract from past situations what is *not* unique about them, and what will be relevant in the future" (Boyer 2009, p. 4). So, given the likely recent emergence of mental time travel, its biological function may involve just the broader self-related and social functions at work in the positive illusions described by M&D. In remembering, correspondence with reality is often trumped by coherence, truth by psychosocial utility (Alea & Bluck 2003; Conway & Pleydell-Pearce 2000; Wilson & Ross 2003). Evolved tendencies for forming memories which lack perfectly secure grounds are useful computational short-cuts, minimizing the heavy cognitive costs of precise source monitoring, of tracking the diverse origins of our representations of what is not present in the immediate environment. The idea parallels the suggestion that forgetting is an adaptive response to the computational challenges of retrieval access in context-sensitive systems (Anderson & Schooler 2000; Michaelian, submitted). But M&D's approach suggests the extra thought that sometimes-false memories can themselves, given the structure of the social and natural environment, be *intelligent* errors. Not only are the costs of trusting and acting on them in general less than those of strenuously applying forensic standards of warrant and justification, but (further) misremembering things in particular positive ways might have direct personal, motivational, and social benefits. Neither social nor motivational influence on memory is intrinsically malign: M&D remind memory researchers that cataloguing our susceptibility to misinformation in non-standard, experimentally skewed environments may blind us to the richer functions of remembering.

Effective untestability and bounded rationality help in seeing religion as adaptive misbelief

doi:10.1017/S0140525X0999135X

Konrad Talmont-Kaminski

Institute of Philosophy. Marie Curie-Sklodowska University, 20-031 Lublin, Poland.

ktalmont@bacon.umcs.lublin.pl

http://bacon.umcs.lublin.pl/~ktalmont

Abstract: McKay & Dennett (M&D) look for adaptive misbeliefs that result from the normal, though fallible, functioning of human cognition. Their account can be substantially improved by the addition of two elements: (1) significance of a belief's testability for its functionality, and (2) an account of reason appropriate to understanding systemic misbelief. Together, these points show why religion probably is an adaptive misbelief.

McKay & Dennett (M&D) think that systematically adaptive misbeliefs are unstable – that over time they are revealed to be false, much as in the case of the boy who cried wolf. This need not be the case. Systematic falsehood can only be discounted or accommodated if its truth can be investigated effectively. Some claims are particularly difficult to investigate in some contexts, however. The reasons are threefold (explored in Talmont-Kaminski 2009). First, their content may be such as to impede investigation. Supernatural claims, for example, typically involve entities that are invisible, shy, or just far, far away. Second, there may be social taboos against investigating such claims. Durkheim's (1912) category of the sacred singles out precisely this kind of social barrier against investigation. Third, which claims can be investigated depends upon which scientific tools are available. Opposition to scientific development ensures that some misbeliefs remain uninvestigated. If people think the wolf is invisible, they won't be surprised when they don't see it. Similarly, if looking for it would break some social mores or would, for example, require heat-sensing cameras they cannot access, people would probably be held back from checking the boy's claims.

Without the possibility of investigation, beliefs become untethered from their truth-value so their popularity and stability are free to be determined by the idiosyncrasies of the human belief-forming system (BFS) as well as any functions that they might have – it is the *untestable* beliefs that can be most readily moulded to best serve their function. The problem, therefore, is not belief that runs counter to evidence but belief without

evidence. Such ungrounded belief is deemed by M&D to fall outside their compass, but their reasons for claiming this are faulty. The first problem is that M&D conflate grounded beliefs with beliefs produced by a BFS that normally produces grounded beliefs. Yet, a BFS that normally produces grounded beliefs may (and in the human case, does) occasionally generate, as a by-product, beliefs that are not properly grounded. The second, and connected, problem is that M&D take Bayes' Theorem to establish what it means for a belief to be grounded. When I have unlimited time and resources available to update beliefs, I'll be sure to do so. In the meanwhile, it is more realistic to use the same standard as that used by Gigerenzer (e.g., Gigerenzer et al. 1999), to whom M&D repeatedly refer. Bounded rationality theory (Simon 1996; Wimsatt 2007) sets a realistic standard because it works with the real limitations of the BFS instead of abstracting away from them. Thanks to that, it is also able to explain why the BFS leads to systematic misbelief, that is, systematic bias – such misbelief being the necessary cost of using cognitive heuristics that are fast enough and frugal enough for humans. Indeed, the theory is robust enough to link particular heuristics with particular kinds of systematic misbelief – this is what Rozin & Nemeroff (2002) does in formulating the contagion heuristic that is responsible for feelings of disgust associated with clean objects. (M&D interpret Rozin in terms of alief instead of belief, but the distinction is not relevant to this point nor does it seem strong enough to carry the weight they place upon it.) Bayes' Theorem cannot cope with any of this except to make post hoc allowances for it.

We now have almost all the tools to see why, despite M&D's less than clear stance, religion provides us with the paradigmatic case of adaptive misbelief (full account to be presented in Talmont-Kaminski, in preparation). Supernatural beliefs, in general, are particularly hard to investigate. Ghosts, fairies, or even just luck are not the kinds of things that would be easy to "catch in the act." The sacred status that such beliefs often have, as well as the opposition to scientific development that holding them often appears to be connected to, further impede investigation, rendering them effectively untestable. The cognitive science of religion, as developed by Atran (2004), Boyer (2001) and others, provides us with a plethora of examples such as Barrett's (2000) hyperactive agent detection device that can be best understood as cognitive heuristics that lead to supernatural beliefs as a by-product. Significantly, this by-product approach seems to explain all manner of supernatural beliefs, including superstitions, without necessarily explaining the specificity of religious beliefs. Life has a habit of putting by-products to work, however.

According to the account of religion put forward by D. S. Wilson (2002), hardly mentioned by M&D even though it does present religion as adaptive, the job of religious beliefs is to improve group cohesion, thereby allowing religious groups to compete more successfully against other groups. However, just as the by-product account was unable to distinguish religion from superstition, so Wilson is unable to distinguish religion from other group-oriented ideologies. Both problems disappear when we consider religion as a cultural phenomenon that exapts the existing cognitive by-products – an ideology that puts superstitions and the cognitive mechanisms underlying them to work. The result is beliefs that, thanks to their content as well as their social and scientific contexts, are well guarded against falsification and, therefore, relatively free to change as their functions require; that, thanks to the bounded and systematically biased nature of human reason, seem highly plausible; and that motivate people to act in ways adaptive for the group as well as, in the best of times, even for its members – an example of just the kind of gene-culture co-evolution that M&D are looking for.

ACKNOWLEDGMENTS

The research behind this commentary was carried out mostly while I was on a fellowship at the Konrad Lorenz Institute for Evolution and Cognition Research. Without my colleagues there it would not have been possible.

Belief in evolved belief systems: Artifact of a limited evolutionary model?

doi:10.1017/S0140525X09991361

Tyler J. Wereha and Timothy P. Racine

Department of Psychology, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada.

tjw4@sfu.ca tracine@sfu.ca

Abstract: Belief in evolved belief systems stems from using a population-genetic model of evolution that misconstrues the developmental relationship between genes and behaviour, confuses notions of "adapted" and "adaptive," and ignores the fundamental role of language in the development of human beliefs. We suggest that theories about the evolution of belief would be better grounded in a developmental model of evolution.

McKay & Dennett (M&D) present a clever, substantive, and thought-provoking analysis of the evolution of mistaken beliefs. Unfortunately, we feel it falls short because it is based on a particular model of evolution that has limited applications for understanding the development of beliefs. Given that M&D are not estimating proportions of phenotypic variance attributable to genes versus environments and are rather trying to account for the origin of a system, the population genetic framework upon which they rely is not helpful. Because the population-genetic ("pop-gen") framework does not speak to the development of organisms, it misconstrues the relationship between genes and behaviour (i.e., posits a passive conception of development), views beliefs as adaptations rather than as products of an organism's adaptability, and is blind to important epigenetic resources – most importantly in the case of the human belief system, language. We now expand upon these problems, and outline an alternative evolutionary developmental ("evo-devo") perspective which would better ground evolutionary claims about beliefs.

We argue that belief in evolutionarily engineered, truth-aiming belief systems is a product of an inadequate pop-gen model of evolution that essentially ignores developmental processes (Callebaut et al. 2007; Racine et al., forthcoming; Roberts 2002; Wereha & Racine 2009). Stated simply, even if one considers that we have evolutionarily derived belief systems, these systems still have to emerge in ontogeny and are not fundamentally separate from developmental processes. Such an uncontroversial statement belies the different conceptualizations of development derived from the pop-gen and evo-devo models of evolution. The overextension of the pop-gen model into the organismic level of analysis leads to a passive characterization of development. That is, development is viewed as directed by a robust genetic system that uses the most minimal "environmental" input as a trigger. Therefore, M&D forgo a developmental analysis of belief when they describe the features of an "evolutionary belief system" by reverse-engineering the beliefs of adult humans. From an evo-devo perspective, collapsing complicated developmental processes to a genetic component terribly misconstrues the nature of development. Such a commitment to a mechanistic, genocentric view of evolution and a passive conception of development is inconsistent with a developmental perspective that emphasizes the ubiquity of multiple levels of influence beyond simply the genetic level (e.g., Gottlieb & Lickliter 2007). We now elaborate on how a narrow genocentric, non-developmental view has little use for understanding the formation of beliefs in humans, and demonstrate how radically different a developmental approach conceptualizes

the issue. This difference is made explicit in describing the importance of "adaptive" and "adapted" in these two models.

In the context of their paper and in light of how these terms are used within the pop-gen model of evolution, it is understandable that M&D conflate the terms "adapted" and "adaptive" (see the target article's Note 3). We argue, however, that these two terms can actually characterize the difference between pop-gen and the developmental models of evolution. What are taken to be evolved belief systems or adaptations in the pop-gen model (genocentric view) are actually the products of organismic adaptability in an evo-devo perspective (organismic view).

In the pop-gen model, evolved belief systems are adaptations. Beliefs are "tools" derived from belief systems that were "engineered" by evolution (sect. 1). These systems, then, are conceived as features that organisms "acquire." In this mechanistic, particulate, and additive genocentric conceptualization, belief systems are features that are in a way fundamentally separate from the organism and "used" by them. These systems (originating at the genetic level) were selected for because in our evolutionary past they conferred a differential reproductive advantage to their bearers (whether they led to the development of accurate representations of the world or not). From an evo-devo perspective, however, beliefs are viewed as products of organismic adaptability, not evolved adaptations. Adaptability is a property of organisms, not a property of any one part (adaptation). Further, beliefs are manifested at the organismic level, not at the genetic level. From this perspective, the ability to form beliefs may be conceived of as a fundamental property of complex organisms. It is not something that is "selected" for as something additional to the organism itself. It is an emergent product of animal biological systems and inherent in certain biological forms. It is based on a form of life that requires successful interaction with one's environment. Thus, at least the basis for "beliefs" is an intrinsic property of certain organisms, not a separate acquired adaptation.

The notion of evolved belief systems as adaptations, we argue, rests on a limited genocentric model of evolution and overused metaphor. "Belief systems" that exist as adaptations stem from using a population level of analysis to understand properties at the individual level. Because adaptability is an organismic property, its crucial importance in the development of belief is invisible to a genocentric view. An evo-devo conception of belief leads to understanding beliefs through developmental analysis rather than through genocentric, non-developmental "adaptational analysis." At the very most, what can be "selected for" are perhaps particular refinements (mechanisms) within a "belief system," not the belief system itself.

Elucidating the myriad factors that can affect development is integral in understanding how an organismic trait such as "beliefs" emerge. One obvious epigenetic resource is the complex rearing environment afforded to the developing human child, including human languages. Indeed, the forms of belief that M&D review emerge as functions and uses of complex forms of such languages. A developmental analysis of forms of belief along with linguistic competence in children is a necessary step to understanding the emergence of more complicated forms of belief through the lifespan. Further, if "language" itself is a crucial resource in such a system, M&D's postulation is strained by the fact that evolutionary pressures are seen to act over a very lengthy time period within a standard pop-gen framework.

Therefore, although we agree with M&D that accounting for the evolution of mistaken beliefs is an important task, we believe, hopefully not mistakenly, that it requires a much broader developmental conceptualization of evolution. In their response to commentaries, we encourage M&D to attempt to integrate their theory with recent work in evolutionary development psychology.

ACKNOWLEDGMENT

Preparation of this commentary was supported by a Social Sciences and Humanities Research Council of Canada (SSHRC) grant to the second author, Timothy P. Racine.

Lamarck, Artificial Intelligence (AI), and belief

doi:10.1017/S0140525X09991385

Yorick Wilks

Oxford Internet Institute, University of Oxford, Oxford, OX1 3JS, United Kingdom.

yorick@dcs.shef.ac.uk

http://www.dcs.shef.ac.uk/~yorick

Abstract: Nothing in McKay & Dennett's (M&D's) target article deals with the issue of how the adaptivity, or some other aspect, of beliefs might become a biological adaptation; which is to say, how the functions discussed might be coded in such a way in the brain that their development was also coded in gametes or sex transmission cells.

There is much of interest in this article on misbeliefs and their "adaptive" value in evolutionary terms, but I have a serious problem in knowing how to assess the arguments: I cannot see what all this has to do with evolution, understood as natural selection of traits inherited though the genome.

McKay & Dennett (M&D) are perfectly well aware of this overall requirement: "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" (sect. 10, para. 5, emphasis in original). The issue of whether beliefs or other mental representations could affect gametes has surfaced in the past in the Artificial Intelligence (AI) context: Longuet-Higgins (in Kenny et al. 1972) discussed how such brain modifications might conceivably affect the gametes, yielding a sort of post-Lamarckianism (cf. Lamarck 1809/1914/1984) with a possible underlying mechanism stated in genetic terms. I am not sure how seriously that discussion was intended to be taken but, unless some such discussion is set out, what can we make of the arguments of M&D except seeing them as simple Lamarckianism without any mechanism at all – that is, mere adaptivity of traits transferred to a successor generation, such as a tendency to create life-prolonging misbeliefs about, say, health (sect. 10) but with no suggestion of how this could be done?

The only alternative account – and M&D do not give this either – is to fall back on a long and honorable tradition of "cultural transmission" from Lovejoy's *Great Chain of Being* (Lovejoy 1936) to Dawkins' memes (Dawkins 1989) and perhaps even Sheldrake's morphic resonance (Sheldrake 1987). The problem with all this is that it has no known empirical, biological, substrate and hence nothing to do with evolution as the latter is normally understood.

Our authors' dilemma is very like Chomsky's in his lifelong campaign to establish a notion of "universal grammar" that is the biological and inherited substrate that allows humans, but not other species, to develop languages of such and such a type. Chomsky (1965) has never established clearly what features constrain the class of languages we call human and which, if found, would presumably correspond to the coding in gametes. Chomsky had additional difficulties because of his distaste for the whole idea of evolution as applied to languages, but the problem is not far from Herb Simon's closely related speculations (Simon 1996) about the need for an inherited capacity to manipulate hierarchically structured representations to explain the mechanisms underlying vision, language, and reasoning. There have also been comic diversions like the discussion following Fodor's *Language of Thought* (Fodor 1981) about the need for concepts as complex as *telephone* to be innate, because they could not be decomposed into simpler concepts without loss of meaning. Again, it is not clear how serious the discussion of the innateness, and inheritability, of the concept *telephone* was intended to be.

But is the discussion of the target article any more firmly based than Fodor's? Can we imagine that special classes of helpful misbelief, and their associated concepts (e.g., "supreme being" in the

article), are any better candidates for heritability through genetic representations than Fodor's *telephone*? And if they are not heritable, why is this discussion taking place in the context of natural selection and Darwinian evolution?

It may be said that the above criticism is unfair and beside the point: If we accept that there is a heritable feature corresponding to human languages, broadly conceived as a class, and that we have the feature and other higher mammals do not, then we can speculate about its content, as Chomsky and his school have, without any knowledge of coding within organic matter of the structures that heritable features require. Similarly, it may be said, we can speculate about the features of a general belief and reasoning system, in just the way that AI has for 50 years, without brain codings, but under general assumptions about which aspects of such a system will aid survivability or reproduction.

But Chomsky did realize that none of this could concern itself with particular languages, since a new brain can learn any language, but only with underlying representational features common to all languages. It seems clear to me that the target article does not observe that constraint for beliefs and therefore has even less plausibility than Chomsky's long-running campaign. The article is largely about particular (mis)beliefs and their function: "I am sexually attractive," "I will get better from this disease," "A supreme being watches my actions." Is their coding any more plausible than that of the concept *telephone*, or some structural feature of Indo-European languages? Is it not much more plausible that all this is learned within some very general structure we can barely specify, though Simon (1996) probably made the best attempt to guess it.

Suppose we stand a little further back, and accept the general Quinean view (see Quine 1953) that we test not individual beliefs against the world but whole belief systems: anthropological, Newtonian, Einsteinian, religious, naive physics (Hayes 1978), folk psychology, and so forth. These are accepted or not, give satisfaction to individuals or not, but none can be completely right or true as a whole; and no scientist thinks any such theory complex is final or correct, though some are plainly better than others, and may have served our species well for millennia. It can be very hard to separate out individual beliefs as misbeliefs, without regard to the whole complex they come from; this is a cliché of the philosophy of science as much as it is of anthropology. It is also very hard to get any clear evidence on how much, if at all, of these complexes could be hard-coded in genetic material: Is jumping out of the way of rapidly approaching large objects innate in humans, as it is in some insects?

I believe one needs to hear more about these issues, and above all about the relationship to a biological substrate, before discussions of the adaptivity of particular beliefs can be more than rewarmed Lamarck – no matter how much fun the discussion is.

Adaptive misbeliefs are pervasive, but the case for positive illusions is weak

doi:10.1017/S0140525X09991543

David Sloan Wilson[a] and Steven Jay Lynn[b]

[a]*Departments of Biology and Anthropology, Binghamton University, Binghamton, NY 12902;* [b]*Department of Psychology, Binghamton University, Binghamton, NY 13902.*

dwilson@binghamton.edu

http://evolution.binghamton.edu/dswilson/

stevenlynn100@gmail.com

http://www2.binghamton.edu/psychology/people/faculty/stevenlynn.html

Abstract: It is a foundational prediction of evolutionary theory that human beliefs accurately approximate reality only insofar as accurate beliefs enhance fitness. Otherwise, adaptive misbeliefs will prevail. Unlike McKay & Dennett (M&D), we think that adaptive belief systems rely heavily upon misbeliefs. However, the case for positive illusions as an example of adaptive misbelief is weak.

Dozens of evolutionists have observed that insofar as beliefs are products of natural selection, either proximally or distally, then they should be designed to enhance fitness, not to perceive the world as it really is. One of us (Wilson 1995; 2002; 2010) has used the terms *practical realism* and *factual realism* to contrast these two criteria for evaluating beliefs. At the most foundational level, an evolutionary approach to epistemology predicts that the mind is designed to apprehend reality only to the extent that factual realism contributes to practical realism (Wilson 1990).

The gist of McKay & Dennett's (M&D's) argument is that a positive tradeoff between factual and practical realism usually does exist, with the exception of positive illusions. We admire the way that M&D interpret various categories of belief from an evolutionary perspective (e.g., by-products vs. adaptations) but we disagree with their conclusions. We think that negative tradeoffs are pervasive but that, ironically, the case for positive illusions is weak.

To begin with genetic evolution, when it comes to misbeliefs, deception begins with perception. All organisms perceive only the environmental stimuli that matter to their fitness. Our species can see only a narrow slice of the sound and light spectrum, cannot sense electrical and gravitational fields at all, and so on. We also distort what we can perceive, for example, by turning the continuous light spectrum into discrete colors. Perception might not qualify as belief, but if the former is so prone to adaptive distortions, it would be surprising if the latter was not prone as well.

Proceeding to cultural evolution, we disagree with M&D's assessment that by-product explanations of religion are prevailing over adaptation explanations. The actual emerging consensus is that the two interpretations are more compatible than previously thought. In particular, when we examine by-products versus adaptations separately for genetic and cultural evolution, the by-product proponents could be largely right for genetic evolution and the adaptationists could be largely right for cultural evolution. All adaptations begin as exaptations, so it is possible that cultural evolution has produced highly adaptive religions – and other cultural systems – out of genetic adaptations that evolved for other purposes (Wilson 2005; 2010).

Cultural evolution at the group level leads to a view of cultures as highly adaptive systems that are designed to orchestrate action for members of the culture. If most behaviors come directly from one's culture and only more distally from one's genes, then cultural systems must approach the sophistication of genomes as far as the replication and expression of traits is concerned (Wilson 2002; 2006).

Regardless of how plausible one regards this view of culture, let's explore its implications for adaptive misbeliefs. Somehow, a culture must provide an elaborate guide for how to behave in the many situations encountered by members of the culture. When it comes to motivating behavior, there is nothing like a putative fact. If I disapprove of what you are doing, I can call you sick or immature. If I think that a woman's place is in the home, I can believe that women are mentally inferior and that it is abnormal for them to become sexually aroused. If I despise my enemy, I might think that an enemy lacks compassion even among its members.

We trust that we don't need to belabor the point. Cultures are awash in putative facts that can be easily explained in terms of the behaviors they motivate and clearly depart from factual reality. People fight to establish adaptive misbeliefs.

Some of the candidate misbeliefs considered by M&D are self-limiting; for example, when lies are effective only when rare. Useful fictions as cultural instructions are different. They are often in everyone's interest, they are taught to young children who have no basis for rejecting them, and disbelievers are

punished. Even those who try their best to disbelieve can be overwhelmed by constant repetition and pressure to conform. As Daniel Gilbert and his colleagues have shown (reviewed in Gilbert 1991), the default assumption is to believe what one hears, and disbelieving requires effort. Propaganda works because the mind's capacity to disbelieve can be overwhelmed if you shovel in falsehoods fast enough.

Often apparent in cases of ideological extremism, marked by strongly held and sometimes patently false beliefs, is the ubiquitous confirmation bias – the tendency to seek out consistent evidence and to ignore, dismiss, or selectively reinterpret contradictory evidence (Lilienfeld et al. 2009). False beliefs can become deeply entrenched when imbued with strong or "hot" affect, reinforced by social interactions (i.e., people seek out others with similar beliefs), and embedded in matrices of affirming consistent beliefs. To compound recalcitrance to challenging false beliefs and allied actions: (a) people often lack both knowledge of how they form beliefs about themselves (T. D. Wilson 2009) and awareness of their own biases (yet they are quick to point out the biases of others: Pronin et al. 2004); (b) people rationalize their actions in terms of existing beliefs (Kermer et al. 2006); and (c) people process information automatically, with minimal, if any, conscious introspection or challenge (Kirsch & Lynn 1999). Even memory conspires to fortify the status quo: We tend to remember events (whether conforming to historical facts or false memories) that are plausible; that is, based on our beliefs about ourselves and the world, as well as on our current moods and action tendencies (see Lynn & McConkey 1998).

Theoretically, misbeliefs can be subjected to a cost-benefit analysis. The benefits, which might accrue to either individuals or groups, are the actions motivated by the putative facts. The costs are the consequences of ignoring the real facts over the long term. The belief that global warming isn't caused by people, for example, leads to short-term benefits and potentially disastrous long-term costs. The title of Al Gore's movie, *An Inconvenient Truth*, says it all.

In a given culture, the costs and benefits can be expected to weigh in favor of apprehending factual reality in some cases and treating fiction as fact in others. According to anthropologists such as Malinowski (1948), all cultures have a mode of thought that can be recognized as rational and proto-scientific, but which is expressed only some of the time. That is exactly what we should expect from an evolutionary perspective, which also explains why science has only a toehold in our own culture and is ignored whenever convenient.

Ironically, while we think that M&D have underestimated the importance of adaptive misbeliefs, we also think that the evidence for the adaptive value of positive illusions is weak. For instance, one survey of over 15,000 studies (Baumeister et al. 2003) revealed that self-esteem is minimally related to interpersonal success, and not consistently related to alcohol abuse, drug abuse, and smoking. A subset of high–self-esteem individuals who are narcissistic, with inflated yet unstable self-esteem, are at highest risk for physical aggression (Baumeister 2001). Relatedly, bullies have inflated self-perceptions (Baumeister et al. 2003), and aggressive children overestimate their popularity (relative to peer ratings) more than non-aggressive children, a tendency especially marked among narcissistic children (Barry et al. 2003). An inflated sense of self can have serious costs.

M&D cite research that denial as a coping strategy decreases the risk of recurrence of breast cancer. Yet it is difficult to comprehend how denial would promote cancer survival if it discouraged treatment-seeking or follow-through. Moreover, the lion's share of findings indicates that positive beliefs have no bearing on cancer survival (Beyerstein et al. 2007; Phillips 2008). A nine-year study (Coyne et al. 2007) of more than a thousand patients with advanced head and neck cancer revealed that hopeful patients were no more likely to live longer than patients who believed they were "losing hope in my fight against my illness."

The weak evidence for positive illusions notwithstanding, adaptive misbeliefs are so pervasive that we wonder how two smart people such as M&D could have missed them. One reason might be that they take a predominately individual approach to beliefs and say little about culture, much less cultures as something comparable to a genome. We know from Dennett's other work (Dennett 2006a) that he doesn't *deny* the possibility that memes can form into "memeplexes" in addition to acting on their own, but M&D don't explore the implications in their target article. Our commentary, which points out the obvious in retrospect, demonstrates the utility of thinking about adaptation above the level of the individual.

Adaptive self-directed misbeliefs: More than just a rarefied phenomenon?

doi:10.1017/S0140525X0999149X

Tadeusz W. Zawidzki

Department of Philosophy, George Washington University, Washington, DC 20052.

zawidzki@gwu.edu

http://www.gwu.edu/~philosop/faculty/Zawidzki.cfm

Abstract: I argue that adaptive, self-directed misbeliefs are likely more prevalent and important than McKay & Dennett (M&D) claim. Humans often falsely interpret their own behavior in terms of culturally afforded categories. Despite their falsity, such self-interpretations are often adaptive because of our disposition to behave consistently with them. This makes us easier to interpret by similarly enculturated interactants.

McKay & Dennett (M&D) argue that misbeliefs are adaptive only "in certain rarefied contexts" (sect. 15, last para.). Only certain "positive illusions" about one's own capacities are adaptive. In the following, I argue that the authors seriously underestimate the prevalence and importance of adaptive self-directed misbeliefs in human populations. One of the most important tasks facing a human being is making oneself as interpretable as possible to one's fellows (Cash 2008). Behaving consistently with self-interpretations afforded by one's culture, even if they are inaccurate, is an important means by which human beings accomplish this task. Religious beliefs constitute an important class of such culturally afforded self-interpretations.

It is surprising that M&D ignore certain prominent kinds of adaptive, self-directed misbelief suggested by Dennett's own corpus. According to Dennett (1991), the widespread belief that human cognition is under the control of a conscious self is, in many respects, *false* – a "user illusion" (p. 216) – yet *adaptive*, enabling cognitive "self-control" (pp. 417–18). Elsewhere, Dennett suggests that all available evidence often fails to determine which of two competing interpretations of a person's behavior is correct: There may be no fact of the matter whether a famous art critic *really* admires his son's mediocre art or is actually a victim of self-deception (Dennett 1978, pp. 39–49). If Dennett is right, then, to the extent that our self-interpretations presume to settle such interpretive indeterminacies, our self-interpretations are often false. If such self-directed misbeliefs can nonetheless help our survival prospects, then they can be adaptive.

Here is how this might work. We often form beliefs about what we want and think based on the categories of interpretation afforded by our language and culture. Such self-interpretations are often false because, even if Dennett is wrong that all the evidence fails to select between competing interpretations, in everyday contexts, we often do not have the time to consult all relevant evidence before settling on one. Do I really think an author I am reading is pretentious, or is this self-deception concealing my

envy of her masterful prose? None of the subjectively or objectively available data to which I realistically have access need settle this question. Indeed, the available data might also fail to rule out some third, incompatible and, perhaps, linguistically inexpressible, self-interpretation. Yet, context often requires that, even in the absence of determinative evidence, we express self-interpretations employing categories familiar to likely interactants. We also have a standing desire to behave consistently with our publicly expressed self-interpretations (Carruthers 2009, p. 127). So, once we pick among such underdetermined self-interpretations, we tend to act consistently with our choice. Why is this adaptive? It makes us easier to interpret by our likely interactants, enabling complex coordination and cooperation.

There is strong evidence for this mechanism from pathological cases, such as split-brain patients. It is possible to induce behavior in such subjects in ways that escape their awareness (Carruthers 2009, p. 126; Gazzaniga 1995, p. 1393). For example, it is possible to induce standing in a sitting split-brain subject by flashing the word "walk" in her left visual field, to which her left, complex-language-encoding hemisphere has no access. The right hemisphere knows enough language to trigger compliance with the one-word command. When such subjects are asked why they stand up, they immediately confabulate a self-interpretation, for example, "I'm going into the house to get a Coke," *and proceed to comply with this completely false self-interpretation*. Presumably, when the subject's left hemisphere processes the question, it cannot access the correct answer, and so, quickly confabulates a culturally afforded self-interpretation, upon which the subject proceeds to act. My suggestion is that, in dynamic, temporally constrained, quotidian contexts, everyone's access to their true motivations is incredibly limited, so our self-interpretations are usually false – based on quick and dirty motivation attributions afforded by our language and culture. Thanks to the desire to act consistently with one's self-interpretations, however, such self-directed misbeliefs can dramatically facilitate interpretation and coordination among similarly enculturated interactants.

If this proposal is on the right track, then it is relatively easy to see why many religious misbeliefs are adaptive, contrary to the "by-product" theory of the evolution of religious belief endorsed in the target article. The reason is that religious interpretive frameworks are widespread in many populations – that is, many cultures afford religious self-interpretations. Thus, for many human beings, the only interpretive tools available in quotidian, interactive contexts are saturated with religious self-conceptions. So the mechanism described above ensures that many human beings develop religious, self-directed misbeliefs that are adaptive. Relative to some populations, coordination is facilitated by self-interpretations such as "I am inspired by the Holy Spirit!" This is not a defense of "belief in belief" (Dennett 2006a). Much as the capacity to learn English is adaptive only relative to an environment that includes other English speakers, the disposition to self-interpret in religious terms is adaptive only relative to a social environment consisting of interpreters that find such interpretations intuitive. Religious self-interpretation might be adaptive relative to the kinds of interpretive frameworks that happen to have dominated human populations historically. This does not mean that it is adaptive relative to any kind of interpretive framework, or that only religious interpretive frameworks are possible for human populations.

If I am right, adaptive, self-directed misbeliefs are much more prevalent and important than M&D acknowledge. In order to coordinate with our fellows, we need to make ourselves as easily interpretable by them as possible. But, usually, we have very poor access to our true motivations in dynamic, quotidian contexts. One mechanism for dealing with this involves the cultural affordance of ready-made self-interpretations with which most members in a population are familiar. We interpret our motivations using such culturally afforded resources, even though such self-interpretations are likely false. Then we act in ways that confirm such self-interpretations, making interpretation by and coordination with others familiar with such interpretations much easier. As Dennett aptly puts it, we are "creatures of our own attempts to make sense of ourselves" (Dennett 1987, p. 91).

ACKNOWLEDGMENTS

I am grateful to Mason Cash, Bryce Huebner, Dan Hutto, Robert Lurz, Joe Paxton, Don Ross, Eric Saidel, Whit Schonbein, and Robert Thompson for helpful comments. Any remaining flaws are entirely my fault.

Authors' Response

Our evolving beliefs about evolved misbelief

doi:10.1017/S0140525X09991555

Ryan T. McKay[a] and Daniel C. Dennett[b]

[a]*Centre for Anthropology and Mind, University of Oxford, Oxford OX2 6PE, United Kingdom;* [b]*The Center for Cognitive Studies, Tufts University, Medford, MA 02155-7059.*

ryantmckay@mac.com http://homepage.mac.com/ryantmckay/
ddennett@tufts.edu http://ase.tufts.edu/cogstud/incbios/dennettd/dennettd.htm

Abstract: The commentaries raise a host of challenging issues and reflect a broad range of views. Some commentators doubt that there is any convincing evidence for adaptive misbelief, and remain (in our view, unduly) wedded to our "default presumption" that misbelief is maladaptive. Others think that the evidence for adaptive misbelief is so obvious, and so widespread, that the label "default presumption" is disingenuous. We try to chart a careful course between these opposing perspectives.

R1. Introduction

We are very gratified by the thoughtful and temperate responses to our target article. Our aims were ambitious, and the commentaries reflect the broad scope of the topic we tackled. In this response we will try to attend to the most important themes that have emerged. We cannot hope to address each and every substantive point our commentators have raised, but we will try not to shy away from the thorny issues.

We began with a "default presumption," or "prevailing assumption" – that veridical beliefs beget reproductive fitness. Simply put, true beliefs are adaptive, and misbeliefs maladaptive. Our aim was to investigate an alternative possibility, the possibility of *adaptive misbelief*. **Liddle & Shackelford** note that the epigraphs that introduce certain sections of our manuscript showcase this alternate perspective; the implication, they suggest, is that the possibility we explore is already well established, in which case our "prevailing assumption" is a straw man. A similar point is made by **Cokely & Feltz**, who note that the argument for adaptive misbelief is not new. We agree with Cokely & Feltz – the argument is not a new one. Had it been our intention to suggest otherwise, we would have been rather unwise to incorporate the aforementioned quotations. The fact that the argument is not new, however,

does not mean that it is accepted. One has only to glance through the commentaries to see that the issue is far from settled. We put forward a somewhat tentative claim about adaptive misbelief – only positive illusions, we argued, fit the bill. Interestingly, while some of our commentators (e.g., **Dunning**; **Dweck**; **Flanagan**; **Frankish**; **Konečni**; **Kruger, Chan, & Roese** [**Kruger et al.**]; **Marcus**; **Millikan**; **Wilks**) appear to think that we went too far here, a slew of others seem to think that we didn't go far enough (e.g., **Ackerman, Shapiro, & Maner** [**Ackerman et al.**]; Cokely & Feltz; **Haselton & Buss**; **Johnson**; **Mishara & Corlett**; **Randolph-Seng**; **Schloss & Murray**; **Talmont-Kaminski**; **Zawidzki**). You can't please everyone. As we see it, one of the main contributions of our article is to reveal these striking differences of opinion and perspective. Our response is ordered roughly as follows: After clarifying some points about evolution that met with confusion or disagreement, we respond first to those who think our claim errs on the generous side, and then turn to those who view our claim as overly cautious and who seek, one way or another, to extend our analysis.

R2. Oversimplify and self-monitor

As several commentators (e.g., **Boyer**; **Sutton**) point out, cognitive systems are necessarily compromises that have to honor competing demands in one way or another. Since time is of the essence, the speed-accuracy tradeoff is critical; cost also matters so "fast and frugal" systems or methods (Gigerenzer & Goldstein 1996; Gigerenzer et al. 1999) are often the order of the day. But these can generate errors in abundance, so if the animal can afford it, it is good to have a meta-system of one kind or another in place, monitoring the results, discarding bad outputs when they arise, and shifting methods if possible. Good advice, then, in both animal design and artifact design, is *oversimplify and self-monitor* (Dennett 1984a).

The *BBS* format enables this strategy, and we followed it in our target article. Our deliberately oversimplified definition of belief set the table for a variety of useful commentaries showing just how complicated these issues truly are. We had a lot of ground to survey, so we decided, pragmatically, to paint with broad strokes, and to come back later (in this response to commentary) with the called-for corrections. As several commentators (e.g. **Cokely & Feltz**; **Gjersoe & Hood**; **Liddle & Shackelford**; **Wereha & Racine**) point out, our hypergeneral definition of belief, as "a functional state of an organism that implements or embodies that organism's endorsement of a particular state of affairs as actual" (target article, sect. 1, para. 1), blurs the oft-proposed boundaries between a range of arguably distinct types of cognitive states. (It is also worth remembering that in the working vocabularies of many people, the everyday term "belief" is restricted to matters of great moment only – religious belief, political creed, and other topics of capital-b Belief – and would not be used to discuss one's current perceptual state or whether there was beer in the fridge.) We did discuss, and approve of, Gendler's (2008) *alief/belief* distinction, and **Ainslie** puts it to good use, also reminding us (in personal correspondence) of Gendler's useful mnemonic characterization, which we should have quoted in the target article:

> [*A*]lief is *a*ssociative, *a*utomatic, and *a*rational. As a class, aliefs are states that we share with nonhuman *a*nimals; they are developmentally and conceptually *a*ntecedent to other cognitive attitudes that the creature may go on to develop. And they are typically also *a*ffect-laden and *a*ction generating. (Gendler 2008, p. 641; emphasis in original)

But we did not even mention, as **Frankish** points out, the acceptance/belief distinction, which, he argues, may turn out to play a key role: A pragmatic acceptance is not, strictly speaking, a misbelief at all; and our prime candidates for adaptive misbeliefs, positive illusions, may be voluntarily adopted policies, not involuntarily imposed biases – in us, if not in other animals incapable of such "metacognitive" evaluations. This leads Frankish to a sketch of an experimental paradigm well worth pursuing. **Flanagan** and **Konečni** raise similar objections. Flanagan comments on the strategic role of *statements of belief* in competitive contexts, but notes that there is nothing epistemically disreputable about believing that one *can* win: "'can' does not entail 'will.'" Further on, however, he makes a telling slip: "Both players, if they are any good, go into the match believing that they can win, *indeed that they will win*" (our emphasis). Flanagan is right that there is no mistake in believing that one can win, or in hoping that one will win. But where both players believe that they *will* win, we have misbelief (although not necessarily unreasonable misbelief; each may have compelling reasons for expecting to win). Insofar as such misbelief boosts confidence and enables honest signaling of such confidence, it may be adaptive. Like Frankish and Flanagan, Konečni suggests that positive illusions may represent doxastically uncommitted action policies. **Haselton & Buss** and **Johnson**, however, take roughly the opposite view, arguing that genuine (mis)beliefs may generate adaptive behavior more effectively than cautious action policies. We return to their commentaries further on.

It is tempting to re-baptize acceptance as *c-lief*, since acceptance stands to belief roughly as belief stands to alief, a more sophisticated and expensive state, reserved now for just one species, humans. (Cf. Dennett's 1978 belief/opinion distinction, which is explicitly modeled on *betting on the truth of a sentence* which one believes [not alieves] to be true.) But this won't help resolve all the confusions, since, as **Krebs & Denton** observe, collaborative positive illusions (e.g., "I'm OK, you're OK") may begin as pragmatic policies or acceptances – we are, in Haidt's (2001) nice observation, intuitive lawyers, not intuitive [truth-seeking] scientists – but among the effects in us are unarticulated cognitive tendencies that may be best seen as akin to aliefs – except for not being *a*ntecedent to all other cognitive attitudes.

R3. Is our evolutionary thinking naïve?

Several commentators challenge our frankly adaptationist reasoning as naïve, and our "reverse engineering" perspective on misbelief is seen as blinkered or distorting. The points they raise are instructive, but serve rather to expose the weaknesses of various standard objections to adaptationism. **Wilks**, for instance, sees ghosts of Lamarck and Sheldrake's morphic resonances (!) in our

project, and suggests that even Chomsky's curious views on evolution have more plausibility than ours. Wilks "cannot see what all this has to do with evolution, understood as natural selection of traits inherited through the genome," and indeed, from that pinched perspective, it is not surprising that he would miss the point. Natural selection is not just about "traits inherited through the genome." Perhaps he was misled by the fact that we carefully distinguished – as some do not – between genetic fitness and human happiness; but we also went to some lengths to note the role of gene-culture coevolution, and nowhere did we restrict natural selection to genetic evolution (see sects. R5 and R7). Wilks expresses doubts about how "brain modifications might conceivably affect the gametes," ignoring the Baldwin Effect (Deacon 1997; Dennett 1991; 1995a; 2003b), a particularly clear path by which surprisingly specific talents can migrate from brain modifications into the genome. (Fear of Lamarckian heresy has prevented many from taking the Baldwin Effect seriously; it is *not* heretical biology.) His comparison with Fodor's notorious example of an innate concept of *telephone* is simply a straw man. There are plenty of well-proven cases in which *ecologically significant* contents of considerable specificity are genetically transmitted. The fear of snakes exhibited by laboratory-raised monkeys and small children who have never seen a snake (Mineka et al. 1984; LoBue & DeLoache 2008), for example, or the species-specific nest-building dispositions of birds that have never seen such a nest being built should temper his incredulity.

Especially in the case of behavior regulators, there is typically an interplay, a coordination, between genetically transmitted features and culturally (or "socially") transmitted elements. **Marcus** is right that it does not follow that if something could be learned, it must be learned; it might well be innate, but also vice versa, as Avital and Jablonka (2000) show: Many long-presumed innate animal "instincts" turn out to be learned behaviors, copied in one way or another from parents' behavior, not part of their genetic legacy, as a host of cross-fostering studies demonstrate. The genes fix the disposition to attend to what the parents do, but the rest is up to environmental transmission. What this fact brings to our attention is that Mother Nature is not a gene-centrist! Where genetic evolution leaves off and developmental and, indeed, "empiricist" (Marcus) psychological learning takes over, is an entirely open option, with large differences between fairly closely related species. (Consider the wide variation in the extent to which species-typical bird song is innate.) Marcus's example of "learning to walk" is useful, since, as he says, there is an innate stepping reflex in humans that exists at birth. On top of this reflex comes something we can still call learning to walk. His point is not that walking is innate in humans – it isn't, when compared with, say, the walking (indeed, *running*) skills found in a newborn antelope. Where innate instinct leaves off and learning begins is not a line that can be, or need be, sharply drawn. It goes without saying, we thought, that belief-generating mechanisms depend critically on environmental input, but we should have said it anyway, as several commentators (e.g., **Dunning**; **Dweck**) chide us for underestimating the importance of environmental variation. So we agree with Dweck, **Liddle & Shackelford,** and **Wilks** that individual, isolated beliefs are unlikely to be the target of genetic selection, but that does not imply that quite specific biases could not be incorporated into our genetically transmitted equipment. We may in effect be primed to *imprint* on whatever in the environment fills a certain fairly specific doxastic role, much as newly hatched ducklings imprint on the first large moving thing they see and follow it.

Similarly, as a number of commentators reveal, the line between by-product and adaptation is not sharp at all. Every adaptation, after all, must emerge from something that varies "randomly" (under no selection) or from some prior arrangement that persists for other reasons, and what had heretofore been a by-product is brought into focus and enhanced and exploited by selective pressure. Showing that something is (likely) a by-product does not rule out the possibility that there is (already, as it were) opportunistic selective pressure on it. The bright colors of autumn foliage of deciduous trees in New England are probably just a by-product of the chemistry of chlorophyll loss after leaf death (though this has recently been challenged by evidence that it signals either inhospitality or vigor to aphids looking for a winter home; see Yamazaki 2008); but whether or not aphids are attracted to, or repelled by, bright autumn colors, assuredly there is now selective advantage to having brilliant autumn color in New England. The economies of Vermont, New Hampshire, and Maine benefit significantly from the autumn "leaf-peepers" (foliage enthusiasts) that invade, and hence there is a pronounced bias against cutting down handsome trees and for planting, or encouraging the growth of, the most colorful variants.

Adaptationists know – or should know, since the classic work of George Williams (1966) – that the evidential demands for establishing an adaptation are greater than the demands for discovering a mere by-product. As **Millikan** says, "If certain kinds of errors are common and also systematically useful, it does not follow that they are common because they are useful." It does not follow, but fortunately there are ways of testing to see if and when such adaptationist hypotheses are true. Sometimes, however, the tests are too impractical to carry out (they might require a few thousand years of observation of evolution, for instance), and often the adaptation is so obvious, once discovered, that nobody bothers challenging the claim. It is interesting that the *charge* of "Just So Story" leveled at adaptationists is almost entirely reserved for hypotheses dealing with features of *human* evolution. A brief canvassing of textbooks of biology will find literally thousands of examples of confidently asserted adaptationist claims that have never been challenged and never been *thoroughly* tested – claims about the functions of enzymes, the functions of organs, the functions of behaviors (of protists, animals, plants. . .). People get touchy when their own organs and behaviors are analyzed from an adaptationist perspective, but unless they are prepared to dismiss the mountains of insight to be found in the rest of biology, they should stop treating "Just So Story" as a handy-dandy wild-card refutation-device. It is no such thing.

Critics of adaptationism are right, however, that there is a perilous amount of free scope in the range of permissible hypotheses. For instance, why does nature so often counteract one bit of flawed design with another, compensatory

one, instead of just "fixing" the first? Maybe there is a constraint – so far unknown – that renders the latter course impossible or more expensive (see discussion of **Haselton & Buss** further on, and see McKay & Efferson [under review] for a discussion of constraints in the context of error management theory). Such chains of reasoning are not just flights of fancy, since there are differential consequences that can usually be tested for, but until such tests are conducted, we are left with merely plausible conjectures. This open-endedness haunts the discussions further on in this response – see, for example, our discussion of **Johnson –** since the question to which commentators continually return is whether it is *cheaper or easier* for the mind to deceive itself with misinformation than to provide accurate information and adjust its prudential policies to fit the risk. Until we can assess this by evaluating known cognitive mechanisms and their evolutionary costs, this question must remain unsettled.

Wereha & Racine chant the standard evo-devo mantra, claiming that "by reverse engineering the beliefs of adult humans" we forgo a developmental analysis – which is true enough, but does it matter in this case? They are right, of course, that environmental interactions, especially those that engage language, are crucial, and for that reason cultural-genetic interactions are necessary. What we disagree with is their claim that evo-devo considerations obviate or even blunt the effectiveness of reverse-engineering approaches. One simply has to do one's reverse-engineering with more attention to the myriad possibilities raised by developmental demands. One way of putting Wereha & Racine's main claim is this: Because development, from embryo on, is a process that has to protect the robustness of the organism at every stage, later (e.g., "adult") features could just as easily be leftovers, fossil traces, of features that paid for themselves in infancy as features that pay for themselves in adulthood. That is, indeed, a distinct possibility that needs to be considered. And **Gjersoe & Hood** provide a possible example: the entrenchment phase in hypothesis formation in childhood development. This oversimplification strategy has a huge payoff: Oversimplify and (eventually) self-monitor, but in those who are not particularly reflective, a tendency to cling uncritically to one's first hypothesis might be a residue of a particularly adaptive bias in childhood that has outlived its usefulness.

R4. Adaptive oversimplifications

Oversimplifications that make cognitive life easier are also proposed by **Zawidzki,** who notes that Dennett himself (1991) has argued that the concept of a self is a benign "user illusion." Zawidzki also notes that much of the specificity of self-interpretation may be an artifact of societal demand, adaptive in the context of complex sociality. We indeed overlooked the role of such oversimplifications as instances of adaptive *misbelief*, probably because, like the concept of a center of gravity, they can quite readily be recast as something like strategic *metaphors* rather than falsehoods (consider the widespread understanding that there is nothing pejorative in the everyday understanding of the "user illusion" that makes laptops so user-friendly). We agree that they make an illuminating further category to explore (see also the discussion of free will later).

In this light, **Bertamini & Casati** could be seen as suggesting that naïve physics is also an instance of oversimplify and self-monitor, on a hugely different time scale. We are only in recent centuries beginning to discover the falsehoods latent in our everyday conception of the world, a conception that is, as they say, "prima facie veridical" in that it does not "interfere with our interactions with the world." This pragmatic effectiveness, of course, is the evolutionary rationale for the default assumption that true beliefs are adaptive and misbeliefs are not. As **Millikan** observes, "it is getting straight about what is in front of our noses that is the first order of importance for us." (**Boyer** says that what matters for adaptive design is "that the circumstances in question be such that decision-making does not lead to excessive vulnerability.") Not bumping into dangerous things, and finding food, shelter, and mates requires a certain amount of effective information-gathering – and not misinformation-gathering. **Wilson & Lynn** view the fact that our senses give us only a narrow window on the physical variation in the available environmental stimuli as "deception," but that is unwarranted; sensory systems that provide a truncated or edited message that informs may not give the whole truth, while giving, normally, nothing but the truth. The obvious norm for information-gathering is *not* to be deflected by the motivational system, since wishful thinking is typically unrealistic and sometimes catastrophically so (**Ainslie**). At the same time, as various commentators note, there can be overriding reasons for editing the information-gathering to accomplish various palliative ends. If the truth hurts too much, it will disable, not enable, the intentional agent.

R5. Illusions and collusions

Although a number of commentators (**Ackerman et al.**; **Brown**; **Gjersoe & Hood**; **Krebs & Denton**) endorse our claim that positive illusions represent sound candidates for adaptive misbelief, others are skeptical. First, there are methodological and statistical concerns. For example, although **Cokely & Feltz** argue that adaptive misbeliefs are in general much more widespread than we allowed, they point out that better-than-average effects can represent statistical artifacts. The fact that driving ability has a negatively skewed distribution means that most drivers simply *are* better-than-average; the mean is clearly an inappropriate measure of central tendency in this case. We do not dispute this, but we note that better-than-average effects are also documented for normally distributed traits, and (as **Kruger et al.** note) the effects replicate when other, similar methodological points are taken into consideration.

A different sort of concern has to do with the contexts in which positive illusions are observed. Some commentators (**Dweck**, **Flanagan**, **Konečni**, and **Kruger et al.**) appear to suggest that if such illusions are a product of genetic evolution, they should not be confined to a particular culture or to a particular historical epoch. Moreover, as **Dunning** observes, they should be particularly evident in tasks with adaptive significance. No one has performed the latter analysis across task contexts, as Dunning notes, but there are data about the cross-cultural replicability of positive illusions. Unfortunately, however, there does

not appear to be consensus on this issue: **Brown** states that "positively-biased self-perceptions are a pervasive, cross-cultural phenomenon," but Dweck, Flanagan, and Kruger et al. express doubts about the cultural universality of positive illusions, noting that they are more reliably documented in Western societies. In any case, we note here that cultural variability is by no means a decisive datum against the evolutionary claim: if cultural evolution plays a coevolutionary role, there may be, in effect, cultural subspecies of evolved misbelief.

Konečni claims that positive illusions are a feature of a particular historical period: "the recently terminated era of easy credit." We enjoyed his commentary, and await empirical substantiation of this claim. We are confused, however, by his methodological critique of studies that purport to demonstrate positive illusions regarding participants' children. On the one hand, Konečni complains that such "studies have presumably not polled the opinions of the parents (including potential ones) who terminated pregnancies – or committed infanticide, physical and/or sexual abuse." His implication at this point seems to be that offspring-directed positive illusions are an artifact of biased sampling. He goes on, however, to suggest that had such parents been polled, they *would* have demonstrated positive illusions regarding their children; this, he suggests, would undermine the suggestion that biased offspring appraisals facilitate parental care. We think Konečni is trying to have his cake and eat it too. Leaving aside pregnancy terminations (which were presumably uncommon in ancestral environments), we agree that demonstrations of offspring-directed positive illusions in abusive parents would undermine the evolutionary argument, but we doubt that such parents would harbor these illusions. This doesn't mean, however, that the finding of widespread offspring-directed positive illusions is a statistical artifact – that depends on whether parental care is normally distributed (abusive parents may represent a "bump" at the lower end of the distribution), and on how much of the distribution was sampled by the relevant studies. But if it *is* an artifact, it's nevertheless a telling one, because it implicates a positive correlation between offspring-directed positive illusions and parental care, consistent with our evolutionary suggestion.

Kruger et al. raise more serious concerns in their measured and informative commentary. They question whether positive illusions are the norm for healthy individuals, and they point out the many instances of systematic *negative* (self) illusions. (**Ackerman et al.** also speak of negative illusions, but they refer to illusions that are negative with respect to others.) We appreciate this point, but we note that the existence of negative self illusions is not in itself problematic for claims about the adaptive significance of positive self illusions – although it may, as **Bertamini & Casati** recognize, demonstrate that the relevant mechanisms are domain-specific rather than domain-general. As Haselton and colleagues have noted, a tendency toward false positives may be adaptive in certain adaptive contexts (as in the male sexual overperception bias that **Haselton & Buss** describe; see sect. R7 for further discussion), whereas a tendency toward false negatives may be adaptive in others (as in the female commitment underperception bias that Haselton & Buss report elsewhere; see Haselton & Buss 2000; see also Ackerman et al.'s comments about the benefits of being "hard to get"). Different domains will call for biases in different directions. It is worth citing Hartung's (1988) speculations about the adaptive value of negative self illusions in certain circumstances, what he calls "deceiving down." It remains to be demonstrated, of course, that the negative illusions that Kruger et al. mention are domain-specific adaptations.

Dunning also emphasizes the role of environmental context (as does **Dweck**), noting that misbeliefs often arise because the environment fails to furnish the information needed to form accurate judgments. Illusions, on this view, reflect forgivable design limitations rather than design features. Pessimistic predictions about the trustworthiness of others may persist not because they are fitness-enhancing, as **Ackerman et al.** suggest, but because they are liable to confirmation but not to refutation. **Marcus** makes a similar point, noting that illusions may reflect the operation of a general confirmation-bias mechanism rather than dedicated domain-specific machinery. We note that even if a general confirmation-bias mechanism generates illusions as a well-entrenched subclass of outputs, the serendipitous benefits that those outputs provide might "protect" the confirmation bias mechanism (which does, after all, output a lot of mistaken cognition) from counter-selection, helping to "pay for" its persistence. If underestimations of others' trustworthiness are less costly than overestimations (Ackerman et al.), then a mechanism that generates underestimations (initially) as a by-product may be a candidate for exaptation.

Wilson & Lynn also mention the confirmation bias, linking it to the motive force of strong or "hot" affect. **Marcus** suggests that positive illusions are potentially underpinned by motivated reasoning, but we see this possibility as a potential generalization of – rather than necessarily an alternative to – the evolutionary claim we defended. As **Ainslie** discusses, in an extraordinarily rich and compressed commentary, motivational and affective forces represent the proximal mechanisms by which natural selection tethers belief to survival. Our abilities to appraise evidence dispassionately may be selectively sabotaged (motivationally biased) in adaptive domains, yielding positive illusions. Motivated reasoning might not be adapted, however – it might be largely a by-product of the selection for increasing intelligence that Ainslie describes, a process which enabled our ancestors to discover the intervening carrots and sticks of reward, and to begin devising ways of getting the carrots without going to the trouble of checking on the world. Human imagination was born, with all its costs and benefits. The result has been "the unhitching of reward from adaptiveness" and in the ensuing holiday of imagination, we have had to create methods of epistemic self-control to protect ourselves from our own freedom. Exploring the further wrinkles Ainslie draws to our attention will have to wait for another occasion.

Wilson & Lynn point out that inflated self-esteem can come at a cost. We acknowledge that the connections between self-esteem and adaptive behaviours are complex, but we didn't suggest that unrestrained self-esteem would be adaptive, and we cited Baumeister (1989) on the "optimal margin of illusion." In fact, because self-esteem is such a heterogeneous concept we avoided using the term at all in our target article. The large survey that Wilson & Lynn cite points out that the

category "self-esteem" encompasses a range of subtypes. For example, Jordan et al. (2003) characterized the defensive subtype as involving a discrepancy between high explicit (conscious) and low implicit (unconscious) self-esteem. These authors found that individuals with this discrepancy were significantly more narcissistic than individuals high in both explicit and implicit self-esteem. Discrepancies between implicit and explicit self-esteem have also been implicated in the formation of persecutory delusions (Bentall & Kaney 1996; Kinderman & Bentall 1996; 1997; McKay et al. 2007b; Moritz et al. 2006). To the extent that illusory positive self-views are adaptive, therefore, we would predict them to be held at both conscious and unconscious levels.

The above-discussed commentaries provide valuable correctives to our enthusiasm for positive illusions and the evolutionary implications thereof. We acknowledge that more research is needed to clarify whether illusional beliefs are reliably observed in specific adaptive contexts, and whether they trend in the expected directions. We also note, however, that a number of commentaries complement and extend our analysis of positive illusions. We have already mentioned **Gjersoe & Hood**'s suggestion that the developmental phase of theoretical entrenchment involves an adaptive positive illusion – "overconfidence in the generalisability of one's theory." **Brown** and **Krebs & Denton** detail the important role that people play in validating and perpetuating the illusions of others (**Wilson & Lynn** make this point about false beliefs more generally). In our target article we noted how the positive illusions of parents with respect to their co-parents and their children could strengthen familial bonds and facilitate parental care. Brown, however, notes that infants also benefit by internalizing the positive illusions of their parents with respect to themselves. Krebs & Denton point out that individuals may manipulate others into validating their own positive self illusions, but they also appreciate (as does Brown) that this process can be collaborative and mutually beneficial.

Boyer and **Sutton** describe how our own memory systems can be co-conspirators in the maintenance of adaptive illusions. Both commentators note that selection does not indulge abstract epistemic concerns – memories need be accurate, therefore, only insofar as they are fitness-enhancing. Memories that are accurate for accuracy's sake are a biological luxury, so adaptive considerations may frequently trump epistemic considerations. The result is that many of our memories, as beliefs about past occurrences, may be examples of adaptive misbelief.

R6. Delusions and doxastic shear pins

To provide a framework for our discussion, we developed a tentative taxonomy of misbelief. We began by distinguishing two general types: misbeliefs arising in the course of normal doxastic functioning, and misbeliefs resulting from some kind of break in normal functioning. **Liddle & Shackelford** draw our attention to a similar analysis by Wakefield (1992). Wakefield's (1992; see also 1999a; 1999b) concern is to provide a rigorous theoretical grounding for the concept of (mental) disorder. His analysis incorporates both a value component (disorders are harmful, where "harm" is judged by the standards of the relevant culture) and an evolutionary component (disorders reflect the failures of internal mechanisms to carry out their naturally selected functions). We endorse Wakefield's analysis – and regret not previously being aware of it – but note that his project is wider and more general than ours, distinguishing function from dysfunction in naturally selected mechanisms insofar as this distinction can guide decisions about candidates for disorder, while we are interested in belief specifically.

As "disorders of belief," delusions represent the key area of overlap between Wakefield's analysis and ours. Our emphasis was on delusions as the output of belief-formation mechanisms that have ceased to perform their normal (naturally selected) functions – Wakefield's evolutionary criterion. His value criterion, however, is clearly also important: Delusions are harmful insofar as they occasion distress and insofar as they jeopardize the social and occupational functioning of individuals who hold them (this is the "clinical significance" criterion in the DSM-IV-TR; American Psychiatric Association 2000). In our target article we were wary of considering delusions as adaptive, and indeed we labeled them instances of "doxastic dysfunction" (although we didn't clearly discriminate between biological and social conceptions of dysfunction). We did, however, speculate about a class of misbeliefs enabled by the action of system components *designed to break*: doxastic shear pins. Several of our commentators (**Langdon**; **Liddle & Shackelford**; **Millikan**; **Mishara & Corlett**) pick up on this concept.

Langdon notes that if doxastic shear pins exist, their shearing should involve some kind of neurocognitive "short-circuit" rather than a stable neuropsychological impairment. We agree with this point. She also distinguishes neuropsychological (deficit) and motivational answers to the question of why deluded individuals cling to their delusions. **Mishara & Corlett** consider this distinction an "overly strict conceptual schism," and it is true that motivational and deficit hypotheses need not be mutually exclusive (Langdon is well aware of this). Although we found it useful to distinguish between misbeliefs that represent functionless departures from normal operation ("culpable design limitations") and those that incorporate some functional component, we did acknowledge the porous nature of such conceptual boundaries. Nevertheless, we remain open to the possibility of misbeliefs with purely "deficit" aetiologies. Mishara & Corlett, however, favor the doxastic shear pin perspective: delusions accommodate aberrant prediction error signaling, disabling flexible conscious processing and enabling the preservation of habitual responses in the context of impaired predictive learning mechanisms. As such they serve a functional, even biologically adaptive, role. We appreciate this perspective, and we think that work on prediction errors represents a key avenue of research into delusions. Inferences about biological adaptiveness, however, may be unjustified here: As **Millikan** notes, the existence of doxastic shear pins "does not imply that failures to function properly are helpful, but only that in some circumstances it is best not to attempt to function at all."

Coltheart raises issues of truth and groundedness with respect to delusions, and asks us to clarify whether we consider well-grounded false beliefs to be misbeliefs. The short answer to this is, Yes. Misbeliefs are simply false

beliefs – they may be grounded or ungrounded. Grounded misbeliefs reflect forgivable design limitations: in contexts of imperfect information (we may be underinformed or even deliberately misinformed), misbeliefs are inevitable. Ungrounded misbeliefs, on the other hand, may result from culpable failures in naturally selected belief mechanisms (delusions), but they might also reflect designed features of such mechanisms (the adaptive misbeliefs we sought in the target article). **Talmont-Kaminski** claims that ungrounded beliefs fall outside our compass, but he seems to have misunderstood our expository strategy. Ungrounded beliefs *are* within our compass, but only insofar as such beliefs are false. This was perhaps overly stipulative, but it made our discussion manageable (for example, it allowed us to skirt moral beliefs and beliefs about norms more generally). As we stated in our target article, we do not expect adaptive misbeliefs to be generated by mechanisms designed to produce beliefs that are false *per se*. Rather, we implicate evolved tendencies for forming domain-specific ungrounded beliefs. Where these beliefs are (contingently) false, we will see adaptive misbelief. Where they are (contingently) true, they fall outside our purview.

Not all ungrounded beliefs, of course, are adaptive: once again, we argue that such beliefs often reflect breakdowns in belief formation machinery, and where such beliefs are harmful (Wakefield's value criterion), they constitute delusions. But here, too, ungrounded beliefs can be contingently true, as in the delusional jealousy example that **Coltheart** elaborates. We are quite happy for such cases of "serendipitously" true belief (or "accidentally" true, as Coltheart prefers) to count as instances of delusion; they are just not instances of misbelief. Where misbelief is concerned, truth is the critical feature, simply by (our) definition; where delusion is concerned, truth may ultimately be irrelevant (see Leeser & O'Donohue 1999; Spitzer 1990). In cases such as the delusional jealousy scenario that Coltheart outlines, truth or falsity may be difficult to establish – a feature that may contribute to the incorrigibility of such beliefs (many religious beliefs also have this feature; see discussion later). It is worth noting, however, that delusions can resist the presentation of manifestly contradictory evidence; indeed, as **Mishara & Corlett** show, such evidence may even *strengthen* delusional conviction through the process of reconsolidation (see Corlett et al. 2009).

Sperber makes the interesting point that most human beliefs are acquired via communication with others. Because of this, he doubts that most human beliefs are grounded in the sense of being "appropriately founded on evidence and existing beliefs." We appreciate Sperber's general point, but our view is that the testimony of others (whether oral or written) is ultimately just another source of evidence that should be weighed up when forming beliefs. We look up at the sky and form a belief about whether it will rain; later we listen to the weather forecast and revise our belief accordingly. The evidence of testimony may be easier to override than direct perceptual evidence (**Langdon** discusses the idea that delusions involve a loss of the ability to override the latter), but it is evidence that can ground belief nevertheless. We don't see any reason to consider beliefs acquired via communication to be ungrounded. Sperber notes that "from a cognitive and social science point of view, a definition of 'belief' that excludes most religious beliefs renders itself irrelevant." We agree with this, and we think the same of any definition of "grounded" that excludes beliefs acquired by communication. Such a definition would guarantee its own irrelevance.

R7. Error management theory and religion

Haselton & Buss and **Johnson** pick up on our point about how adaptive behavioral biases need not reflect adaptive biases in belief. We do not doubt that when the costs of relevant errors in a given domain are recurrently asymmetric, selection should implement a bias toward committing less costly errors (Haselton & Nettle 2006). Our point was that such biases need not involve a systematic departure from Bayesian belief revision, but merely judiciously biased action policies (see McKay & Efferson [under review] for a more thorough, technical treatment of these issues). A second point we made was that even when selection in accordance with the error management principle plausibly results in biased belief-forming processes, such processes may produce misbeliefs as tolerable by-products rather than as adaptations (such biased systems may be adaptive not by virtue of the misbeliefs they produce, but by virtue of the fact that they minimize misbeliefs of a certain type). **Millikan** endorses this point.

Haselton & Buss provide a valuable counterpoint to our skepticism regarding whether certain error management examples might qualify as examples of adaptive misbelief. They observe that a demonstration that selection *can* solve such adaptive problems without misbelief is not a demonstration that selection *has* solved such problems without misbelief. Selection might have followed any number of design trajectories, subject to the physical, economic, historical, and topographical constraints that we mentioned; it is an empirical question which trajectory was in fact followed. Haselton & Buss go on to suggest several reasons why biased beliefs might have featured in the solution to such adaptive problems. We are not convinced by their first suggestion, that such beliefs "could provide the motivational impetus for courtship behavior." Judicious action policies, after all, would also provide that. Their next suggestion, that male misbeliefs about the sexual intent of women might help allay fears of rejection, is not obviously different from their first: presumably this is also a point about motivational impetus. It is not clear why selection would go to the trouble of instilling fears of rejection and then installing biased beliefs to allay those fears, but again, it is an empirical matter which trajectory was in fact followed. Haselton & Buss's final suggestion seems more promising to us: The confidence boost that biased beliefs provide might be attractive to females in and of itself. In a related analysis, **Ackerman et al.** imply that female misbeliefs about the commitment intentions of men might heighten the desires of potential suitors, leading to increased male investment and ultimately boosting the romantic returns to the females concerned. We have already discussed the similar points that **Brown** and **Krebs & Denton** make about how misbeliefs can transform the psychological states of others.

Johnson takes an error management approach to supernatural belief, and argues that such belief is adaptive. His claim is that selection should favor belief in

supernatural agents because such beliefs would yield exaggerated estimates of the risk of one's social transgressions being detected. In our target article we indicated that we did not think there was strong evidence for this theory. Johnson has several points to make about the priming evidence we reviewed, but none of these points seem to help his case. First, he notes that the religious primes used by researchers tend to be culturally specific – typically derived from Western Judeo-Christian traditions. The issue of cultural specificity is important, especially as regards genetic evolutionary claims (see our remarks earlier concerning the cross-cultural validity of positive illusions), but how should it apply here? We did not contest the findings that religious primes increase prosocial behavior – instead we queried whether such primes exert their effects by activating reputational concerns involving supernatural agents, and we also queried whether such effects are mediated by religious belief. Johnson then states that experiments may not "differentiate the behavior of 'believers' and 'non-believers' – Joe Bloggs may be an avowed atheist who, on his way to Las Vegas, is nevertheless very concerned about seeing a black cat or wearing his lucky jacket or what his grandmother would have said." We are not sure we follow this – we don't see the relevance of such superstitious beliefs to the supernatural watcher hypothesis that Johnson advocates. We do, however, acknowledge Johnson's point that many different belief systems might play the role of his "supernatural watcher" – karmic beliefs in comeuppance might inhibit social transgressions just as effectively as beliefs in personal punitive deities.

A further point that **Johnson** makes concerns the conclusions that can be drawn from priming studies. Evidence that supernatural primes promote prosocial behavior does not, he says, prove that supernatural beliefs are adaptive – such effects "could be evidence that religious primes turn people into suckers who give away precious resources." We are confused by this point. The supernatural watcher hypothesis states that belief in supernatural agents inhibits antisocial behavior and is adaptive by virtue of that fact. Priming studies enable demonstrations of a causal link between religious priming and prosocial behavior. What kind of evidence would Johnson think relevant if not this? He doesn't specify. Perhaps the problem is that Shariff and Norenzayan (2007) reported an increase in prosocial behavior (Dictator Game donations) following religious priming, whereas Johnson's theory requires a *decrease* in *antisocial* behavior. If so, we draw attention to Randolph-Seng and Nielsen's (2007) study, which found that participants primed with religious words cheated significantly less than controls on a subsequent task. The problem, from our perspective, is that this study could not empirically adjudicate between the supernatural watcher hypothesis and an alternative, behavioral priming, interpretation (**Randolph-Seng** does not appear to dispute this point). The same limitation, we argued, applies to the studies of Pichon et al. (2007) and Shariff and Norenzayan (2007).

Norenzayan, Shariff, & Gervais (**Norenzayan et al.**) pick up on this point, noting that supernatural watcher and behavioral priming mechanisms need not be mutually exclusive; they might well operate in tandem, and could even be mutually reinforcing. Nevertheless, these authors marshal evidence that provides support for the supernatural watcher account and yet resists a behavioral-priming interpretation. We appreciate their reference to the study of Dijksterhuis et al. (2008), although we worry that the baby is discarded with the bathwater here: This study disambiguates the felt presence of a supernatural agent from prosocial outcomes, certainly, but only by dispensing with a prosocial component altogether (this is not a gripe about the study itself, but about its interpretation vis-à-vis the supernatural watcher hypothesis). In general, however, we find the arguments of Norenzayan et al. to be quite persuasive. In particular, we are impressed by the results of the Gervais and Norenzayan (2009) study that they mention. The finding that religious primes activate public self-awareness is exactly the kind of result that is needed to substantiate the supernatural watcher hypothesis. We are keen to learn whether such reputational awareness moderates the magnitude of the primes' effect on prosocial behavior.

Norenzayan et al. attribute (mis)belief in supernatural agents to cultural rather than genetic evolution. Although, by their lights, religion does not therefore supply a case of evolved misbelief, we did not intend to restrict our analysis of adaptive misbelief to cases of genetic evolution. On the contrary, we are open, at least in principle, to the possibility that culturally selected religious beliefs constitute adaptive misbeliefs. **Talmont-Kaminski**, **Wilson & Lynn**, and **Zawidzki** provide related analyses. The accounts of Talmont-Kaminski and Wilson & Lynn are in fact almost identical – like Norenzayan et al., they view "religion as a cultural phenomenon that exapts existing cognitive by-products" (Talmont-Kaminski). Wilson & Lynn thus suggest that the tension between by-product and adaptation explanations of religion can be defused: Both camps might be right – the by-product proponents where genetic evolution is concerned and the adaptation proponents where cultural evolution is concerned.

Along with **Johnson**, **Talmont-Kaminski** remarks upon the lack of falsifiability of religious beliefs, and outlines several barriers, physical and social, to the exposure of religious belief as false. As Talmont-Kaminski notes, it is precisely because of such barriers to testability that supernatural beliefs are well suited to serving a functional role. **Sperber** provides an indispensable analysis of an additional class of barriers: barriers to comprehension. For a belief to be open to epistemic evaluation, he notes, it must have a propositional content, a truth value. Many religious beliefs, however, have only "semi-propositional" content – they are mysterious and obscure, permitting manifold exegeses. (Sperber's concept of semi-propositional attitudes has not, alas, been influential among the philosophers who have devoted their careers to elucidating "classical" propositional attitudes. We can hope that a new generation of more empirically minded philosophers will eventually see the utility, indeed the inescapability, of acknowledging this set of at least *belief-like* phenomena.) According to Sperber, such beliefs are better suited to playing an adaptive role than many beliefs with ordinary propositional content: "content unproblematically open to epistemic evaluation might either raise objections within the relevant social group, or, on the contrary, be too easily shared beyond that group."

Bulbulia & Sosis propose yet another variety of beliefs (or belief-like states) whose function is not strictly to

inform (or misinform) the believers about the layout of their world: cooperative commitments. Following Schelling, they suggest that a certain sort of commitment problem might be solved by something like a group myth that gets everybody on the same page, as one says. The commitment problem is this: Getting individuals to cooperate can be like herding cats, but if the cats can be transformed into something more like sheep, by inculcating a religious myth in them all, this may create points of salience that engender the sorts of uniformity of attitude and synchrony of response that make large scale cooperative projects feasible. Once initiated, such a phenomenon might become more or less self-sustaining without any knowing supervision. Indeed, too much knowingness might subvert the whole enterprise, breaking the spell and tumbling everyone back into their feline individuality. It is important to note that if such a phenomenon did evolve (mainly by cultural evolution, one must suppose, with perhaps some genetic predisposition favoring it), individuals could be strongly motivated to resist any developments that threatened to undermine their obliviousness to the motivational source of their "conviction" or "faith" – without needing to know why they were so motivated. As usual, those who were blessed (by natural selection) with the disposition to behave in this way would be the beneficiaries of this clever arrangement without anybody needing to understand the cleverness of it all – until Schelling came along.

Wilson & Lynn give a vivid account of the ubiquity of deception in human culture, but seem to forget the adaptiveness of deceiving *others*. In the target article we set this topic aside as too obvious to need more than a brief review: Of course, it is often "adaptive" for kings to deceive their subjects, for generals to deceive their troops, for everyone to deceive their enemies. Who benefits – *cui bono* (Dennett 1995a; 2006a) – from the false religious and social propaganda that they describe? Wilson & Lynn apparently assume that if it is not the individuals themselves whose fitness is enhanced by believing these falsehoods, it is the groups to which they belong – an instance of group selection utilizing cultural, not genetic, evolution. Again, we are open to this possibility – but we note that these authors overlook the other possibility proposed and defended by Dennett (1995a, 2006a): It may be the memes' *own* fitness that is enhanced by these adaptations, in which case these are instances of other-deception or host-manipulation, not group selection at all. One of the benefits of the memetic perspective is that it exposes the non sequitur in any argument that claims that some features are ubiquitous among groups and (hence) must be adaptive to those groups that have them. In order to establish religion as a case of (culturally selected) adaptive misbelief, one must show that individuals or groups that acquire religious cultural variants have an advantage over those not similarly "infected." We think the jury is still out, and await evidence of this selective advantage.

R8. Truth or consequences

The ways in which the truth of beliefs can be divorced from their consequences for survival may be myriad, but they do not extend as far as **Schloss & Murray** propose. Like some other commentators, they think the case for adaptive (or fitness-neutral) misbelief is stronger than we allow. Our view, Schloss & Murray claim, "requires the falsity of" the radical claims of Churchland, Plantinga, and Stich, and we agree; we think those views are clearly false, for reasons presented elsewhere (on Churchland and Plantinga, see Dennett 2009 and forthcoming; on Stich, see Dennett 1981; 1985). As **Millikan** notes, to say – as Stich does – that natural selection does not care about truth, is like saying that:

> Natural selection "does not care about" digesting food, pumping blood, supplying oxygen to the blood, walking, talking, attracting mates, and so forth. For each of these activities can either be (biologically purposefully) set aside (the vomiting reflex, holding one's breath under water, sleeping) or simply fails to occur in many living things. Nonetheless, surely the main function for which the stomach was selected was the digestion of food, the lungs for supplying oxygen, and so forth, and a main function for which our cognitive systems were selected was the acquisition and use of knowledge – that is, true belief. (Millikan's commentary, first paragraph)

Schloss & Murray also, we think, underestimate the force of Quine's observations on systematic falsehood discussed by us, and their thought experiment about the robot competition can nicely expose the issue:

> While one would surely seek to program competing robots to form beliefs that provided an isomorphic "map" of the external environment, would one further seek to program beliefs about that environment that were true? Not obviously. Indeed, there are numerous ways of programming the robot to "conceptualize" its environment that, while representationally biased or even radically false, are nonetheless (a) appropriately isomorphic and (b) reliably adaptive behavior-inducing. Such programs would be adaptive. (Schloss & Murray's commentary, para. 7)

Schloss & Murray are apparently imagining something like this: First the roboticist writes a program that captures all the relevant information in a behavioral "map" – and, to make the software development easier, all the nodes and action-representations are given *true* labels ("cliff" means *cliff* and "wall" means *wall* and "go left" means *go left*, etc.); and then, once the system is up and running and well tested, the roboticist goes back and systematically replaces "cliff" with "street" and "go left" with "jump" and so forth, for all the terms in the program. *Now*, it seems, the robot believes that when it reaches the street it should jump, where before it believed that when it reached the cliff it should go left – but since the "isomorphism" is preserved, it actually turns left when approaching the cliff, just as before – it is like the nearsighted Mr. Magoo, only more so! All its "false beliefs" conspire to keep it out of harm's way. But, as Quine (among others) observed, what the nodes mean, what content they actually have, is not determined by their labels, but by their myriad connections with each other and the world. The robot still has mainly true beliefs, but they are misleadingly "expressed" in the imagined internal labels. We think it is failure to appreciate this point that underlies much of the skepticism about the force of our default presumption. The explanation of the behavioral success of any successful organism must be in terms of how its sense organs *inform* it about its behavioral

environment. Misinformation can only "work" against a broad background of information.

A final example: Suppose people started saying, to everybody they encountered, "You're the most wonderful person I've met!" Perhaps initially this would have a benign effect, perking everyone up a little, but of course the effect would soon fade and the utterance would become the one-word synonym for "hello": "urthemoswunnerfulpersonimet." Why? Because utterances can only mean, in the long run, what their hearers *take* them to mean, and when utterers can no longer reasonably expect their hearers to take them to mean what their words "literally" mean, they can no longer have the intention of communicating by those words what the words used to mean, and then the words can no longer mean what they used to mean ("literally"). There is no way of divorcing what the subject believes, overall, from how the subject acts, so if an internal "danger to the left!" warning reliably leads the animal to jump left, not right, then the meaning of "left" and "right" in the animal's representation system must have reversed – *or* it must have inverted its "policies" somehow. So for evolution to discover a move, a design, that reliably misleads an organism (in an adaptive direction) it must be that the organism for one reason or another cannot make the Quinean adjustment, or it is evolutionarily cheaper, more robust, for the organism to actually lie to itself than to make the policy adjustments that would do the adaptive thing, given the truth about the situation.

R9. The "illusion of conscious will"?

We had initially hoped to devote space in the target article to belief in free will as a candidate for adaptive misbelief, but the topic is huge and space limitations obliged us to postpone it altogether, so we are pleased that **Mishara & Corlett** and **Randolph-Seng** raise the issue. As in our treatment of the "user-illusion" (see earlier, on **Zawidzki**), we think that there is a strong case to be made that this is best seen not as a useful falsehood, an enabling *myth* that we expose at our peril, but rather, as simply an important *true* belief, once it is properly unpacked and laundered of obsolete connotations.

Some (e.g., Blackmore 1999; Crick 1994; Wegner 2002) have argued that science has shown that we don't have free will. Others are *compatibilists* (e.g., Dennett 1984b; 2003a; Fischer 1994; Fischer & Ravizza 1998; Frankfurt 1988; Mele 1995). Dennett, for instance, has argued that although there are varieties of free will that science has plausibly shown not to exist, there are others that are unscathed, and they are the varieties that matter. Belief in them is indeed crucial to our mental health (to put it crudely) but these are true beliefs, compatible with what science has discovered, and is likely to discover, about the mechanisms of human choice. Does what one believes about the reality of free will make a discernible difference? Vohs and Schooler (2008) show that students who read a passage (from Crick 1994) assuring them that free will is a myth are more likely to cheat in a subsequent opportunity to win money. Like **Dweck**'s results, this finding might motivate a policy of deliberate myth-making – to try to preserve whatever shreds of responsibility remain in the wake of scientific self-knowledge – but since myth-maintenance is probably a losing battle even in the short run, for the reasons we have reviewed, a more stable policy might be to wean ourselves from the brittle traditional concepts; so that Crick's message turns into a socially bland observation about the emptiness of an obsolete concept, not a subversive blow to the integrity of our self-image as responsible agents. The fact that this healthy perspective is a hard sell, perennially challenged by the all too obvious *intuition* that "real" free will requires something like a miracle, may be indirect evidence that we are not just "natural-born dualists" (Bloom 2004) but natural-born believers in incompatibilist versions of free will as well. Such (false) beliefs may indeed have been adaptive in the past, enabling our ancestors to face life's decisions unburdened by misbegotten worries about causation and fatalism, but that does not make them necessary for mental health or effectiveness today.

R10. Conclusion

What is an adaptive misbelief? In essence, it is a false belief that has a recurrently positive effect on the reproductive fitness of its consumers. (Of course, for better or worse we conflated adapt*ive* with adapt*ed* or "evolved" in our target article; so false beliefs that *were* adaptive in the evolutionary past, but are not so nowadays, were of equal interest to us.) Let us briefly recap each of these features. First, an adaptive misbelief must be a bona fide belief. It cannot be merely an alief, and it cannot be merely a pragmatic acceptance reflecting a judicious policy for action. Second, an adaptive misbelief must be false, at least in part (it must at least *exaggerate* the truth). It cannot have morphed into a mere metaphor that no longer means what it would have to mean to be false (as in the case of free will, the self's user illusion and the cases of "content erosion" we have discussed).

Third, an adaptive misbelief must be adaptive (or have *been* adaptive, in the case of adapt*ed* misbelief – that pesky conflation again). Moreover, it must be adaptive for its consumers – lies that are adaptive for misinformants but harmful to the misinformed don't count, nor do parasitic misbeliefs that evolve simply because they can evolve (see Dennett & McKay 2006). Adaptive misbeliefs can't just represent the tolerated outputs of adaptive systems, by-products that are carried along for the ride despite being useless or even harmless. And they can't reflect the wholesale failures of internal mechanisms to carry out their naturally selected functions – at least not directly (we leave open here the possibility of naturally selected doxastic shear pins). Their effects must be recurrently positive – not lucky one-offs as in Stich's (1990) case of "Harry." Finally, their positive effects must be biologically beneficial, not just (or not necessarily) psychologically beneficial: they must enhance the reproductive fitness of their consumers. The mechanism of inheritance, however, can be genetic or cultural (natural selection can operate via either channel, as we remind **Wilks**).

We identified positive illusions as the best candidates for adaptive misbelief. In doing so we did not seek to undermine the "default presumption" that true belief is adaptive. Although we remain open to the possibility of adaptive misbelief, our position is that misbelief will, for the most part, lead to costly missteps: Misbelief can be

adaptive only against a broad background of true belief. Some commentators (e.g., **Dweck**, **Wilson & Lynn**) suggest that we held religious beliefs to a stricter standard than positive illusions, and we accept that, pending further research, religious beliefs may represent an important cultural subspecies of evolved misbelief. But as **Ainslie** notes, we are the endlessly tinkering, self-prospecting species, and such myths as we – or natural selection – may devise for ourselves are vulnerable to our insatiable curiosity. The tragic abyss that now opens before us is familiar from hundreds of tales, from Eve's fatal apple and Pandora's box, through Faust's bargain, Bluebeard's Castle and Dostoyevsky's Grand Inquisitor: What price knowledge? Are we better off not knowing the truth? This question presupposes, implausibly, that we might have a choice, but it is probably too late in the day to opt for blissful ignorance. Science has seen to that, letting the cat out of the bag (to cite one more version of the tale). Now that skepticism is ubiquitous, "practically realistic" myths (Wilson & Lynn; see also Wilson 2002) are in danger of losing whatever effectiveness accounts for their preservation up to now. The frequency in the social world of recursive meta-examinations (such as this article, along with thousands of others) has changed the selective pressures acting on such myths, making their extinction more likely, and not at all incidentally jeopardizing whatever benefits to us, their vectors, these myths may have provided.

References

[The letters "a" and "r" before author's initials stand for target article and response references, respectively]

Abbey, A. (1982) Sex differences in attributions for friendly behavior: Do males misperceive females' friendliness? *Journal of Personality and Social Psychology* 42:830–38. [MGH, aRTM]

Ackerman, J. M., Becker, D. V., Mortensen, C. R., Sasaki, T., Neuberg, S. L. & Kenrick, D. T. (2009) A pox on the mind: Disjunction of attention and memory in processing physical disfigurement. *Journal of Experimental Social Psychology* 45:478–85. [JMA]

Ackerman, J. M., Griskevicius, V. & Li, N. (submitted) Let's get serious: Communicating commitment in romantic relationship formation. [JMA]

Ackerman, J. M. & Kenrick, D. T. (2008) The costs of benefits: Help-refusals highlight key trade-offs of social life. *Personality and Social Psychology Review* 12:118–40. [JMA]

Ackerman, J. M. & Kenrick, D. T. (2009) Cooperative courtship: Helping friends raise and raze relationship barriers. *Personality and Social Psychology Bulletin* 35:1285–300. [JMA]

Ackerman, J. M., Shapiro, J. R., Neuberg, S. L., Kenrick, D. T., Becker, D. V., Griskevicius, V., Maner, J. K. & Schaller, M. (2006) They all look the same to me (unless they're angry): From out-group homogeneity to out-group heterogeneity. *Psychological Science* 17:836–40. [JMA]

Adams, C. D. & Dickinson, A. (1981) Actions and habits: Variations in associative representations during instrumental learning. In: *Information processing in animals: Memory mechanisms*, ed. N. E. Spear & R. R. Miller, pp. 143–66. Erlbaum. [ALM]

Ainslie, G. (2001) *Breakdown of will.* Cambridge University Press. [GA]

Ainslie, G. (2005) Précis of *Breakdown of will. Behavioral and Brain Sciences* 28(5):635–73. [GA]

Ainslie, G. (2010) Procrastination, the basic impulse. In: *The thief of time: Philosophical essays on procrastination*, ed. C. Andreou & M. White. pp. 11–27. Oxford University Press. [GA]

Akins, K. (1996) Of sensory systems and the "aboutness" of mental states. *The Journal of Philosophy* 93(7):337–72. [aRTM, JS]

Alcorta, C. S. & Sosis, R. (2005) Ritual, emotion, and sacred symbos. *Human Nature* 16(4):323–59. [JB]

Alea, N. & Bluck, S. (2003) "Why are you telling me that?" A conceptual model of the social function of autobiographical memory. *Memory* 11(2):165–78. [JS]

Alexander, R. D. (1979) *Darwinism and human affairs*. University of Washington Press. [aRTM]

Alexander, R. D. (1987) *The biology of moral systems*. Aldine de Gruyter. [aRTM]

Alicke, M. D. (1985) Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology* 49:1621–30. [JDB, aRTM]

Alloy, L. B. & Abramson, L. Y. (1988) Depressive realism: Four theoretical perspectives. In: *Cognitive processes in depression*, ed. L. B. Alloy, pp. 223–65. Guilford Press. [aRTM]

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

Anderson, J. R. & Schooler, L. J. (2000) The adaptive nature of memory. In: *The Oxford handbook of memory*, ed. E. Tulving & F. I. M. Craik, pp. 557–70. Oxford University Press. [PB, JS]

Aristotle (1985) *Nicomachean Ethics*. trans. T. Irwin. Hackett. [OF]

Armor, D. A. & Taylor, S. E. (1998) Situated optimism: Specific outcome expectancies and self-regulation. In: *Advances in experimental social psychology, vol. 30*, ed. M. Zanna, pp. 309–79. Academic Press. [DD]

Aronson, E. (1969) A theory of cognitive dissonance: A current perspective. In: *Advances in experimental social psychology, vol. 4*, ed. L. Berkowitz, pp. 1–34. Academic Press. [NLG]

Asch, S. E. (1956) Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs* 70 (Whole No. 416). [VJK]

Atran, S. (2004) *In Gods we trust: The evolutionary landscape of religion*. Oxford University Press. [aRTM, KT-K]

Atran, S. & Norenzayan, A. (2004) Religion's evolutionary landscape: Counterintuition, commitment, compassion, communion. *Behavioral and Brain Sciences* 27:713–70. [aRTM, DS]

Avital, E. & Jablonka, E. (2000) *Animal traditions: Behavioural inheritance in evolution*. Cambridge University Press. [rRTM]

Baer, D. & McEachron, D. L. (1982) A review of selected sociobiological principles: Application to hominid evolution I: The development of group structure. *Journal of Social and Biological Structures* 5:69–90. [JMA]

Baker, S. T., Murray, K. & Hood, B. M. (2009) Children's expectations about weight and speed of falling objects; the younger the judge the better? Poster presented at the Society for Research into Children's Development meeting, Denver, CO, April 2–4, 2009. [NLG]

Balcetis, E. & Dunning, D. (2006) See what you want to see: Motivational influences on visual perception. *Journal of Personality and Social Psychology* 91:612–25. [BR-S]

Bandura, A. (1989) Human agency in social cognitive theory. *American Psychologist* 44:1175–84. [aRTM]

Barber, L. (2009) God's banker: Interview of Stephen Green. In: *The Financial Times*, June 27/28, 2009, "Life and Arts" section, p. 3. (Online posting June 26, 2009. Available at: http://www.ft.com/cms/s/2/224b507c-61de-9e03-00144feabdc0.html) [VJK]

Bargh, J. A. (2008) Free will is un-natural. In: *Are we free? The psychology of free will*, ed. J. Baer, J. Kaufman & R. Baumeister, pp. 128–54. Oxford University Press. [BR-S]

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* 71:230–44. [aRTM]

Bargh, J. A. & Earp, B. (2009) The will is caused, not "free." *Dialogue: Newsletter of the Society for Personality and Social Psychology* 24:13, 15. [BR-S]

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K. & Troetschel, R. (2001) Automating the will: Nonconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology* 81:1014–27. [AN]

Barnier, A. J., Sutton, J., Harris, C. B. & Wilson, R. A. (2008) A conceptual and empirical framework for the social distribution of cognition: the case of memory. *Cognitive Systems Research* 9(1):33–51. [JS]

Barrett, J. L. (2000) Exploring the natural foundations of religion. *Trends in Cognitive Sciences* 4(1):29–34. [aRTM, KT-K]

Barry, C. T., Frick, P. J. & Killian, A. L. (2003) The relation of narcissism and self-esteem to conduct problems in children: A preliminary investigation. *Journal of Clinical and Adolescent Psychology* 32:139–52. [DSW]

Bartlett, F. C. (1932) *Remembering. A study in experimental and social psychology*. Cambridge University Press. [PB]

Baumeister, R. F. (1989) The optimal margin of illusion. *Journal of Social and Clinical Psychology* 8:176–89. [JDB, arRTM]

Baumeister, R. F. (2001) Violent pride: Do people turn violent because of self hate or self love? *Scientific American* 284(4):96–101. [DSW]

Baumeister, R. F. (2008) Free will in scientific psychology. *Perspectives on Psychological Science* 3:14–19. [BR-S]

Baumeister, R. F., Campbell, J. D., Krueger, J. I. & Vohs, K. D. (2003) Does high self-esteem cause better performance, interpersonal success, happiness, or healthier life styles? *Psychological Science in the Public Interest* 4:1–44. [DSW]

Baumeister, R. F., Masicampo, E. J. & DeWall, C. N. (2009) Prosocial benefits of feeling free: Disbelief in free will increases aggression and reduces helpfulness. *Personality and Social Psychology Bulletin* 35:260–68. [BR-S]

Baumgartner, T., Lutz, K., Schmidt, C. F. & Jäncke, L. (2006) The emotional power of music: How music enhances the feeling of affective pictures. *Brain Research* 1075(1):151–64. [JB]

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. [aRTM]

Bayne, T. & Fernández, J. (2009) Delusion and self-deception: Mapping the terrain. In: *Delusion and self-deception: Affective and motivational influences on belief formation*, ed. T. Bayne & J. Fernández, pp. 1–21. Psychology Press. [aRTM]

Bayne, T. & Pacherie, E. (2005) In defence of the doxastic conception of delusions. *Mind and Language* 20(2):163–88. [aRTM]

Becker, D. V., Mortensen, C. R., Ackerman, J. M., Shapiro, J. R., Anderson, U. S., Sasaki, T., Maner, J. K., Neuberg, S. L. & Kenrick, D. T. (submitted) Self-protection and revenge-mindedness modulate detection of enemy insignia. [JMA]

Beja-Pereira, A., Luikart, G., England, P. R., Bradley, D. G., Jann, O. C., Bertorelle, G., Chamberlain, A. T., Nunes, T. P., Metodiev, S., Ferrand, N. & Erhardt, G. (2003) Gene-culture coevolution between cattle milk protein genes and human lactase genes. *Nature Genetics* 35:311–13. [aRTM]

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

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):4315–23. [aRTM]

Bentall, R. P. & Kaney, S. (1996) Abnormalities of self-representation and persecutory delusions: A test of a cognitive model of paranoia. *Psychological Medicine* 26:1231–37. [arRTM]

Bering, J. M. (2002) The existential theory of mind. *Review of General Psychology* 6:3–24. [aRTM]

Bering, J. M. (2006) The folk psychology of souls. *Behavioral and Brain Sciences* 29:453–98. [aRTM]

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(1/2):118–42. [aRTM, DDPJ]

Bering, J. M., McLeod, K. A. & Shackelford, T. K. (2005) Reasoning about dead agents reveals possible adaptive trends. *Human Nature* 16:360–81. [aRTM, AN]

Bernstein, D. M. & Loftus, E. F. (2009) How to tell if a particular memory is true or false. *Perspectives on Psychological Science* 4:370–74. [JS]

Berrios, G. E. (1991) Delusions as "wrong beliefs": A conceptual history. *British Journal of Psychiatry* 159:6–13. [aRTM]

Bertamini, M. & Parks, T. E. (2005) On what people know about images on mirrors. *Cognition* 98:85–104. [MB]

Beyerstein, B., Sampson, W. I., Stojanovic, Z. & Handel, J. (2007) Can mind conquer cancer? In: *Tall tales about the mind and brain: Separating fact from fiction*, ed. S. Dallea Sala, pp. 440–60. Oxford University Press. [DSW]

Bianchi, I. Savardi, U. & Bertamini, M. (2008) Estimation and representation of head size (People overestimate their own head, evidence starting from the 15th century). *British Journal of Psychology* 99:513–31. [MB]

Billig, M. & Tajfel, J. (1973) Social categorization and similarity in intergroup behavior. *European Journal of Social Psychology* 3:27–52. [JDB]

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. [aRTM]

Binswanger, L. (1965) *Wahn*. Neske. [ALM]

Bjork, E. L. & Bjork, R. A. (1988) On the adaptive aspects of retrieval failure in autobiographical memory. In: *Practical aspects of memory: Current research and issues*, ed. M. Gruneberg, P. Morris & R. Sykes, pp. 283–88. Wiley. [JS]

Blackmore, S. (1999) *The meme machine*. Oxford University Press. [rRTM]

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* 78:246–63. [aRTM]

Bloom, P. (2004) *Descartes' baby: How child development explains what makes us human*. Arrow Books. [GA, JRL, arRTM]

Bloom, P. (2005) Is God an accident? *Atlantic Monthly* 296:105–12. [aRTM]

Bloom, P. (2007) Religion is natural. *Developmental Science* 10(1):147–51. [aRTM]

Boden, M. (1984) Animal perception from an Artificial Intelligence viewpoint. In: *Minds, machines and evolution*, ed. C. Hookway, pp. 153–74. Cambridge University Press. [aRTM]

Bouchard, T. J. (1994) Genes, environment, and personality. *Science* 264:1700–701. [ETC]

Bowles, S., Choi, J.-K. & Hopfensitz, A. (2003) The co-evolution of individual behaviours and social institutions. *Journal of Theoretical Biology* 223(2):135–47. [aRTM]

Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. (2003) The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences USA* 100(6):3531–35. [aRTM]

Boyer, P. (1994) *The naturalness of religious ideas: A cognitive theory of religion*. University of California Press. [DS]

Boyer, P. (2001) *Religion explained: The evolutionary origins of religious thought*. Basic Books. [aRTM, KT-K]

Boyer, P. (2003) Religious thought and behaviour as by-products of brain function. *Trends in Cognitive Sciences* 7(3):119–24. [aRTM]

Boyer, P. (2008a) Evolutionary economics of mental time travel. *Trends in Cognitive Sciences* 12(6):219–24. [JS]

Boyer, P. (2008b) Religion: Bound to believe? *Nature* 455(23):1038–39. [aRTM]

Boyer, P. (2009) What are memories for? Functions of recall in cognition and culture. In: *Memory in mind and culture*, ed. P. Boyer & J. Wertsch, pp. 3–28. Cambridge University Press. [JS]

Bozzi, P. (1958) Analisi fenomenologica del moto pendolare armonico. *Rivista di Psicologia* 52:281–302. [MB]

Brainerd, C. J. & Reyna, V. F. (2005) *The science of false memory*. Oxford University Press. [PB]

Bratman, M. E. (1992) Practical reasoning and acceptance in a context. *Mind* 101(401):1–15. [KF, aRTM]

Breen, N., Caine, D. & Coltheart, M. (2001) Mirrored-self misidentification: Two cases of focal onset dementia. *Neurocase* 7:239–54. [aRTM]

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. [aRTM]

Brewer, M. B. (1979) In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin* 86:307–24. [JDB]

Brown, J. D. (1986) Evaluations of self and others: Self-enhancement biases in social judgments. *Social Cognition* 4:353–76. [JDB]

Brown, J. D. (1991) Accuracy and bias in self-knowledge. In: *Handbook of social and clinical psychology: The health perspective*, ed. C. R. Snyder & D. F. Forsyth, pp. 158–78. Pergamon Press. [JDB]

Brown, J. D. (1998) *The self*. McGraw-Hill. [JDB]

Brown, J. D. (2003) The self-enhancement motive in collectivistic cultures: The rumors of my death have been greatly exaggerated. *Journal of Cross-Cultural Psychology* 34:603–605. [JDB]

Brown, J. D. (2007) Positive illusions. In: *Encyclopedia of social psychology*, ed. R. Baumeister & K. Vohs, pp. 615–17. Sage. [JDB]

Brown, J. D., Cai, H., Oakes, M. A. & Deng, C. (2009) Cultural similarities in self-esteem functioning: East is East and West is West, but sometimes the twain do meet. *Journal of Cross-Cultural Psychology* 40:140–57. [JDB]

Brown, J. D. & Kobayashi, C. (2002) Self-enhancement in Japan and America. *Asian Journal of Social Psychology* 5:145–67. [JDB]

Brown, J. D. & Kobayashi, C. (2003) Motivation and manifestation: The cross-cultural expression of the self-enhancement motive. *Asian Journal of Social Psychology* 6:85–88. [JDB]

Brüne, M. (2001) De Clerambault's syndrome (erotomania) in an evolutionary perspective. *Evolution and Human Behavior* 22(6):409–15. [aRTM]

Brüne, M. (2003a) Erotomania (De Clerambault's Syndrome) revisited – Clues to its origin from evolutionary theory. In: *Advances in psychology research, vol. 21*, ed. S. P. Shohov, pp. 185–212. Nova Science. [aRTM]

Brüne, M. (2003b) Erotomanic stalking in evolutionary perspective. *Behavioral Sciences and the Law* 21(1):83–88. [aRTM]

Bruner, J. S. (1957) On perceptual readiness. *Psychological Review* 64:123–52. [BR-S]

Buckman, R. & Sabbagh, K. (1993) *Magic or medicine: An investigation of healing and healers*. MacMillan. [aRTM]

Bulbulia, J. (2004a) Religious costs as adaptations that signal altruistic intention. *Evolution and Cognition* 10(1):19–38. [JB]

Bulbulia, J. (2004b) The cognitive and evolutionary psychology of religion. *Biology and Philosophy* 19:655–86. [aRTM]

Bulbulia, J. (2009) Religiosity as mental time travel: Cognitive adaptations for religious behavior. In: *The believing primate: Scientific, philosophical and theological perspectives on the evolution of religion*, ed. J. Schloss & M. Murray, pp. 44–75. Oxford University Press. [JB]

Bulbulia, J. (in press) Coordination by sacred cues :|. *Journal for the Study of Religion, Nature, Culture*. [JB]
Bulbulia, J. & Mahoney, A. (2008) Religious solidarity: The hand grenade experiment. *Journal of Cognition and Culture* 8:295–320. [JB]
Burger, J. M., Messian, N., Patel, S., del Prado, A. & Anderson, C. (2004) What a coincidence! The effects of incidental similarity on compliance. *Personality and Social Psychology Bulletin* 30:35–43. [MB]
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. [aRTM]
Buss, D. M. & Haselton, M. G. (2005) The evolution of jealousy: A response to Buller. *Trends in Cognitive Sciences* 9(11):506–507. [aRTM]
Buss, D. M., Haselton, M. G., Shackelford, T. K., Bleske, A. L. & Wakefield, J. C. (1998) Adaptations, exaptations, and spandrels. *American Psychologist* 53:533–48. [JRL]
Butler, P. V. (2000) Reverse Othello syndrome subsequent to traumatic brain injury. *Psychiatry: Interpersonal and Biological Processes* 63(1):85–92. [aRTM]
Byrne, C. C. & Kurland, J. A. (2001) Self-deception in an evolutionary game. *Journal of Theoretical Biology* 212:457–80. [ETC]
Cacioppo, J. T., Berntson, G. G., Sheridan, J. F. & McClintock, M. K. (2000) Multilevel integrative analyses of human behavior: Social neuroscience and the complementing nature of social and biological approaches. *Psychological Bulletin* 126:829–43. [BR-S]
Cai, H., Brown, J. D., Deng, C. & Oakes, M. A. (2007) Self-esteem and culture: Differences in cognitive self-evaluations or affective self-regard? *Asian Journal of Social Psychology* 10:162–70. [JDB]
Cai, H., Wu, Q. & Brown, J. D. (2009) Is self-esteem a universal need? Evidence from the People's Republic of China. *Asian Journal of Social Psychology* 12:104–20. [JDB]
Callebaut, W., Müller, G. B. & Newman, S. A. (2007) The organismic systems approach: Evo-devo and the streamlining of the naturalistic agenda. In: *Integrating evolution and development: From theory to practice*, ed. R. Sansom & R. N. Brandon, pp. 25–92. MIT Press. [TJW]
Camerer, C. (2003) *Behavioral game theory*. Princeton University Press. [aRTM]
Campbell, D. T. (1965) Ethnocentric and other altruistic motives. In: *Nebraska Symposium on Motivation*, pp. 283–311, ed. D. Levine. University of Nebraska Press. [JMA]
Caputo, D. D. & Dunning, D. (2005) What you don't know: The role played by errors of omission in imperfect self-assessments. *Journal of Experimental Social Psychology* 41:488–505. [DD]
Carruthers, P. (2009) How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences* 32(2):121–82. [TWZ]
Casati, R. (2008) The copycat solution to the shadow correspondence problem. *Perception* 37(4):495–503. [MB]
Cash, M. (2008) Thoughts and oughts. *Philosophical Explorations* 11(2):93–119. [TWZ]
Casperson, L. W. (1999) Head movement and vision in underwater-feeding birds of stream, lake, and seashore. *Bird Behavior* 13:31–46. [aRTM]
Cassia, V. M., Turati, C. & Simion, F. (2004) Can a non-specific preference for top-heavy patterns explain newborns' face preference? *Psychological Science* 15:379–83. [CSD]
Cauvin, J. (2000) *The birth of the Gods and the origins of agriculture*. Cambridge University Press. [AN]
Cavanagh, P. (2005) The artist as neuroscientist. *Nature* 434:301–307. [MB]
Chambers, J. R. & Windschitl, P. D. (2004) Biases in social comparison judgments: The role of nonmotivated factors in above-average and comparative-optimism effects. *Psychological Bulletin* 130:813–38. [JK]
Chomsky, N. (1965) *Aspects of the theory of syntax*. MIT Press. [YW]
Chow, Y. C., Dhillon, B., Chew, P. T. & Chew, S. J. (1990) Refractive errors in Singapore medical students. *Singapore Medical Journal* 31:472–73. [aRTM]
Churchland, P. (1987) Epistemology in the age of neuroscience. *Journal of Philosophy* 84(10):544–53. [JPS]
Cimpian, A., Arce, H., Markman, E. M. & Dweck, C. S. (2007) Subtle linguistic cues impact children's motivation. *Psychological Science* 18:314–16. [CSD]
Clark, A. (1994) Beliefs and desires incorporated. *Journal of Philosophy* 91:404–25. [JS]
Cohen, L. J. (1989) Belief and acceptance. *Mind* 98:367–89. [KF]
Cohen, L. J. (1992) *An essay on belief and acceptance*. Oxford University Press. [KF]
Cokely, E. T. & Feltz, A. (2009a) Adaptive variation in folk judgment and philosophical intuition. *Consciousness and Cognition* 18:355–57. [ETC]
Cokely, E. T. & Feltz, A. (2009b) Individual differences, judgment biases, and Theory-of-Mind: Deconstructing the intentional action side effect asymmetry. *Journal of Research in Personality* 43:18–24. [ETC]
College Board (1976–1977) *Student descriptive questionnaire*. Educational Testing Service, College Board. [JK]
Coltheart, M. (1996) Are dyslexics different? *Dyslexia* 2:79–81. [aRTM]
Coltheart, M. (2002) Cognitive neuropsychology. In: *Stevens' handbook of experimental psychology, vol. 4: Methodology in experimental psychology*, 3rd edition, ed. H. Pashler & J. Wixted, pp. 139–74. Wiley. [aRTM]
Coltheart, M. (2007) Cognitive neuropsychiatry and delusional belief. *Quarterly Journal of Experimental Psychology* 60:1041–1062. [RL]
Coltheart, M., Menzies, P. & Sutton, J. (in press) Abductive inference and delusional belief. In: *Delusion and confabulation: Overlapping or distinct psychopathologies of reality distortion*, ed. R. Langdon & M. Turner. Macquarie Monographs in Cognitive Science Series. Series editor: M. Coltheart. Psychology Press. [aRTM]
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. [aRTM]
Conrad, K. (1958) *Die beginnende Schizophrenie. Versuch einer Gestaltanalyse des Wahns*. Thieme. [ALM]
Conway, M. A. (2005) Memory and the self. *Journal of Memory and Language* 53:594–628. [JS]
Conway, M. A. & Pleydell-Pearce, C. W. (2000) The construction of autobiographical memories in the self-memory system. *Psychological Review* 107(2):261–88. [JS]
Corlett, P. R., Frith, C. D. & Fletcher, P. C. (2009a) From drugs to deprivation: A Bayesian framework for understanding models of psychosis. *Psychopharmacology* 206(4):515–30. [ALM]
Corlett, P. R., Honey, G. D. & Fletcher, P. C. (2007) From prediction error to psychosis: Ketamine as a pharmacological model of delusions. *Journal of Psychopharmacology* 21(3):238–52. [ALM]
Corlett, P. R., Krystal, J. K., Taylor, J. R. & Fletcher, P. C. (2009b) Why do delusions persist? *Frontiers in Human Neuroscience 3*. Available at: http://www.frontiersin.org/humanneuroscience/paper/10.3389/neuro.09/012.2009/html/ [rRTM, ALM]
Coser, L. A. (1956) *The functions of social conflict*. Free Press. [JMA]
Cosmides, L. & Tooby, J. (2000) Consider the source: The evolution of adaptations for decoupling and metarepresentation. In *Metarepresentations: A multidisciplinary perspective*, ed. D. Sperber, pp. 53–115. Oxford University Press. [PB]
Cottrell, C. A. & Neuberg, S. L. (2005) Different emotional reactions to different groups: A sociofunctional threat-based approach to "prejudice." *Journal of Personality and Social Psychology* 88:770–89. [JMA]
Coyne, J. C. & Kitcher, P. (2007) Letter to the editor re: Fodor. *London Review of Books*, November 15, p. 29. [aRTM]
Coyne, J. C., Pajak, T. F., Harris, J., Konski, A., Movsas, B. & Ang, K. (2007) Emotional well-being does not predict survival in head and neck cancer patients: A radiation therapy oncology group study. *Cancer* 110:2568–75. [DSW]
Crick, F. (1994) *The astonishing hypothesis: The scientific search for the soul*. Scribner. [rRTM]
Cromer, A. (1993) *Uncommon sense: The heretical nature of science*. Oxford University Press. [JPS]
Cronk, L. (1994) Evolutionary theories of morality and the manipulative use of signals. *Zygon* 4:117–35. [DDPJ]
Cross, P. (1977) Not can but will college teaching be improved? *New Directions for Higher Education* 17:1–15. [aRTM]
Currie, G. (2000) Imagination, delusion and hallucinations. *Mind and Language* 15:168–83. [aRTM]
Currie, G. & Jureidini, J. (2001) Delusion, rationality, empathy: Commentary on Davies et al. *Philosophy, Psychiatry, and Psychology* 8(2–3):159–62. [aRTM]
David, A. S. (1999) On the impossibility of defining delusions. *Philosophy, Psychiatry, and Psychology* 6(1):17–20. [RL, aRTM]
David, A. S. & Halligan, P. W. (1996) Editorial. *Cognitive Neuropsychiatry* 1:1–3. [aRTM]
Davidson, D. (1994) Radical interpretation interpreted. In: *Philosophical perspectives, vol. 8: Logic and Language*, ed. J. E. Tomberlin, pp. 121–28. Ridgeview. [aRTM]
Davidson, D. (2001) *Inquiries into truth and interpretation*, 2nd edition. Clarendon Press. [aRTM]
Davies, M. & Coltheart, M. (2000) Introduction: Pathologies of belief. In: *Pathologies of belief*, ed. M. Coltheart & M. Davies, pp. 1–46. Blackwell. [aRTM, ALM]
Davies, M., Coltheart, M., Langdon, R. & Breen, N. (2001) Monothematic delusions: Towards a two-factor account. *Philosophy, Psychiatry, and Psychology* 8(2–3): 133–58. [RL, aRTM]
Daw, N. D., Niv, Y. & Dayan, P. (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience* 8(12):1704–11. [ALM]

Dawkins, R. (1982) *The extended phenotype*. Freeman/Oxford University Press. [JRL, aRTM]
Dawkins, R. (1986) *The blind watchmaker*. W. W. Norton. [aRTM]
Dawkins, R. (1989) *The selfish gene*, second edition. Oxford University Press. [RGM, YW]
Dawkins, R. (1996) *The blind watchmaker: Why the evidence of evolution reveals a universe without design*. W.W. Norton. [JPS]
Dawkins, R. (2006a) *The God delusion*. Bantam Press. [aRTM]
Dawkins, R. (2006b) *The selfish gene: 30th anniversary edition*. Oxford University Press. [aRTM]
Deacon, T. (1997) *The symbolic species: The coevolution of language and the brain*. Norton. [rRTM]
Dean, C. & Surtees, P. G. (1989) Do psychological factors predict survival in breast cancer? *Journal of Psychosomatic Research* 33(5):561–69. [aRTM]
DeBruine, L. M. (2002) Facial resemblance enhances trust. *Proceedings of the Royal Society of London B* 269:1307–12. [MB]
Deeley, P. Q., Identity, D. & Identity, R. P. (2003) Social, cognitive, and neural constraints on subjectivity and agency: Implications. *Project Muse* 10(2):161–67. Available at: muse.jhu.edu [JB]
Dennett, D. C. (1971) Intentional systems. *Journal of Philosophy* 68(2):87–106. [aRTM]
Dennett, D. C. (1978) *Brainstorms: Philosophical essays on mind and psychology*. MIT Press/A Bradford Book. [arRTM, TWZ]
Dennett, D. C. (1981) Making sense of ourselves (Reply to S. Stich, "Dennett on Intentional Systems"). *Philosophical Topics* 12:63–81. Reprinted in: *Mind, brain, and function: Essays in the philosophy of mind*, ed. J. I. Biro & R. W. Shahan, pp. 63–81. University of Oklahoma Press, 1982. [rRTM]
Dennett, D. C. (1982) Beyond belief. In: *Thought and object: Essays on intentionality*, ed. A. Woodfield, pp. 1–96. Clarendon Press. [aRTM]
Dennett, D. C. (1984a) A route to intelligence: Oversimplify and self-monitor. Available at: http://ase.tufts.edu/cogstud/papers/oversimplify.pdf. [rRTM]
Dennett, D. C. (1984b) *Elbow room*. MIT Press. [rRTM]
Dennett, D. C. (1985) Why believe in belief? (Review of S. Stich, *From folk psychology to cognitive science: The case against belief*). *Contemporary Psychology* 30:949. [rRTM]
Dennett, D. C. (1987) *The intentional stance*. MIT Press. [aRTM, TWZ]
Dennett, D. C. (1990a) Attitudes about ADHD: Some analogies and aspects. In: *ADHD: Attention Deficit Hyperactivity Disorders*, ed. K. Conners & M. Kinsbourne, pp. 11–16. MMV Medizin Verlag. [aRTM]
Dennett, D. C. (1990b) The interpretation of texts, people and other artifacts. *Philosophy and Phenomenological Research* 50(Supplement):177–94. [aRTM]
Dennett, D. (1991) *Consciousness explained*. Little, Brown. [JB, rRTM, JS, TWZ]
Dennett, D. C. (1995a) *Darwin's dangerous idea: Evolution and the meanings of life*. Simon & Schuster/Penguin. [arRTM]
Dennett, D. C. (1995b) How to make mistakes. In: *How things are*, ed. J. Brockman & K. Matson, pp. 137–44. William Morrow. [aRTM]
Dennett, D. C. (1998) *Brainchildren – Essays on designing minds*. MIT Press/ Bradford Books and Penguin. [aRTM]
Dennett, D. C. (2003a) *Freedom evolves*. Viking Press. [rRTM]
Dennett, D. C. (2003b) The Baldwin Effect: A crane, not a skyhook. In: *Evolution and learning: The Baldwin Effect reconsidered*, ed. B. H. Weber & D. J. Depew, pp. 60–79. MIT Press/Bradford Books. [rRTM]
Dennett, D. C. (2005) Show me the science. *The New York Times*, August 28, p. 11. [aRTM]
Dennett, D. C. (2006a) *Breaking the spell: Religion as a natural phenomenon*. Viking/Penguin Press. [arRTM, JPS, DSW, TWZ]
Dennett, D. C. (2006b) Thank Goodness! *Edge: The Third Culture*, November 3, 2006. http://edge.org/3rd_culture/dennett06/dennett06_index.html. [aRTM]
Dennett, D. C. (2007) Letter to the Editor re Fodor. *London Review of Books*, November 15, 2007, p. 29. [aRTM]
Dennett, D. C. (2008) Fun and games in Fantasyland. *Mind and Language* 23(1):25–31. [aRTM]
Dennett, D. C. (2009) Darwin's "strange inversion of reasoning." *Proceedings of the National Academy of Sciences USA* 106 (Suppl. 1):10061–65. [rRTM]
Dennett, D. C. (forthcoming) *Science and religion: Are they compatible? A debate with Alvin Plantinga*. Oxford University Press. [rRTM]
Dennett, D. C. & McKay, R. T. (2006) A continuum of mindfulness (Commentary on Mesoudi et al). *Behavioral and Brain Sciences* 29(4):353–54. [rRTM]
Denrell, J. (2005) Why most people disapprove of me: Experience sampling in impression formation. *Psychological Review* 112:951–78. [DD]
Denton, K. & Zarbatany, L. (1996) Age differences in support processes in conversations between friends. *Child Development* 67:1360–73. [DLK]
Dickinson, A. & Shanks, D. (1995) Instrumental action and causal representation. In: *Causal cognition: A multidisciplinary debate*, ed. D. Sperber, D. Premack & A. J. Premack, pp. 5–25. Oxford University Press. [ALM]
Dijksterhuis, A., Chartrand, T. L. & Aarts, H. (2007) Effects of priming and perception on social behavior and goal pursuit. In: *Social psychology and the unconscious: The automaticity of higher mental processes*, ed. J. A. Bargh, pp. 51–131. Psychology Press. [aRTM]
Dijksterhuis, A., Preston, J., Wegner, D. M. & Aarts, H. (2008) Effects of subliminal priming of self and God on self-attribution of authorship for events. *Journal of Experimental Social Psychology* 44:2–9. [rRTM, AN]
Dillard, A. J., McCaul, K. D. & Klein, W. M. P. (2006) Unrealistic optimism in smokers: Implications for smoking myth endorsement and self-protective motivation. *Journal of Health Communication* 11:93–102. [DD]
Dukas, R. (1999) Costs of memory: ideas and predictions. *Journal of theoretical biology* 197(1):41–50. [PB]
Dunning, D. (2005) *Self-insight: Roadblocks and detours on the path to knowing thyself*. Psychology Press. [DD]
Dunning, D., Heath, C. & Suls, J. M. (2004) Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest* 5:69–106. [JDB, DD, CSD]
Dunning, D., Meyerowitz, J. A. & Holzberg, A. D. (1989) Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology* 57:1082–90. [JK]
Duntley, J. & Buss, D. M. (1998) Evolved anti-homicide modules. Paper presented at the Human Behavior and Evolution Society Conference, Davis, CA, July 1998. [DD, aRTM]
Durkheim, E. (1912/1995) *The elementary forms of religious life*. Free Press. (Original work published in 1912). [KT-K]
Dweck, C. S. (1999) *Self-theories: Their role in motivation, personality and development*. Psychology Press. [aRTM]
Easton, J. A., Schipper, L. D. & Shackelford, T. K. (2007) Morbid jealousy from an evolutionary psychological perspective. *Evolution and Human Behavior* 28:399–402. [aRTM]
Eisenhardt, D. & Menzel, R. (2007) Extinction learning, reconsolidation and the internal reinforcement hypothesis. *Neurobiology of Learning and Memory* 87(2):167–73. [ALM]
Ellis, A. W. & Young, A. W. (1988) *Human cognitive neuropsychology*. Erlbaum. [aRTM]
Ellis, H. D. (2003) Book review: Uncommon psychiatric syndromes. *Cognitive Neuropsychiatry* 8(1):77–79. [aRTM]
Engel, P. (1998) Believing, holding true, and accepting. *Philosophical Explorations* 1:140–51. [KF]
Engel, P., ed. (2000) *Believing and accepting*. Kluwer Academic. [KF]
Enoch, M. D. & Ball, H. N. (2001) *Uncommon psychiatric syndromes*, 4th edition. Arnold. [aRTM]
Fehr, E. & Fischbacher, U. (2003) The nature of human altruism. *Nature* 425:785–91. [aRTM, AN]
Fehr, E. & Fischbacher, U. (2004) Social norms and human cooperation. *Trends in Cognitive Sciences* 8(4):185–90. [aRTM]
Fehr, E., Fischbacher, U. & Gaechter, S. (2002) Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature* 13:1–25. [aRTM]
Fehr, E. & Gaechter, S. (2002) Altruistic punishment in humans. *Nature* 415:137–40. [aRTM]
Feinberg, T. E. (2001) *Altered egos: How the brain creates the self*. Oxford University Press. [aRTM]
Feinberg, T. E. & Shapiro, R. M. (1989) Misidentification-reduplication and the right hemisphere. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology* 2(1):39–48. [aRTM]
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: *Mathematical evolutionary theory*, ed. M. W. Feldman, pp. 145–73. Princeton University Press. [aRTM]
Feltz, A. & Cokely, E. T. (2008) The fragmented folk: More evidence of stable individual differences in moral judgments and folk intuitions. In: *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, ed. B. C. Love, K. McRae & V. M. Sloutsky, pp. 1771–76. Cognitive Science Society. [ETC]
Feltz, A. & Cokely, E. T. (2009) Do judgments about freedom and responsibility depend on who you are? Personality differences intuitions about compatibilism and incompatibilism. *Consciousness and Cognition* 24:342–50. [ETC]
Fernandes, M., Ross, M., Wiegand, M. & Schryer, E. (2008) Are the memories of older adults positively biased? *Psychology and Aging* 23(2):297–306. [JS]
Festinger, L. (1954) A theory of social comparison processes. *Human Relations* 7:117–40. [JK]
Festinger, L., Reicken, H. & Schachter, S. (1956) *When prophecy fails*. New York Press. [JB]
Fetchenhauer, D. & Dunning, D. (2009) Do people trust too much or too little? *Journal of Economic Psychology* 30:263–76. [DD]

Fetchenhauer, D. & Dunning, D. (in press) Why so cynical? Asymmetric feedback underlies misguided skepticism in the trustworthiness of others. *Psychological Science*. [DD]
Fischer, J. M. (1994) *The metaphysics of free will*. Blackwell. [rRTM]
Fischer, J. M. & Ravizza, M. (1998) *Responsibility and control: An essay on moral responsibility*. Cambridge University Press. [rRTM]
Fisher, H. (2006) The drive to love: The neural mechanism for mate selection. In: *The new psychology of love*, 2nd edition, ed. R. J. Sternberg & K. Weis, pp. 87–115. Yale University Press. [aRTM]
Flanagan, O. (1991) *Varieties of moral personality: Ethics and psychological realism*. Harvard University Press. [OF]
Flanagan, O. (2007) *The really hard problem: Meaning and the material world*. MIT Press. [OF]
Fletcher, P. C. & Frith, C. D. (2009) Perceiving is believing: A Bayesian approach. *Nature Reviews Neuroscience* 10(1):48–58. [ALM]
Fodor, J. A. (1981) *Representations: Philosophical essays on the foundations of cognitive science*. MIT Press. [YW]
Fodor, J. A. (1983) *The modularity of mind*. MIT Press. [aRTM]
Fodor, J. A. (1986) Précis of *The modularity of mind. Meaning and cognitive structure*. *Behavioral and Brain Sciences* 8(1):1–42. [aRTM]
Fodor, J. A. (2007) Why pigs don't have wings. *London Review of Books*, October 18, 2007. [aRTM]
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. [aRTM]
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. [aRTM]
Frankfurt, H. (1988) *The importance of what we care about*. Cambridge University Press. [rRTM]
Frankish, K. (2004) *Mind and supermind*. Cambridge University Press. [KF]
Frankish, K. (2009) Partial belief and flat-out belief. In: *Degrees of belief*, ed. F. Huber & C. Schmidt-Petri, pp. 75–93. Springer. [KF]
Freud, S. (1920) *On metapsychology: The theory of psychoanalysis*, trans. J. Strachey. Penguin Books. [JK]
Friedrich, J. (1996) On seeing oneself as less self-serving than others: The ultimate self-serving bias? *Teaching of Psychology* 23(2):107–109. [aRTM]
Fuster, J. M. (2006) The cognit: A network model of cortical representation. *International Journal of Psychophysiology* 60(2):125–32. [ALM]
Gagné, F. M. & Lydon, J. E. (2004) Bias and accuracy in close relationships: An integrative review. *Personality and Social Psychology Review* 8(4):322–38. [aRTM]
Gailliot, M. T., Baumeister, R. F., DeWall, C. N., Maner, J. K., Plant, E. A., Tice, D. M., Brewer, L. E. & Schmeichel, B. J. (2007) Self-control relies on glucose as a limited energy source: Willpower is more than a metaphor. *Journal of Personality and Social Psychology* 92:325–36. [BR-S]
Gazzaniga, M. (1995) Consciousness and the cerebral hemispheres. In: *The cognitive neurosciences*, first edition, ed. M. Gazzaniga, pp. 1391–400. MIT Press. [TWZ]
Gendler, T. S. (2008) Alief and belief. *Journal of Philosophy* 105(10):634–63. [arRTM]
Gergen, K. J. (1973) Social psychology as history. *Journal of Personality and Social Psychology* 26(2):309–20. [VJK]
Gervais, W. & Norenzayan, A. (2009) Priming God increases public self-awareness. Unpublished raw data, University of British Columbia. [Author's note: The data are available upon request from Ara Norenzayan at: ara@psych.ubc.ca] [rRTM, AN]
Ghiselin, M. T. (1974) *The economy of nature and the evolution of sex*. University of California Press. [JRL, aRTM]
Gibbon, J. (1977) Scalar expectancy theory and Weber's law in animal timing. *Psychological Review* 84:279–325. [GA]
Gigerenzer, G. & Brighton, H. J. (2009) Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science* 1:107–43. [ETC]
Gigerenzer, G. & Goldstein, D. G. (1996) Reasoning the fast and frugal way: Models of bounded rationality *Psychological Review* 103(4):650–69. [arRTM]
Gigerenzer, G. & Goldstein, D. G. (1999) Betting on one good reason: The take the best heuristic. In: *Simple heuristics that make us smart*, ed. G. Gigerenzer, P. M. Todd, & the ABC Research Group, pp. 75–95. Oxford University Press. [aRTM]
Gigerenzer, G., Hertwig, R., Hoffrage, U. & Sedlmeier, P. (2008) Cognitive illusions reconsidered. In: *Handbook of experimental economics results: Vol. 1 (Handbooks in Economics, No. 28)*, ed. C. R. Plott & V. L. Smith, pp. 1018–34. North-Holland. [ETC]
Gigerenzer, G., Todd, P. M. & the ABC Research Group (1999) *Simple heuristics that make us smart*. Oxford University Press. [ETC, arRTM, KT-K]
Gilbert, D. T. (1991) How mental systems believe. *American Psychologist* 46:107–19. [DSW]
Gilovich, T. (1991) *How we know what isn't so: The fallibility of human reason in everyday life*. Free Press. [JK]
Gintis, H. (2000) Strong reciprocity and human sociality. *Journal of Theoretical Biology* 206:169–79. [aRTM]
Gintis, H. (2003) The hitchhiker's guide to altruism: Gene-culture co-evolution and the internalization of norms. *Journal of Theoretical Biology* 220:407–18. [aRTM]
Gintis, H., Bowles, S., Boyd, R. & Fehr, E. (2003) Explaining altruistic behavior in humans. *Evolution and Human Behavior* 24(3):153–72. [aRTM]
Gintis, H., Smith, E. & Bowles, S. (2001) Costly signalling and cooperation. *Journal of Theoretical Biology* 213:103–19. [aRTM]
Glenberg, A. M. (1997) What memory is for. *Behavioral and Brain Sciences* 20(1):1–55. [JS]
Goldstein, D. G. & Gigerenzer, G. (2002) Models of ecological rationality: The recognition heuristic. *Psychological Review* 109(1):75–90. [aRTM]
Goleman, D. (1987) Who are you kidding? *Psychology Today* 21(3):24–30. [aRTM]
Gottlieb, G. & Lickliter, R. (2007) Probabilistic epigenesis. *Developmental Science* 10:1–11. [TJW]
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):581–98. [aRTM]
Gould, S. J. & Vrba, E. S. (1982) Exaptation: A missing term in the science of form. *Paleobiology* 8(1):4–15. [aRTM]
Govern, J. M. & Marsch, L. A. (2001) Development and validation of the Situational Self-Awareness Scale. *Consciousness and Cognition* 10:366–78. [AN]
Green, L. & Myerson, J. (2004) A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin* 130:769–92. [GA]
Guthrie, S. E. (1993) *Faces in the clouds: A new theory of religion*. Oxford University Press. [aRTM]
Haggard, P., Aschersleben, G., Gehrke J. & Prinz, W. (2002) Action, binding and awareness. In *Attention and performance XIX: Common mechanisms in perception and action*, ed. B. Hommel & W. Prinz, pp. 266–85. Oxford University Press. [ALM]
Haidt, J. (2001) The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 108:814–34. [DLK, rRTM]
Hamilton, A. (2007) Against the belief model of delusion. In: *Reconceiving schizophrenia*, ed. M. C. Chung, K. W. M. Fulford & G. Graham, pp. 217–34. Oxford University Press. [aRTM]
Hamilton, W. D. (1964) Genetical evolution of social behavior, I and II. *Journal of Theoretical Biology* 7:1–52. [aRTM]
Hammond, R. A. & Axelrod, R. (2006) The evolution of ethnocentrism. *Journal of Conflict Resolution* 50:926–36. [JMA]
Hartung, J. (1988) Deceiving down: Conjectures on the management of subordinate status. In: *Self-deception: An adaptive mechanism?*, ed. J. S. Lockard & D. L. Paulhus, pp. 170–85. Prentice Hall. [rRTM]
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. [aRTM]
Haselton, M. G. (2007) Error management theory. In: *Encyclopedia of social psychology, vol. 1*, ed. R. F. Baumeister & K. D. Vohs, pp. 311–12. Sage. [aRTM]
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. [JMA, ETC, MGH, DDPJ, arRTM]
Haselton, M. G. & Buss, D. M. (2003) Biases in social judgment: Design flaws or design features? In: *Responding to the social world: Implicit and explicit processes in social judgments and decisions*, ed. J. P. Forgas, K. D. Williams & W. von Hippel, pp. 23–43. Cambridge University Press. [aRTM]
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. [MGH, DDPJ, JRL, arRTM]
Hayes, P. J. (1978) The Naive Physics manifesto. In: *Expert systems in the micro-electronic age*, ed. D. Michie, pp. 242–70. Edinburgh University Press. [MB, YW]
Hecht, H. H. & Bertamini, M. (2000) Understanding projectile acceleration. *Journal of Experimental Psychology: Human Perception and Performance* 26:730–46. [MB]
Hecht, H. H. & Proffitt, D. R. (1995) The price of expertise: Effects of experience on the water-level task. *Psychological Science* 6:90–95. [MB]
Heine, S. J. & Lehman, D. R. (1995) Cultural variation in unrealistic optimism: Does the West feel more vulnerable than the East? *Journal of Personality and Social Psychology* 68:595–607. [CSD]
Heine, S. J., Lehman, D. R., Marcus, H. R. & Kitayama, S. (1999) Is there a universal need for positive self-regard? *Psychological Review* 106(4):766–94. [OF, JK]
Hemsley, D. R. & Garety, P. A. (1986) The formation of maintenance of delusions: A Bayesian analysis. *British Journal of Psychiatry* 149(July):51–56. [ALM]

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. [aRTM]

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D. & Ziker, J. (2009) Markets, religion, community size and the evolution of fairness and punishment. Unpublished manuscript, University of British Columbia. (Not available online; available upon request from Joseph Henrich at: joseph.henrich@gmail.com) [AN]

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

Henrich, J., Heine, S. J. & Norenzayan, A. (in press) The weirdest people in the world? *Behavioral and Brain Sciences*. [JK, AN]

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

Hitchcott, P. K., Quinn, J. J. & Taylor, J. R. (2007) Bidirectional modulation of goal directed actions by prefrontal cortical dopamine. *Cerebral Cortex* 17(12):2820–27. [ALM]

Hoffrage, U., Hertwig, R. & Gigerenzer, G. (2000) Hindsight bias: A by-product of knowledge updating? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26(3):566–81. [PB]

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

Hood, B. M. (1995) Gravity rules for 2- to 4-year olds? *Cognitive Development* 10:577–98. [NLG]

Hove, M. & Risen, L. (in press) The in-synch effect: Interpersonal synchrony increases affiliation. *Social Cognition*. [JB]

Humphrey, N. (2002) *The mind made flesh*. Oxford University Press. Available at: www.humphrey.org.uk. [aRTM]

Humphrey, N. (2004) The placebo effect. In: *Oxford companion to the mind*, 2nd edition, ed. R. L. Gregory, pp. 735–36. Oxford University Press. Available at: www.humphrey.org.uk [GA, aRTM]

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):801–12. [aRTM]

Irons, W. (2008) Why people believe (what other people see as) crazy ideas. In: *The evolution of religion: Studies, theories, and critiques*, ed. J. Bulbulia, R. Sosis, R. Genet, E. Harris, K. Wyman & C. Genet, pp. 51–60. Collins Foundation Press. [JB]

Jaccard, J., Dodge, T. & Guilamo-Ramos, V. (2005) Metacognition, risk behavior, and risk outcomes: The role of intelligence and perceived knowledge. *Health Psychology* 24:161–70. [DD]

Jahoda, M. (1953) The meaning of psychological health. *Social Casework* 34:349–54. [aRTM]

Jahoda, M. (1958) *Current concepts of positive mental health*. Basic Books. [VJK, aRTM]

James, W. (1890) *The principles of psychology, vol. 1.* Holt. [JDB]

Jaspers, K. (1963) *General psychopathology*, 7th edition, trans. J. Hoenig & M. W. Hamilton. The Johns Hopkins University Press. (Originally published in 1946.) [aRTM, ALM]

Johnson, D. D. P. (2004) *Overconfidence and war: The havoc and glory of positive illusions*. Harvard University Press. [DDPJ]

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):410–46. [aRTM]

Johnson, D. D. P. (2008) Gods of war: The adaptive logic of religious conflict. In: *The evolution of religion: Studies, theories, and critiques*, ed. J. Bulbulia, R. Sosis, C. Genet, R. Genet, E. Harris & K. Wyman, pp. 111–117. Collins Foundation Press. [DDPJ]

Johnson, D. D. P. (2009) The error of God: Error management theory, religion, and the evolution of cooperation. In: *Games, groups, and the global good*, ed. S. A. Levin, pp. 169–180. Springer. [DDPJ, JPS]

Johnson, D. D. P. & Bering, J. M. (2006) Hand of God, mind of man: Punishment and cognition in the evolution of cooperation. *Evolutionary Psychology* 4:219–33. [DDPJ, aRTM, AN]

Johnson, D. D. P. & Krüger, O. (2004) The good of wrath: Supernatural punishment and the evolution of cooperation. *Political Theology* 5(2):159–76. [DDPJ, aRTM]

Johnson, D. D. P., Stopka, P. & Knights, S. (2003) The puzzle of human cooperation. *Nature* 421:911–12. [aRTM]

Johnson, D. J. & Rusbult, C. E. (1989) Resisting temptation: Devaluation of alternative partners as a means of maintaining commitment in close relationships. *Journal of Personality and Social Psychology* 57:967–80. [JMA]

Johnson, E. J. & Goldstein, D. G. (2003) Do defaults save lives? *Science* 302:1338–39. [ETC]

Johnson, S., Dweck, C. S. & Chen, F. (2007) Evidence for infants' internal working models of attachment. *Psychological Science* 18:501–502. [CSD]

Jordan, C. H., Spencer, S. J., Zanna, M. P., Hoshino-Browne, E. & Correll, J. (2003) Secure and defensive high self-esteem. *Journal of Personality and Social Psychology* 85(5):969–78. [rRTM]

Juslin, P. & Olsson, H. (1997) Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review* 104:344–66. [ETC]

Juslin, P., Winman, A. & Olsson, H. (2000) Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review* 107:384–96. [ETC]

Kamins, M. & Dweck, C. S. (1999) Person vs. process praise and criticism: Implications for contingent self-worth and coping. *Developmental Psychology* 35:835–47. [CSD]

Kapur, S. (2003) Psychosis as a state of aberrant salience: A framework linking biology, phenomenology, and pharmacology in schizophrenia. *American Journal of Psychiatry* 160(1):13–23. [ALM]

Karmiloff-Smith, A. (1986) From meta-processes to conscious access: Evidence from children's metalinguistic and repair data. *Cognition* 23:95–147. [NLG]

Karmiloff-Smith, A. (1992) *Beyond modularity: A developmental perspective in cognitive science*. MIT Press. [NLG]

Karmiloff-Smith, A. & Inhelder, B. (1974) If you want to get ahead, get a theory. *Cognition* 3:195–212. [NLG]

Katzir, G. & Howland, H. C. (2003) Corneal power and underwater accommodation in great cormorants (*Phalacrocorax carbo sinensis*). *The Journal of Experimental Biology* 206:833–41. [aRTM]

Katzir, G. & Intrator, N. (1987) Striking of underwater prey by a reef heron, *Egretta gularis schistacea*. *Journal of Comparative Physiology A* 160:517–23. [aRTM]

Kelemen, D. (2004) Are children "intuitive theists"? *Psychological Science* 15:295–301. [aRTM]

Kennedy, Q., Mather, M. & Carstensen, L. L. (2004) The role of motivation in the age-related positivity effect in autobiographical memory. *Psychological Science* 15(3):208–14. [JS]

Kenny, A. J. P., Wadding, C. H., Longuet-Higgins, H. C. & Lucas, J. R. (1972) *The nature of mind.* Edinburgh University Press. [YW]

Kenrick, D. T., Griskevicius, V., Neuberg, S. L. & Schaller, M. (in press) Renovating the pyramid of needs: Contemporary extensions built upon ancient foundations. *Perspectives on Psychological Science*. [JMA]

Kermer, D. A., Driver-Linn, E., Wilson, T. D. & Gilbert, D. T. (2006) Loss aversion is an affective forecasting error. *Psychological Science* 17:649–53. [DSW]

Kimura, M. (1991) The neutral theory of molecular evolution: A review of recent evidence. *Japanese Journal of Genetics* 66:367–86. [JPS]

Kinderman, P. & Bentall, R. P. (1996) Self-discrepancies and persecutory delusions: Evidence for a model of paranoid ideation. *Journal of Abnormal Psychology* 105(1):106–13. [arRTM]

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):341–45. [arRTM]

Kirby, K. N. (1997) Bidding on the future: Evidence against normative discounting of delayed rewards. *Journal of Experimental Psychology: General* 126:54–70. [GA]

Kirsch, I. & Lynn, S. J. (1999) The automaticity of behaviour and clinical psychology. *American Psychologist* 54:504–15. [DSW]

Kitayama, S., Markus, H. R., Matsumoto, H. & Norasakkunkit, V. (1997) Individual and collective process in the construction of the self: Self-enhancement in the U.S. and self-criticism in Japan. *Journal of Personality and Social Psychology* 72:1245–67. [CSD]

Klahr, D. & Dunbar, K. (1988) Dual search space during scientific reasoning. *Cognitive Science* 12:1–48. [NLG]

Klein, S. B., Cosmides, L., Tooby, J. & Chance, S. (2002) Decisions and the evolution of memory: Multiple systems, multiple functions. *Psychological Review* 109(2):306–29. [PB]

Kobayashi, C. & Brown, J. D. (2003) Self-esteem and self-enhancement in Japan and America. *Journal of Cross-Cultural Psychology* 34:567–80. [JDB]

Kraus, A. (1997) Technical delusions as form of stabilisation [Le Delire technique comme forme de stabilisation]. *Evolution Psychiatrique* 62(2):381–99. [ALM]

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

Krebs, D. L. & Laird, P. (1998) Judging yourself as you judge others: Perspective-taking moral development, and exculpation. *Journal of Adult Development* 5:1–12. [DLK]

Krueger, J. & Mueller, R. A. (2002) Unskilled, unaware, or both? The contribution of social-perceptual skills and statistical regression to self-enhancement biases. *Journal of Personality and Social Psychology* 82:180–88. [ETC]

Kruger, J. (1999) Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology* 77:221–232. [JK]

Kruger, J. & Burrus, J. (2004) Egocentrism and focalism in unrealistic optimism (and pessimism). *Journal of Experimental Social Psychology* 40:332–40. [JK]

Kruger, J. & Dunning, D. (1999) Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 77:1121–34. [DD]

Kruger, J. & Savitsky, K. (2009) On the genesis of inflated (and deflated) judgments of responsibility: Egocentrism revisited. *Organizational Behavior and Human Decision Processes* 108:143–52. [JK]

Kugeares, S. (2002) Social anxiety in dating initiation: An experimental investigation of an evolved mating-specific anxiety mechanism. Unpublished doctoral dissertation, Department of Psychology, University of Texas, Austin, Texas. Available at: http://dspace.lib.utexas.edu/bitstream/2152/100/1/kugeares|029.pdf [MGH]

Kunda, Z. (1990) The case for motivated reasoning. *Psychological Bulletin* 108(3):480–98. [GFM]

Lamarck, J.-B. (1809/1914/1984) *Philosophie zoologique, ou Exposition des considerations relatives a l'histoire naturelle des animaux. . .* Paris. Published in English (trans. Hugh Eliot) Macmillan, London, 1914. Reprinted: University of Chicago Press, 1984. (Original work written in 1809). [YW]

Langdon, R. & Bayne, T. (in press) Delusion and confabulation: Mistakes of perceiving, remembering and believing. *Cognitive Neuropsychiatry*. [RL]

Langdon, R. & Coltheart, M. (2000) The cognitive neuropsychology of delusions. *Mind and Language* 15(1):183–216. [RL, aRTM]

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. [aRTM]

Langdon, R., McKay, R. & Coltheart, M. (2008) The cognitive neuropsychological understanding of persecutory delusions. In: *Persecutory delusions: Assessment, theory, and treatment*, ed. D. Freeman, R. Bentall & P. Garety, pp. 221–36. Oxford University Press. [RL, aRTM]

Langston, C. & Sykes, W. (1997) Beliefs and the Big Five: Cognitive bases of broad individual differences in personality. *Journal of Research in Personality* 31:141–65. [ETC]

Larrick, R. P., Burson, K. A. & Soll, J. B. (2007) Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not). *Organizational Behavior and Human Decision Processes* 102:76–94. [ETC]

Lea, S. E. G. & Webley, P. (2006) Money as tool, money as drug: The biological psychology of a strong incentive. *Behavioral and Brain Sciences* 29:161–209. [GA]

Leeser, J. & O'Donohue, W. (1999) What is a delusion? Epistemological dimensions. *Journal of Abnormal Psychology* 108:687–94. [rRTM]

Lilienfeld, S. O., Ammirati, R. & Landfield, K. (2009) Giving debiasing away: Can psychological research on correcting cognitive errors promote human welfare? *Perspectives on Psychological Science* 4:390–99. [DSW]

Lipmann, O. & Bogen, H. (1923) *Naive Physik. Arbeiten aus dem Institut für angewandte Psychologie in Berlin. Theoretische und experimentelle Untersuchungen Über die Fähigkeit zu intelligentem Handeln*. Johann Ambrosius Barth. [MB]

LoBue, V. & DeLoache, J. S. (2008) Detecting the snake in the grass: Attention to fear-relevant stimuli by adults and young children. *Psychological Science* 19(3):284–89. [rRTM]

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

Lockard, J. S. (1980) Speculations on the adaptive significance of self-deception. In: *The evolution of human social behavior*, ed. J. S. Lockard, pp. 257–76. Elsevier. [aRTM]

Lockard, J. S. & Paulhus, D. L., ed. (1988) *Self-deception: An adaptive mechanism?* Prentice Hall. [aRTM]

Lord, C. G., Ross, L. & Lepper, M. R. (1979) Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology* 37(11):2098–109. [GFM]

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* 42:341–46. [aRTM]

Lovejoy, A. O. (1936) *The great chain of being: A study of the history of an idea.* Harvard University Press. [YW]

Lynn, S. J. & McConkey, K. eds. (1998) *Truth in memory*. Guilford Press. [DSW]

Malinowski, B. (1948) *Magic, science and religion and other essays*. Beacon. [DSW]

Maner, J. K., Gailliot, M. T. & Miller, S. L. (2009) The implicit cognition of relationship maintenance: Inattention to attractive alternatives. *Journal of Experimental Social Psychology* 45:174–79. [JMA]

Maner, J. K., Kenrick, D. T., Neuberg, S. L., Becker, D. V., Robertson, T., Hofer, B., Delton, A., Butner, J. & Schaller, M. (2005) Functional projection: How fundamental social motives can bias interpersonal perception. *Journal of Personality and Social Psychology* 88:63–78. [JMA]

Maner, J. K., Rouby, D. A. & Gonzaga, G. (2008) Automatic inattention to attractive alternatives: The evolved psychology of relationship maintenance. *Evolution and Human Behavior* 29:343–49. [JMA]

Marcus, G. F. (2008) *Kluge: The haphazard construction of the human mind.* Houghton Mifflin. [GFM]

Marcus, G. F. (2009) How *does* the mind work? Insights from biology. *Topics in Cognitive Science* 1:145–72. [GFM]

Marshall, M. A. & Brown, J. D. (2007) On the psychological benefits of self-enhancement. In: *Self-enhancement and self-criticism: Theory, research, and clinical implications*, ed. E. Chang, pp. 19–35. American Psychological Association. [JDB]

Mascaro, O. & Sperber, D. (2009) The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition* 112(3):367–80. [DS]

Maslow, A. H. (1950) Self-actualizing people: A study of psychological health. *Personality Symposium* 1:11–34. [aRTM]

Massey, C. & Gelman, R. (1988) Preschoolers' ability to decide whether pictured or unfamiliar objects can move themselves. *Developmental Psychology* 24:307–17. [NLG]

Mather, M. & Carstensen, L. L. (2005) Aging and motivated cognition: The positivity effect in attention and memory. *Trends in Cognitive Sciences* 9(10):496–502. [JS]

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

McCloskey, M. (1983) Intuitive physics. *Scientific American* 248(4):114–22. [MB]

McCloskey, M., Caramazza, A. & Green, B. (1980) Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science* 210(5):1139–41. [MB]

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

McGhie, A. & Chapman, J. (1961) Disorders of attention and perception in early schizophrenia. *British Journal of Medical Psychology* 34:103–16. [ALM]

McKay, R. & Cipolotti, L. (2007) Attributional style in a case of Cotard delusion. *Consciousness and Cognition* 16(2):349–59. [JB]

McKay, R. & Efferson, C. (under review) The subtleties of error management. [rRTM]

McKay, R. & Kinsbourne, M. (in press) Confabulation, delusion, and anosognosia: Motivational factors and false claims. *Cognitive Neuropsychiatry*. [Also to appear in: *Delusion and confabulation: Overlapping or distinct psychopathologies of reality distortion*, ed. R. Langdon & M. Turner. *Macquarie Monographs in Cognitive Science* series. (Series editor, M. Coltheart.) Psychology Press. [aRTM]

McKay, R., Langdon, R. & Coltheart, M. (2007a) Models of misbelief: Integrating motivational and deficit theories of delusions. *Consciousness and Cognition* 16:932–41. [aRTM, ALM]

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. [arRTM]

McKay, R., Langdon, R. & Coltheart, M. (2009) "Sleights of mind": Delusions and self-deception. In: *Delusion and self-deception: Affective and motivational influences on belief formation*, ed. T. Bayne & J. Fernández, pp. 165–85. Psychology Press. [aRTM]

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

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. [aRTM]

Mele, A. R. (1995) *Autonomous agents*. Oxford University Press. [rRTM]

Meyer, D., Leventhal, H. & Gutmann, M. (1985) Common-sense models of illness: The example of hypertension. *Health Psychology* 4:115–35. [DD]

Michaelian, K. (submitted) The epistemology of forgetting. [JS]

Miller, G. (2000) Mental traits as fitness indicators: Expanding evolutionary psychology's adaptationism. *Annals of the New York Academy of Sciences* 907:62–74. [JPS]

Miller, G. (2001) *The mating mind: How sexual choice shaped the evolution of human nature.* Anchor Press. [JPS]

Miller, R. (2008) *A neurodynamic theory of schizophrenia (and related disorders).* Lulu. [ALM]

Millikan, R. G. (1984a) *Language, thought and other biological categories*. MIT Press. [aRTM]

Millikan, R. G. (1984b) Naturalistic reflections on knowledge. *Pacific Philosophical Quarterly* 65(4):315–34. [aRTM]
Millikan, R. G. (1993) *White queen psychology and other essays for Alice*. MIT Press. [aRTM]
Millikan, R. G. (1998) Cognitive luck: Externalism in an evolutionary frame. In: *Philosophy and the sciences of mind*, ed. P. Machamer & M. Carrier, pp. 207–19. [*Pittsburgh-Konstanz Series in the Philosophy and History of Science*]. Pittsburgh University Press and Universitätsverlag Konstanz. [RGM]
Millikan, R. G. (2004) *Varieties of meaning: The 2002 Jean Nicod lectures*. MIT Press/A Bradford Book. [aRTM]
Milton, F., Patwa, V. K. & Hafner, R. J. (1978) Confrontation vs. belief modification in persistently deluded patients. *British Journal of Medical Psychology* 51(2):127–30. [ALM]
Mineka, S., Davidson, M., Cook, M. & Keir, R. (1984) Observational conditioning of snake fear in rhesus monkeys. *Journal of Abnormal Psychology* 93:355–72. [rRTM]
Mishara, A. L. (2007a) Is minimal self preserved in schizophrenia? A subcomponents view. *Consciousness and Cognition* 16(3):715–21. [ALM]
Mishara, A. L. (2007b) Missing links in phenomenological clinical neuroscience? Why we are still not there yet. *Current Opinion in Psychiatry* 60(20):559–69. [ALM]
Mishara, A. L. (in press a) Kafka's doubles, paranoia and the brain: Hypnagogic vs. hyper-reflexive models of disruption of self in neuropsychiatric disorders and anomalous conscious states. *Philosophy, Ethics, and Humanities in Medicine (PEHM)*, PubMed Central Open Access Journal. Available at: http://www.peh-med.com/ [ALM]
Mishara, A. L. (in press b) The unconscious in paranoid delusional psychosis? Phenomenology, nuroscience, psychoanalysis. In: *Founding psychoanalysis*, ed. D. Lohmar & J. Bruzinska. Springer. [ALM]
Mishara, A. L. (2010) Klaus Conrad (1905–1961): Delusional mood, psychosis and beginning schizophrenia. *Schizophrenia Bulletin* 36(1):9–13. [ALM]
Moore, D. A. (2007) Not so above average after all: When people believe they are worse than average and its implications for theories of bias in social comparison. *Organizational Behavior and Human Decision Processes* 102:42–58. [JK]
Moore, D. A. & Healy, P. J. (2008) The trouble with overconfidence. *Psychological Review* 115:502–17. [ETC, JK]
Moore, D. A. & Kim, T. G. (2003) Myopic social prediction and the solo comparison paradox. *Journal of Personality and Social Psychology* 85:1121–35. [JK]
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):402–15. [arRTM]
Morrison, D. (2009) Cordon blues: Review of *Au revoir to all that: The rise and fall of French cuisine*, by M. Steinberger. In: *The Financial Times*, June 27/28, 2009, "Life and Arts" section, p. 16. [VJK]
Mortensen, C. R., Becker, D. V., Ackerman, J. M., Neuberg, S. L. & Kenrick, D. T. (in press) Infection breeds reticence: The effects of disease salience on self-perceptions of personality and behavioral avoidance tendencies. *Psychological Science*. [JMA]
Mowat, R. R. (1966) *Morbid jealousy and murder*. Tavistock. [aRTM]
Mueller, C. M. & Dweck, C. S. (1998) Intelligence praise can undermine motivation and performance. *Journal of Personality and Social Psychology* 75:33–52. [CSD]
Murdock, G. P. & White, D. R. (1969) Standard cross-cultural sample. *Ethnology* 8:329–69. [aRTM]
Murphy, M. C. & Dweck, C. S. (2009, in press) A culture of genius: How an organization's lay theories shape people's cognition, affect, and behavior. *Personality and Social Psychology Bulletin*. [CSD]
Murray, M. & Moore, L. (2009) Costly signaling and the origin of religion. *Journal of Cognition and Culture* 9:225–45. [JPS]
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. [JDB, aRTM]
Myers, D. (2002) *Social psychology*, 7th edition. McGraw-Hill. [aRTM]
Nairne, J. S., Pandeirada, J. N. S. & Thompson, S. G. (2008) Adaptive memory. The comparative value of survival processing. *Psychological Science* 19(2):176–80. [PB]
Nairne, J. S., Thompson, S. R. & Pandeirada, J. N. S. (2007) Adaptive memory: Survival processing enhances retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33(2):263–73. [JS]
Navarrete, C. D., Fessler, D. M. T. & Eng, S. J. (2007) Increased ethnocentrism in the first trimester of pregnancy. *Evolution and Human Behavior* 28:60–65. [JMA]
Nettle, D. (2004) Adaptive illusions: Optimism, control and human rationality. In: *Emotion, evolution and rationality*, ed. D. Evans & P. Cruse, pp. 193–208. Oxford University Press. [DDPJ, aRTM]
Neuberg, S. L. & Cottrell, C. A. (2006) Evolutionary bases of prejudices. In: *Evolution and social psychology*, pp. 163–87, ed. M. Schaller, J. A. Simpson & D. T. Kenrick. Psychology Press. [JMA]
Neuhoff, J. G. (1998) A perceptual bias for rising tones. *Nature* 395:123–24. [MGH]
Neuhoff, J. G. (2001) An adaptive bias in the perception of looming auditory motion. *Ecological Psychology* 13:87–110. [aRTM]
Newport, E. L. (1981) Constraints on structure: Evidence from American Sign Language and language learning. In: *Aspects of development of competence: Minnesota Symposium on Child Psychology, vol. 14*, ed. W. A. Collins, pp. 93–258. Erlbaum. [NLG]
Nickerson, R. S. (1998) Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* 2:175–220. [GFM]
Nisbett, R. E. & Ross, L. (1980) *Human inference: Strategies and shortcomings of social judgment*. Prentice-Hall. [DD]
Noë, A. (2004) *Action in perception*. MIT Press. [aRTM]
Norenzayan, A. (in press) Why we believe: Religion as a human universal. In: *Human morality and sociality: Evolutionary and comparative perspectives*, ed. H. Hogh-Oleson. Palgrave/Macmillan. [AN]
Norenzayan, A. & Shariff, A. F. (2008) The origin and evolution of religious prosociality. *Science* 322:58–62. [DDPJ, aRTM, AN]
Nuttin, J. M. J. R. (1985) Narcissism beyond Gestalt and awareness: The name letter effect. *European Journal of Social Psychology* 15(3):353–61. [MB]
Origgi, G. (2004) Is trust an epistemological notion? *Episteme* 1(1):61–72. [DS]
Pascal, B. (1670/1995) *Pensées*, trans. H. Levi. (Original work published in 1670; H. Levi. English translation, 1995). Oxford University Press. [aRTM]
Pasupathi, M. & Carstensen, L. L. (2003) Age and emotional experience during mutual reminiscing. *Psychology and Aging* 18(3):430–42. [JS]
Paulhus, D. L. (1988) *Manual for the balanced inventory of desirable responding*. Multi-Health Systems. [aRTM]
Peck, M. S. (1978) *The road less traveled*. Simon & Schuster. [aRTM]
Pelham, B. W., Carvallo, M. & Jones, J. T. (2005) Implicit egotism. *Current Directions in Psychological Science* 14:106–10. [MB]
Phillips, K.-A. (2008) Psychosocial factors and survival of young women with breast cancer. Paper presented at the Annual Meeting of the American Society of Clinical Oncology, Chicago, IL, June 2008. [DSW]
Pichon, I., Boccato, G. & Saroglou, V. (2007) Nonconscious influences of religion on prosociality: A priming study. *European Journal of Social Psychology* 37:1032–45. [arRTM, BR-S]
Pinker, S. (1997) *How the mind works*. W. W. Norton. [aRTM]
Pittenger, J. B. (1989) Detection of violation of the law of pendulum motion: Observers' sensitivity to the relation between period and length. *Ecological Psychology* 2:55–81. [MB]
Plantinga, A. (2002) Evolutionary argument against naturalism. In: *Naturalism defeated?*, ed. J. Beilby, pp. 1–12. Cornell University Press. [JPS]
Plotkin, H. (1997) *Darwin machines and the nature of knowledge*. Harvard University Press. [JPS]
Povinelli, D. J. & Bering, J. M. (2002) The mentality of apes revisited. *Current Directions in Psychological Science* 11:115–19. [aRTM]
Premack, D. & Woodruff, G. (1978) Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1(4):515–26. [aRTM]
Proffitt, D. R. (1999) Naive physics. In: *The MIT encyclopedia of the cognitive sciences*, ed. R. Wilson & F. Keil, pp. 577–79. MIT Press. [MB]
Pronin, E., Gilovich, T. & Ross, L. (2004) Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychological Review* 111(3):781–99. [aRTM, DSW]
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):369–81. [aRTM]
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):717–67. [aRTM]
Quine, W. V. O. (1953) *From a logical point of view*. Harvard University Press. [YW]
Quine, W. V. O. (1960) *Word and object*. MIT Press. [aRTM]
Quine, W. V. O. & Ullian, J. S. (1978) *The web of belief*, 2nd edition. Random House. [aRTM]
Quinones-Vidal, E., Lopez-Garcia, J. J., Penaranda-Ortega, M. & Tortosa-Gil, F. (2004) The nature of social and personality psychology as reflected in JPSP, 1965–2000. *Journal of Personality and Social Psychology* 86:435–52. [JK]
Racine, T. P., Wereha, T. J. & Leavens, D. A. (forthcoming) To what extent non-human primates are intersubjective and why. In: *Moving ourselves, moving others: The role of (e)motion in intersubjectivity*, ed. A. Foolen, U. Lüdtke, J. Zlatev & T. P. Racine. John Benjamins. [TJW]
Ramachandran, V. S. (1994a) Phantom limbs, neglect syndromes, repressed memories, and Freudian psychology. *International Review of Neurobiology* 37:291–333. [aRTM]
Ramachandran, V. S. (1994b) Phantom limbs, somatoparaphrenic delusions, neglect syndromes, repressed memories and Freudian psychology.

In: *Neuronal group selection*, ed. O. Sporns & G. Tononi. Academic Press. [aRTM]
Ramachandran, V. S. (1995) Anosognosia in parietal lobe syndrome. *Consciousness and Cognition* 4(1):22–51. [aRTM]
Ramachandran, V. S. (1996a) The evolutionary biology of self-deception, laughter, dreaming and depression: Some clues from anosognosia. *Medical Hypotheses* 47(5):347–62. [aRTM]
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* 61:115–34. [aRTM]
Ramachandran, V. S., Altschuler, E. L. & Hillyer, S. (1997) Mirror agnosia. *Proceedings of the Royal Society of London B: Biological Sciences* 264:645–47. [aRTM]
Ramachandran, V. S. & Blakeslee, S. (1998) *Phantoms in the brain: Human nature and the architecture of the mind*. Fourth Estate. [aRTM]
Randolph-Seng, B. (2009) Nonconscious vigilance: Preconscious control over the influence of subliminal priming. Unpublished doctoral dissertation, Texas Tech University. [BR-S]
Randolph-Seng, B. & Nielsen, M. E. (2007) Honesty: One effect of primed religious representations. *The International Journal for the Psychology of Religion* 17(4):303–15. [arRTM, BR-S]
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):119–22. [aRTM, AN]
Rappaport, R. A. (1999) *Ritual and religion in the making of humanity.* Cambridge University Press. [DDPJ]
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* 18:354–63. [aRTM]
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. [aRTM]
Robert, J. S. (2002) How developmental is evolutionary developmental biology? *Biology and Philosophy* 17:591–611. [TJW]
Roediger, H. L. I. (1996) Memory illusions. *Journal of Memory and Language* 35:76–100. [PB]
Roes, F. L. & Raymond, M. (2003) Belief in moralizing gods. *Evolution and Human Behavior* 24(2):126–35. [aRTM, AN]
Roese, N. J. & Olson, J. M. (2007) Better, stronger, faster: Self-serving judgment, affect regulation, and the optimal vigilance hypothesis. *Perspectives on Psychological Science* 2:124–41. [JK]
Ross, L. & Nisbett, R. E. (1991) *The person and the situation.* McGraw-Hill. [CSD]
Ross, M. & Wilson, A. E. (2002) It feels like yesterday: Self-esteem, valence of personal past experiences, and judgments of subjective distance. *Journal of Personality and Social Psychology* 82(5):792–803. [JS]
Ross, M. & Wilson, A. E. (2003) Autobiographical memory and conceptions of self: Getting better all the time. *Current Directions in Psychological Science* 12(2):66–69. [PB]
Rossano, M. J. (2007) Supernaturalizing social life: Religion and the evolution of human cooperation. *Human Nature* 18:272–94. [aRTM]
Rozin, P. & Fallon, A. E. (1987) A perspective on disgust. *Psychological Review* 94:23–41. [aRTM]
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):383–84. [aRTM]
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):703–12. [aRTM]
Rozin, P. & Nemeroff, C. (2002) Sympathetic magical thinking: The contagion and similarity "heuristics." In: *Heuristics and biases. The psychology of intuitive judgment*, ed. T. Gilovich, D. Griffin & D. Kahneman, pp. 201–16. Cambridge University Press. [KT-K]
Rozinblit, L. & Keil, F. (2002) The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science* 26(216):521–62. [NLG]
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. [aRTM]
Saffran, J. R., Aslin, R. N. & Newport, E. L. (1996) Statistical learning in 8-month-old infants. *Science* 274:1926–28. [CSD]
Schacter, D. L. (2001) *The seven sins of memory.* Houghton Mifflin. [JS]
Schacter, D. L. & Coyle, J. T. (1995) *Memory distortion: How minds, brains, and societies reconstruct the past*. Harvard University Press. [PB]
Schelling, T. (1960) *The strategy of conflict*. Oxford University Press. [JB]
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. [aRTM]
Schjødt, U., Stødkilde-Jørgensen, H., Geertz, A. & Roepstorff, A. (submitted) The power of charisma: Perceived charisma inhibits the attentional and excecutive systems of believers in intercessory prayer. [JB]
Schloss, J. P. (2007) He who laughs best: Religious affect as a solution to recursive cooperative defection. In: *The evolution of religion: Studies, theories, critiques*, ed. J. Bubulia, R. Sosis, E. Harris, R. Genet, C. Genet & K. Wyman, pp. 205–15. Collins Foundation Press. [JPS]
Sears, D. (1986) College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology* 51:515–30. [AN]
Sedikides, C. & Gregg, A. P. (2008) Self-enhancement: Food for thought. *Perspectives on Psychological Science* 3(2):102–16. [MB]
Shapiro, J. R., Ackerman, J. M., Neuberg, S. L., Maner, J. K., Becker, D. V. & Kenrick, D. T. (2009) Following in the wake of anger: When not discriminating is discriminating. *Personality and Social Psychology Bulletin* 35:1356–67. [JMA]
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. [DDPJ, arRTM, AN, BR-S]
Shariff, A. F., Norenzayan, A. & Henrich, J. (2010) The birth of high gods. In: *Evolution, culture, and the human mind*, ed. M. Schaller, A. Norenzayan, S. J., Heine, T. Yamagishi & T. Kameda, pp. 119–36. Psychology Press/Taylor & Francis. [AN]
Shea, C. (2004) The power of positive illusions. *The Boston Globe*, September 26, 2004. [DDPJ]
Sheldrake, R. (1987) Mind, memory, and archetype morphic resonance and the collective unconscious – Part I. *Psychological Perspectives* 18(1):9–25. [YW]
Sherif, M., Harvey, O. J., White, B. J., Hood, W. R. & Sherif, C. W. (1961) *Intergroup conflict and cooperation: The Robbers Cave experiment*. Institute of Group Relations, University of Oklahoma. [JMA]
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* 26:607–23. [aRTM]
Simon, H. (1996) *The sciences of the artificial*, third edition. MIT Press. [KT-K, YW]
Simon, H. A. (1990) A mechanism for social selection and successful altruism. *Science* 250(4988):1665–68. [aRTM]
Simpson, J. A., Gangestad, S. W. & Lerma, M. (1990) Perception of physical attractiveness: Mechanisms involved in the maintenance of romantic relationships. *Journal of Personality and Social Psychology* 59:1192–201. [JMA]
Smith, D. L. (2006) In praise of self-deception. *Entelechy: Mind and Culture* 7. (Online publication, available at: http://www.entelechyjournal.com/davidli-vingstonesmith.htm) [aRTM]
Smith, E. R. & Collins, E. C. (2009) Contextualizing person perception: Distributed social cognition. *Psychological Review* 116:343–64. [DD]
Smullyan, R. (1983) *5000 B.C. and other philosophical fantasies*. St. Martin's Press. [aRTM]
Sosis, R. (2000) Religion and intragroup cooperation: Preliminary results of a comparative analysis of utopian communities. *Cross-Cultural Research* 34(1):77–88. [JB]
Sosis, R. (2003) Why aren't we all Hutterites? *Human Nature* 14(2):91–127. [JB]
Sosis, R. (2004) The adaptive value of religious ritual. *American Scientist* 92:166–72. [aRTM]
Sosis, R. (2005) Does religion promote trust? The role of signaling, reputation, and punishment. *Interdisciplinary Journal of Research on Religion* 1(1):1–30. [JB]
Sosis, R. & Alcorta, C. (2003) Signaling, solidarity, and the sacred: The evolution of religious behavior. *Evolutionary Anthropology* 12:264–74. [DDPJ]
Sosis, R. & Bressler, E. R. (2003) Cooperation and commune longevity: A test of the costly signaling theory of religion. *Cross-Cultural Research* 37(2):211–39. [DDPJ]
Spangenberg, K. B., Wagner, M. T. & Bachman, D. L. (1998) Neuropsychological analysis of a case of abrupt onset mirror sign following a hypotensive crisis in a patient with vascular dementia. *Neurocase* 4:149–154. [aRTM]
Spelke, E. S. & Kinzler, K. D. (2007) Core knowledge. *Developmental Science* 10:89–96. [CSD]
Sperber, D. (1982) Apparently irrational beliefs. In: *Rationality and relativism*, ed. S. Lukes & M. Hollis, pp. 149–80. Blackwell. [DS]
Sperber, D. (1985) Anthropology and psychology: Towards an epidemiology of representations. (The Malinowski Memorial Lecture, 1984). *Man* (New Series) 20:73–89. [DS]
Sperber, D. (1990) The epidemiology of beliefs. In: *The social psychological study of widespread beliefs*, ed. C. Fraser & G. Gaskell, pp 25–44. Clarendon Press. [DS]
Sperber, D. (1997) Intuitive and reflexive beliefs. *Mind and Language* 12(1):67–83. [DS]
Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G. & Wilson D. (forthcoming) Epistemic vigilance. *Mind and Language*. [DS]

Spitzer, M. (1990) On defining delusions. *Comprehensive Psychiatry* 31(5):377–97. [rRTM]
Stalnaker, R. C. (1984) *Inquiry*. MIT Press. [KF]
Stephens, G. L. & Graham, G. (2004) Reconceiving delusion. *International Review of Psychiatry* 16(3):236–41. [aRTM]
Stich, S. (1990) *The fragmentation of reason*. MIT Press. [JRL, arRTM, RGM, JPS]
Stillman, T. F., Baumeister, R. F., Vohs, K. D., Lambert, N. M., Fincham, F. D. & Brewer, L. E. (in press) Personal philosophy and personal achievement: Belief in free will predicts better job performance. *Social Psychological and Personality Science*. [BR-S]
Stone, T. & Young, A. W. (1997) Delusions and brain injury: The philosophy and psychology of belief. *Mind and Language* 12:327–64. [aRTM]
Suddendorf, T. & Corballis, M. C. (2007) The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences* 30(3):299–313. [PB]
Sutton, J. (2009) Remembering. In: *The Cambridge handbook of situated cognition*, ed. P. Robbins & M. Aydede, pp. 217–35. Cambridge University Press. [JS]
Tajfel, H., Billig, M. G., Bundy, R. F. & Flament, C. (1971) Social categorization and intergroup behavior. *European Journal of Social Psychology* 1:149–77. [JDB]
Tajfel, H. & Turner, J. (1986) The social identity theory of intergroup behavior. In: *Psychology of intergroup relations*, ed. S. Worchel & W. G. Austin, pp. 7–24. Nelson-Hall. [JMA]
Tallis, F. (2005) *Love sick*. Arrow Books. [aRTM]
Talmont-Kaminski, K. (2009) The fixation of superstitious beliefs. *Teorema* 28(3):81–95. [KT-K]
Talmont-Kaminski, K. (in preparation) *In a mirror, darkly: How superstition and religion reflect rationality*. [KT-K]
Taves, A. (2009) *Religious experience reconsidered: A building block approach to the study of religion and other special things*. Princeton University Press. [JB]
Taylor, S. E. (1989) *Positive illusions: Creative self-deception and the healthy mind*. Basic Books. [JDB, KF, VJK, aRTM]
Taylor, S. E. & Brown, J. D. (1988) Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin* 103(2):193–210. [JDB, OF, KF, MGH, JK, aRTM, JS]
Taylor, S. E. & Brown, J. D. (1994a) Illusion of mental health does not explain positive illusions. *American Psychologist* 49:972–73. [JDB]
Taylor, S. E. & Brown, J. D. (1994b) Positive illusions and well-being revisited: Separating fact from fiction. *Psychological Bulletin* 116(1):21–27. [MB, JDB, aRTM]
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. [aRTM]
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):605–15. [aRTM]
Todd, P. M. & Gigerenzer, G. (2007) Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science* 16:167–71. [ETC]
Tooby, J. & Cosmides, L. (1992) The psychological foundations of culture. In: *The adapted mind: Evolutionary psychology and the generation of culture*, ed. J. Barkow, L. Cosmides & J. Tooby, pp. 19–136. Oxford University Press. [JRL]
Trimble, D. E. (1997) The religious orientation scale: Review and meta-analysis of social desirability effects. *Educational and Psychological Measurement* 57:970–86. [AN]
Trivers, R. L. (1971) The evolution of reciprocal altruism. *The Quarterly Review of Biology* 46:35–57. [aRTM]
Trivers, R. L. (1974) Parent-offspring conflict. *American Zoologist* 14:249–64. [aRTM]
Trivers, R. L. (1985) *Social evolution*. Benjamin-Cummings. [ETC, aRTM]
Trivers, R. L. (2000) The elements of a scientific theory of self-deception. In: *Evolutionary perspectives on human reproductive behavior. Annals of the New York Academy of Sciences, vol. 907*, ed. D. LeCroy & P. Moller, pp. 114–31. New York Academy of Sciences. [ETC, DDPJ, aRTM]
Trivers, R. L. (2006) Foreword to Richard Dawkins' *The selfish gene*. In: *The selfish gene: 30th anniversary edition*, pp. xix–xx. Oxford University Press. [aRTM]
Tsuang, M. T., Faraone, S. V. & Day, M. (1988) Schizophrenic disorders. In: *The new Harvard guide to psychiatry*, ed. A. Nicholi, Jr., pp. 259–95. Belknap Press. [ALM]
Uhlhaas, P. J. & Mishara A. L. (2007) Perceptual anomalies in schizophrenia: Integrating phenomenology and cognitive neuroscience. *Schizophrenia Bulletin* 33(1):142–56. [ALM]
Vaillant, G. (1977) *Adaptation to life*. Little, Brown. [aRTM]
Vaish, A., Grossmann, T. & Woodward, A. (2008) Not all emotions are created equal: The negativity bias in early development. *Psychological Bulletin* 134:383–403. [CSD]
Van Leeuwen, D. S. N. (2007) The spandrels of self-deception: Prospects for a biological theory of a mental phenomenon. *Philosophical Psychology* 20(3):329–48. [aRTM]
Van Vugt, M., De Cremer, D. & Janssen, D. (2007) Gender differences in competition and cooperation: The male warrior hypothesis. *Psychological Science* 18:19–23. [JMA]
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* 39:587–99. [aRTM]
Velleman, J. D. (2006) The self as narrator. In: *Self to self: Selected essays*, by J. D. Velleman, pp. 203–23. Cambridge University Press. [JS]
Vohs, K. D. & Schooler, J. (2008) The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science* 19(1):49–54. [rRTM, BR-S]
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 plenary session of the XIX Biennial Conference of the International Society for Human Ethology (ISHE08), Bologna, Italy, July 13–18, 2008. Conference website: http://www.ishe08.org/ [aRTM]
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. [aRTM]
Voltaire, F. M. A. (1759/1962) *Candide*, trans. T. G. Smollett. Washington Square Press. [aRTM]
von Weizsäcker, V. (1950) *Der Gestaltkreis. Theorie der Einheit von Wahrnnehmen und Bewegen* 4. Aufl. (4th edition). Georg Thieme Verlag. [ALM]
von Weizsäcker, V. (1956) *Pathosophie*. Vandenhoeck & Ruprecht. [ALM]
Vouloumanos, A. & Werker, J. F. (2007) Listening to language at birth: Evidence for a bias for speech in neonates. *Developmental Science* 10:159–64. [CSD]
Wagner, A. (2005) Robustness, evolvability, and neutrality. *FEBS Letters* 579(8):1772–78. [JPS]
Wakefield, J. C. (1992) The concept of mental disorder: On the boundary between biological facts and social values. *American Psychologist* 47:373–88. [JRL, rRTM]
Wakefield, J. C. (1999a) Evolutionary versus prototype analyses of the concept of disorder. *Journal of Abnormal Psychology* 108(3):374–99. [rRTM]
Wakefield, J. C. (1999b) Mental disorder as a black box essentialist concept. *Journal of Abnormal Psychology* 108(3):465–72. [rRTM]
Wallace, B. (1973) Misinformation, fitness and selection. *American Naturalist* 107:1–7. [aRTM]
Watson, J. S. (1985) Contingency perception in early social development. In: *Social perception in infants*, ed. T. M. Field & N. A. Fox, pp. 157–76. Ablex. [CSD]
Weber, E. U. & Johnson, E. J. (2009) Mindful judgment and decision making. *Annual Review of Psychology* 60:53–85. [ETC]
Wegner, D. M. (2002) *The illusion of conscious will*. Bradford Books/MIT Press. [rRTM]
Wegner, D. M. (2004) Précis of *The illusion of conscious will*. *Behavioral and Brain Sciences* 27(5):649–59; discussion 660–92. [ALM]
Wegner, D. M. (2005) Who is the controller of controlled processes? In: *The new unconscious*, ed. R. Hassin, J. Uleman & J. A. Bargh, pp. 19–58. Oxford University Press. [BR-S]
Wenger, A. & Fowers, B. J. (2008) Positive illusions in parenting: Every child is above average. *Journal of Applied Social Psychology* 38(3):611–34. [aRTM]
Wereha, T. J. & Racine, T. P. (2009) Evolutionary psychology at a crossroads? A review of *Moral psychology, volume 1: The evolution of morality: Adaptations and innateness*. *Journal of Research on Character Education* 6:95–99. [TJW]
Williams, A. F. (2003) Views of U.S. drivers about driving safety. *Journal of Safety Research* 34(5):491–94. [aRTM]
Williams, G. C. (1966) *Adaptation and natural selection*. Princeton University Press. [rRTM]
Wilson, A. E. & Ross, M. (2001) From chump to champ: People's appraisals of their earlier and current selves. *Journal of Personality and Social Psychology* 80(4):572–84. [JS]
Wilson, A. E. & Ross, M. (2003) The identity function of autobiographical memory: Time is on our side. *Memory* 11(2):137–49. [JS]
Wilson, D. S. (1990) Species of thought: A comment on evolutionary epistomology. *Biology and Philosophy* 5:37–62. [DSW]
Wilson, D. S. (1995) Language as a community of interacting belief systems: A case study involving conduct toward self and others. *Biology and Philosophy* 10:77–97. [DSW]
Wilson, D. S. (2002) *Darwin's cathedral: Evolution, religion and the nature of society*. University of Chicago Press. [DDPJ, arRTM, KT-K, DSW]
Wilson, D. S. (2005) Testing major evolutionary hypotheses about religion with a random sample. *Human Nature* 16(4):382–409. [DSW]

Wilson, D. S. (2006) Human groups as adaptive units: Toward a permanent consensus. In: The innate mind: Culture and cognition, ed. P. Carruthers, S. Laurence & S. Stich, pp. 78–90. Oxford University Press. [DSW]

Wilson, D. S. (2010) Rational and irrational beliefs from an evolutionary perspective. In: *Rational and irrational beliefs*, ed. D. David, S. J. Lynn & A. Ellis, pp. 63–74. Oxford University Press. [DSW]

Wilson, T. D. (2009) Know thyself. *Perspectives on Psychological Science* 384–89. [DSW]

Wiltermuth, S. S. & Heath, C. (2008) Synchrony and cooperation. *Psychological Science* 20:1–5. [JB]

Wimsatt, W. (2007) *Re-engineering philosophy for limited beings*. Harvard University Press. [KT-K]

Windschitl, P. D., Kruger, J. & Simms, E. N. (2003) The influence of egocentrism and focalism on people's optimism in competitions: When what affects us equally affects me more. *Journal of Personality and Social Psychology* 85:389–408. [JK]

Wolf, M., van Doorn, G. S., Leimar, O. & Weissing, F. J. (2007) Life-history trade-offs favour the evolution of animal personalities. *Nature* 447:581–84. [ETC]

Wolpert, L. (2000) *The unnatural nature of science.* Harvard University Press. [JPS]

Wong, T. Y., Foster, P. J., Hee, J., Ng, T. P., Tielsch, J. M., Chew S. J., Johnson, G. J. & Seah, S. K. L. (2000) Prevalence and risk factors for refractive errors in adult Chinese in Singapore. *Investigative Ophthalmology and Visual Science* 41:2486–94. [aRTM]

Woodward, A. & Needham, A. (Ed.) (2009) *Learning and the infant mind*. Oxford University Press. [CSD]

Wrangham, R. W. (1999) Is military incompetence adaptive? *Evolution and Human Behaviour* 20:3–17. [DDPJ]

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):259–91. [aRTM]

Yamazaki, K. (2008) Colors of young and old spring leaves as a potential signal for ant-tended hemipterans. *Plant Signaling and Behavior* 3(11):984–85. [rRTM]

Yates, J., Bessman, M., Dunne, M., Jertson, D., Sly, K. & Wendelboe, B. (1988) Are conceptions of motion based on a naive theory or on prototypes? *Cognition* 29:251–75. [MB]

Young, A. W. (1999) Delusions. *The Monist* 82(4):571–89. [aRTM]

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

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. [aRTM]

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

Zolotova, J. & Brüne, M. (2006) Persecutory delusions: Reminiscence of ancestral hostile threats? *Evolution and Human Behavior* 27(3):185–92. [aRTM]

https://doi.org/10.1017/S0140525X09991440 Published online by Cambridge University Press