Ryan McKay and Daniel Dennett

OUR EVOLVING BELIEFS ABOUT EVOLVED MISBELIEF

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" - 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 and 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 and Feltz**, who note that the argument for adaptive misbelief is not new. We agree with Cokely and 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; 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;** 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 hyper-general 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," 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 and 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 below.

It is tempting to re-baptize acceptance as *c-lief,* since it stands to belief roughly as belief stands to alief, a more sophisticated and expensive state, reserved now for just one species, us. (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 and 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. He "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 below, sections R5 & R7). Wilks expresses doubts about how "brain modifications might conceivably affect the gametes," ignoring the Baldwin Effect (Deacon, 1997; Dennett, 1991, 1995, 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 that have never seen a snake (Mineka, Davidson, Cook & Keir, 1984; LoBue and DeLoache, 2008), for instance, 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' 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 and 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** below, 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 below - see, e.g., 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 this can be assessed by evaluating known cognitive mechanisms and their evolutionary costs, this has to remain an unsettled question.

Wereha and 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 their 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 and 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" and also 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 below).

In this light, **Bertamini and 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 and 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 and 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. The latter analysis across task contexts has not been done, 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 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.

More serious concerns are raised by **Kruger et al.** in their measured and informative commentary. Kruger et al. 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 and 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 below 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 and 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 and 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 and Lynn cite points out that the category "self-esteem" encompasses a range of subtypes. For

example, Jordan, Spencer, Zanna, Hoshino-Browne and Correll (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, Langdon, & Coltheart, 2007; Moritz, Werner, & von Collani, 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 and 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 and 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 and 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 and Shackelford** draw our attention

to a similar analysis by Wakefield (1992). Wakefield's (1992; see also 1999a,b) 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 and 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 and 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're 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 below). It is worth noting, however, that delusions can resist the presentation of manifestly contradictory evidence; indeed, as **Mishara and Corlett** show, such evidence may even *strengthen* delusional conviction through the process of reconsolidation (see Corlett, Krystal, Taylor & Fletcher, 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 and 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 and 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 and 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's 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. Their 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 and 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 above 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're 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, Boccato and Saroglou (2007) and Shariff and Norenzayan (2007).

Norenzayan, Shariff & Gervais 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 yet that 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 & 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 and Lynn** and **Zawidzki** provide related analyses. The accounts of Talmont-Kaminski and Wilson and 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 and 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 and 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 and 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, 1995, 2006) - from the false religious and social propaganda that they describe? Wilson and 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 (1995, 2006): 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 do not extend as far as **Schloss and Murray** propose. As with other commentators, they think the case for adaptive (or fitness-neutral) misbelief is stronger than we allow. Our view, they 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. None the less, 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.

Schloss and 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 the 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.

They 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 and Corlett** and **Randolph-Seng** raise the issue. As in our treatment of the "user-illusion" (see above, 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's 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's briefly recap each of these features. First, an adaptive misbelief must be a bona fide belief. It can't be merely an alief, and it can't 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 can't 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 only be adaptive 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 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

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

Avital, E. & Jablonka, E. (2000). *Animal Traditions: Behavioural Inheritance in Evolution*. Cambridge, UK: Cambridge University Press.

Baumeister, R. F. (1989). The optimal margin of illusion. *Journal of Social and Clinical Psychology, 8*, 176-89.

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

Blackmore, S. (1999). *The meme machine*. Oxford: Oxford University Press.

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

Corlett, P. R., Krystal, J. H., Taylor, J. R. & Fletcher, P. C. (2009). Why do delusions persist? *Frontiers in Human Neuroscience, 3* http://www.frontiersin.org/humanneuroscience/paper/10.3389/neuro.09/012.2009/html/

Crick, F. (1994). *The astonishing hypothesis: The scientific search for the soul*. New York: Scribner.

Deacon, T. (1997). The Symbolic Species: The Coevolution of Language and the Brain. New York: Norton.

Dennett, D. C. (1978). *Brainstorms: Philosophical Essays on Mind and Psychology.* Cambridge, MA: MIT Press.

Dennett, D. C. (1981). Making sense of ourselves (reply to S. Stich, Dennett on Intentional Systems). *Philosophical Topics, 12*, 63-81. Reprinted in J. I. Biro and R. W. Shahan (Eds.) *Mind, Brain, and Function: Essays in the Philosophy of Mind*, University of Oklahoma Press, 1982.

Dennett, D. C. (1984a). A route to intelligence: oversimplify and self-monitor. Available on the web at http://ase.tufts.edu/cogstud/incpages/publctns.shtml

Dennett, D. C. (1984b). *Elbow room*. Cambridge, MA: MIT Press.

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.

Dennett, D. C. (1991). *Consciousness explained.* New York: Little, Brown & Co.

Dennett, D. C. (1995). *Darwin's Dangerous Idea: Evolution and the Meanings of Life.* New York: Simon & Schuster.

Dennett, D. C. (2003a). *Freedom evolves*. Viking Press.

Dennett, D. C. (2003b). The Baldwin Effect: A Crane, not a Skyhook. In B. H. Weber & D. J. Depew (Eds.) *Evolution and Learning: The Baldwin Effect Reconsidered* (pp. 60-79). MIT Press, Bradford Books.

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

Dennett, D. C. (2009). Darwin's 'strange inversion of reasoning'. *PNAS*, *106*, suppl. 1, 10061-10065.

Dennett, D. C. (forthcoming). *Science and Religion: Are they Compatible? A Debate With Alvin Plantinga*. Oxford University Press.

Dennett, D. & McKay, R. (2006). A continuum of mindfulness (Commentary on Mesoudi et al). *Behavioral and Brain Sciences, 29*(4), 353-354.

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.

Fischer, J. M. (1994). *The metaphysics of free will*. Oxford: Blackwell.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: An essay on moral responsibility*. Cambridge: Cambridge University Press.

Frankfurt, H. (1988). *The importance of what we care about*. Cambridge: Cambridge University Press.

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

Gervais, W. & Norenzayan, A. (2009). Priming God increases public self-awareness. Unpublished raw data. University of British Columbia.

Gigerenzer, G. & Goldstein, D. G. (1996) Reasoning the fast and frugal way: Models of bounded rationality *Psychological Review* 103(4):650–69.

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

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108,* 814–834.

Hartung, J. (1988). Deceiving down: Conjectures on the management of subordinate status. In J. S. Lockard & D. L. Paulhus (Eds.) *Self-deception: An adaptive mechanism?* (pp. 170-185). Englewood Cliffs, NJ: Prentice Hall.

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

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

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), 969-978.

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

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

Leeser, J., & O'Donohue, W. (1999). What is a delusion? Epistemological dimensions. *Journal of Abnormal Psychology, 108*, 687-694.

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), 284-289.

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

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

Mele, A. R. (1995). *Autonomous Agents*. New York: Oxford University Press.

Mineka, S., Davidson, M., Cook, M. & Keir, R. (1984). Observational conditioning of snake fear in rhesus monkeys. *Journal of Abnormal Psychology, 93*, 355-372.

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

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

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

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.

Spitzer, M. (1990). On defining delusions. *Comprehensive Psychiatry, 31*(5), 377-397.

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

Vohs, K. D., & Schooler, J. W. (2008). The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science, 19*(1), 49-54.

Wakefield, J. C. (1992). The concept of mental disorder: On the boundary between biological facts and social values. *American Psychologist, 47*, 373-388.

Wakefield, J. C. (1999a). Evolutionary versus prototype analyses of the concept of disorder. *Journal of Abnormal Psychology, 108*(3), 374-399.

Wakefield, J. C. (1999b). Mental disorder as a black box essentialist concept. *Journal of Abnormal Psychology, 108*(3), 465-472.

Wegner, D. (2002). *The illusion of conscious will*. Cambridge, MA: Bradford Books/MIT.

Williams, G. C. (1966). *Adaptation and Natural Selection*. Princeton, N.J: Princeton University Press.

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

Yamazaki, K. (2008). Colors of young and old spring leaves as a potential signal for ant-tended hemipterans. *Plant Signaling & Behavior*, 3(11), 984-985.