Alief in Action (and Reaction)

TAMAR SZABÓ GENDLER

Abstract: I introduce and argue for the importance of a cognitive state that I call *alief*. An alief is, to a reasonable approximation, an innate or habitual propensity to respond to an apparent stimulus in a particular way. Recognizing the role that alief plays in our cognitive repertoire provides a framework for understanding reactions that are governed by non-conscious or automatic mechanisms, which in turn brings into proper relief the role played by reactions that are subject to conscious regulation and deliberate control.

> Certain complex actions are of direct or indirect service under certain states of the mind, in order to relieve or gratify certain sensations, desires, and whenever the same state of mind is induced, however feebly, there is a tendency through the force of habit and association for the same movements to be performed, though they may not then be of the least use.
>
> Charles Darwin 'The principle of servicable associated Habits' (Darwin, 1898, p. 28).

Introduction

A frog laps up the BB that bounces past its tongue. A puppy bats at the 'young dog' in the mirror in front of him. A sports fan watching a televised rerun of a baseball game loudly encourages her favorite player to remain on second base.

My thinking about the matters discussed in this paper has been profoundly influenced by the groundbreaking discussions of belief and related attitudes by Michael Bratman, Patricia Churchland, Paul Churchland, Daniel Dennett, Donald Davidson, Fred Dretske, Jerry Fodor, H. H. Price, Ruth Garrett Millikan, Steven Stich, and Bernard Williams, and by more recent work of Andy Egan, Sally Haslanger, Richard Holton, Susan Hurley, David Owens, Eric Schwitzgebel, Michael Smith, David Velleman and Ralph Wedgwood. For those familiar with the work of these thinkers, evidence of their influence should be apparent throughout. For specific discussion of the issues addressed in this essay, I am grateful to John Bargh, John Bengson, Paul Bloom, Daniel Bonavec, Charles Brittain, Richard Brooks, Carolyn Caine, David Chalmers, Troy Cross, Greg Currie, Paul Davies, Michael Della Rocca, Gil Diesendruck, Andy Egan, Emily Esch, Sam Guttenplan, Verity Harte, Shelly Kagan, Jill North, Elliot Paul, J. Brendan Ritchie, Eric Schwitzgebel, Matthew Noah Smith, Zoltán Gendler Szabó, Kara Weisman, Ralph Wedgwood and Ken Winkler, and to audiences at Princeton University (March 2007), the Central APA Chicago (April 2007), the *Mind & Language* Pretence and Imagination Conference at the University of London (June 2007) (with commentator Greg Currie), Connecticut College (November 2007), MIT (November 2007), the University of Texas at Austin Graduate Philosophy Conference (April 2008) (with commentator Daniel Bonavec), the Yale Philosophy Faculty Discussion Group (May 2008) and Paul Bloom's Mind and Development Lab Meeting (May 2008) where excellent questions, comments, objections and suggestions were raised in response to talks and presentations where some of these ideas were explored.

Address for correspondence: Philosophy Department, Yale University, PO BOX 208306, New Haven, CT 06520-8306, USA.
Email: tamar.gendler@yale.edu

Mind & Language, Vol. 23 No. 5 November 2008, pp. 552–585.

A cinema-goer watching a horror film 'emits a shriek and clutches desperately at his chair'. A man suspended safely in an iron cage above a cliff 'trembl[es] when he surveys the precipice below him'. An avowed anti-racist exhibits differential startle responses when Caucasian and African faces are flashed before her eyes.

In each of these cases, an experience or behavior is elicited that violates an apparent norm. At least in some sense of 'should', the dog *should* yelp only at an actual dog, the movie-goer *should* exhibit fear only in the face of actual danger, the anti-racist *should* react similarly to members of different racial groups. (Whether this is the same sense of *should* is an issue that I will return to later in the paper.) Moreover, in each of these cases, there is no easy way for the agent to override the norm-violating tendency. 'Frogs continue to snap at (and ingest) bee-bees even when they have plenty of evidence that the bee-bees they're snapping at aren't flies' (Fodor, 1999).[1] The sports fan will feel the temptation to shout at the screen even if she emphatically reminds herself that she is watching a rerun. And the caged 'man … cannot forbear trembling … tho' he knows himself to be perfectly secure from falling, by his experience of the solidity of the iron, which supports him' (Hume, 1739/1978, p. 148).

My goal in this paper is to offer a general account of these sorts of recalcitrant norm-discordant responses. In each of these cases, I will contend, the reactions in question are best understood as resulting from a mental state that I will call *alief*.[2] An alief is, to a reasonable approximation, an innate or habitual propensity to respond to an apparent stimulus in a particular way. Recognizing the role that alief plays in our cognitive repertoire brings out the connection between a number of otherwise apparently discrepant issues, and renders unmysterious a number of otherwise perplexing phenomena. By providing a framework for understanding reactions that are governed by non-conscious or automatic mechanisms, it brings reactions that are subject to conscious regulation and deliberate control into proper relief. And by directing philosophical attention to responses that are governed by habit and instinct, it encourages a new appreciation of a number of important insights from the ancient and early modern traditions.

Overview

The paper has four main sections. In the first and second, I offer a general account of alief. I present a number of examples of alief, enumerate its key characteristics, and explain the ways in which alief is like and unlike classic mental states like

[1] Indeed, as Fodor remarks, it's not 'just that frogs sometimes go for bee-bees … they are prepared to go on going for bee-bees *forever*' (Fodor, 1999, p. 241.)

[2] I introduce and argue for the utility of such a notion in a companion article, 'Alief and Belief' (Gendler, forthcoming). Because I have tried to make each piece self-standing, there is inevitably some overlap between them, though I have tried to keep this to a minimum.

belief, desire, and imagination. I also distinguish among various kinds of alief, bringing out the ways that aliefs resulting from innate propensities do and do not differ from aliefs resulting from acquired habits, and considering the implications this has for the account that I am offering.

In the third, I turn to a discussion of norm-concordant and norm-discordant aliefs. When a subject's environment is stable, typical, and desirable, and the subject is attentive to its relevant features, her salient occurrent aliefs will be largely in accord with her occurrent reality-reflective attitudes (that is, beliefs and their teleofunctional analogues.) But when a subject's environment is unstable, atypical, or undesirable, or when a subject is reality-inattentive in certain ways, her salient occurrent aliefs will come apart from her occurrent reality-reflective attitudes. Sometimes this discord is deliberate and welcome: daydreaming, roller-coasters and therapy all exploit our capacity for belief-discordant alief. But sometimes this discord is inadvertent and unwelcome: superstition, phobias and bad habits are all supported by the same capacity. In the third section, I discuss these issues.

In the final section of the paper, I turn to the topic of how we might regulate and respond to discordant alief in cases where discord is unwelcome. As beings who are simultaneously embodied and capable of rational agency, the challenge is one that we face repeatedly. This has not gone unnoticed. It is the challenge that the ancients explored when they considered the problem of *harmonizing the parts of the soul*,[3] and that the moderns discussed when they examined the *conflict between reason and the passions*.[4] And it is one that contemporary cognitive and social psychology (among other disciplines) have been exploring under many rubrics—both behavioral and neurological. Crudely put, there are two strategies for regulating alief. One—stressed by Aristotle among others (especially in the *Nicomachean Ethics*)—involves the cultivation of alternative habits through deliberate rehearsal. The other—stressed by Descartes among others (especially in the *Passions of the Soul*)—involves the refocusing of attention through directed imagination.

---

[3] Examples include the following: Plato: *Gorgias* 492d-494e; *Republic*, Bks. 4-10, especially 434d-445b, 601b-606e and 611a-612a; *Timaeus* 34b-37c, 41a-44d, 69a-72d; *Phaedrus* 245c-256e (all in Cooper, 1997.) Aristotle: *Nicomachean Ethics* Bks 1.13 and 7.1-10, 10.6-8; *On the Soul* Bk 1.1, 3 (esp. 3.9) (all in Barnes, 1984.) Alcinous: *Handbook* chapters 23-25 (in Alcinous, 1993.) Galen: *On the Doctrines of Hippocrates and Plato*, esp. bks 3-6 (in Galen, 2005.) Plutarch: *On the Generation of the Soul in the Timaeus* (in Plutarch, 1976.) Plotinus: e.g. *Enneads* 1.1, 1.2, 4.3 (in Plotinus, 1984.) (Many thanks to Charles Brittain and Verity Harte for guidance on these sources.) It is also a major theme in Confucian ethics. The ancient (3rd century BCE) Chinese philosopher Xunzi (sometimes transliterated 'Hsun-tzu'), for example, apparently advocates the view that the sage is one whose second nature is in accord with virtue. (Thanks to Eric Schwitzgebel to alerting me to the Chinese literature.)

[4] For numerous examples, see James, 1999 (which also includes discussions of Aristotle and Aquinas). Cf. also, among others, Baier, 1991; Hirschman, 1977; and the papers collected in Jenkins, Whiting and Williams, 2005. (Many thanks to Ken Winkler for help with these references.)

Both of these strategies have important analogues in the contemporary literature on racism. In the final section of the paper, I offer some preliminary remarks on this topic.

Two Caveats

Two final caveats. This project is provisional in several senses. First: I make no claim to alief's being a fundamental mental category, one that will be part of our 'final theory' of how the mind makes sense of the world. Rather, I am making a parity argument: that any theory that makes appeal to notions like belief, desire and pretense in order to explain behavior needs to make appeal to (something like) alief in order to make sense of a wide range of otherwise perplexing phenomena. Introducing the notion of alief into our descriptive repertoire provides a useful alternative way of answering 'why?' questions when confronted with a behavior or tendency that we seek to explain. It provides an alternative that falls somewhere in between a classic reason-based explanation (of the sort offered by belief/desire accounts) and a simple physical-cause explanation (of the sort offered by accounts that appeal to physical or chemical descriptions.) Without the availability of such a notion in our present framework, we are likely to misattribute mental states (for example, by crediting or blaming a subject for a belief when only an alief is present, or by suggesting that her belief is somehow partial or weak), overlook important similarities (for example, by failing to recognize the resemblances among domains such as fictional emotions, superstition, heuristics-based errors, and residual racism), neglect certain continuities (for example, between our own mental states, and those of non-human animals with whom we are evolutionarily continuous), and lack explanations for certain evaluations (for example, for *why* it might seem disrespectful to treat an image of someone in a way that we would not want to treat the person herself, even though she is obviously distinct from the image.) Of course, these issues may not arise in an alternative explanatory framework—one that does not make use of notions like belief, desire and pretense; but that is not my concern here.

Second: I am fully open to the possibility that, even given the first caveat, I have misdrawn the boundaries of the mental state that I am interested in, either by characterizing it too narrowly, or by classing together cases that ought to be treated as distinct. If so, I have no doubt that critics will help to characterize more precisely the notion that I am grasping at.

1. Introducing Alief

What explains the tendency of a person who has set her watch five minutes fast to rush, even when she is explicitly aware of the fact that the time is not what the watch indicates it to be? What explains her reluctance to eat fudge shaped to look

like dog feces, to drink lemonade served in a sterilized bedpan, to throw darts at a picture of a loved one—even when she explicitly acknowledges that the behaviors are harmless (Rozin *et al.*, 1986)? What makes her hesitant to sign a 'pact' giving her soul away to the devil—even if she is an atheist, and even if the pact says explicitly at the bottom 'this is not a real pact with the devil; it is just a prop in a psychology experiment' (Haidt, p.c.)? What explains the tendency of the Hitchcock expert to experience suspense as the shower scene proceeds, even though she has written a book detailing *Psycho* frame-by-frame? What explains the tendency of a chef who has recently rearranged his kitchen to walk towards the old knife drawer to get his cleaver, even as talks about how happy he is with the new set-up? What explains the propensity of subjects whose aim is to select a red ball to go with frequency (choosing from a bag with 9 red and 91 white balls) rather than probability (choosing from a bag with 1 red and 9 white balls)—even when the comparative likelihoods are prominently displayed (Denes-Raj and Epstein, 1994)?

Is it relevantly similar to what explains the tendency of the frog to snap at the BB, and the puppy to bat at the image of the dog—even when they have ample evidence that the lure and the BB aren't edible, and that the dog-image isn't a fellow canine? To what explains the tendency of the cinema-goer to cower and of the baseball fan to exhort—even as they assure their interlocutor that the images before them are nothing more than patterns of pixels on a screen? To what explains the tendency of the man in the cage to tremble—even as he acknowledges that the precipice is of no danger to him? And to what explains the tendency of the avowed anti-racist to respond differentially to Blacks and Caucasians—even as she reiterates her commitment to their equality?

It is natural to think that different explanations are called for in the different cases. Some are due to instinct, some to habit. Some are due to the operation of 'system I', others to forgetfulness. Some are due to vivid imagining, others to false belief. Some are due to superstition, others to hypocrisy. Call the outlook that underlies this cluster of explanations—along with the assumption that different explanations are needed for the different cases—the *classic cognitivist picture*.

For reasons that will become easier to articulate once I present my proposed alternative, I think the classic cognitivist picture is an unhelpful way to make sense of the cases I am interested in. It is insufficiently sensitive to certain key differences among belief, imagination, habit, and instinct, and—correspondingly—insufficiently sensitive to certain important similarities among cases that it classes as falling under those categories. Some of these explanations it proposes are simply wrong—those that credit the subjects with false belief, for example. Others are merely incomplete—such as those that appeal to vivid imagining. Yet others draw distinctions that—for the purposes of recognizing important patterns of similarity—are best overlooked in the context. (I have in mind those that appeal to habit and instinct.) By contrast, the picture that lies behind the explanation that I propose—that, at base, all of these cases involve instances of norm-discordant behavior arising

as the result of the mental state that I call *alief*[5]—helps us attend to certain important similarities among the cases, and to be sensitive to important differences among other cognitive states.[6]

The value of the account I offer lies in its reframing of the explanatory terrain, enabling us to notice similarities where there had appeared to be only differences, and differences where there appeared to be only similarities. In making this claim, I am not claiming that alief has a localized substrate, or that it is associated with distinctive neural firing patterns, or that it can be selectively inhibited, or whatever other criterion you think is required for something to be included in our 'mature science' of the mind. Rather, I am claiming that it is of explanatory utility: it helps us make sense of things.

1.1 Overview

So what is alief? To have an alief is, to a reasonable approximation, to have an innate or habitual propensity to respond to an apparent stimulus in a particular way. It is to be in a mental state that is (in a sense to be specified) *a*ssociative, *a*utomatic and *a*rational. As a class, aliefs are states that we share with non-human *a*nimals; they are developmentally and conceptually *a*ntecedent to other cognitive attitudes that the creature may go on to develop. Typically, they are also *a*ffect-laden and *a*ction-generating.[7]

Examples will follow, but in the meantime, a bit of clarification about each of the features just identified:

- *Associative*: Aliefs encode patterns of responses to particular (internally or externally prompted) mental images.
- *Automatic*: Though a subject may be consciously aware of her aliefs, aliefs operate without the intervention of conscious thought.
- *Arational*: Though aliefs may be useful or detrimental, laudable or contemptible, they are neither rational nor irrational.

---

[5] My resort to neologism gives rise to the utterly reasonable question—raised explicitly by Sylvain Bromberger (p.c.)—of why I should have been so fortunate to have discovered a category of thought that has evaded the eyes of philosophers for two millennia. As the discussion below will make apparent, however, I make no claim to novelty: at most, I am claiming to have noticed a certain commonality across some lines of thought that might otherwise have appeared disparate. A related worry—raised by J. Brendan Ritchie (p.c.)—is why we lack a folk psychological notion for alief, given that we have such notions for attitudes like belief, desire, imagination and perception. I have no convincing response to this worry, though the discussion in section 2.1.1 below may go some way towards offering an answer.

[6] This fact—that having a blanket term for a phenomenon can be philosophically useful by making us attentive to certain patterns—is itself due to the phenomenon for which I now offer a blanket term. For some very preliminary thoughts on this matter, see Gendler, 2007.

[7] Thanks to Paul Bloom, Emily Esch, Elliot Paul, and Ralph Wedgwood for pushing me to clarify the relations among these characteristics.

- *Shared by human and non-human animals*: Any creature capable of responding differentially to features of its environment that impinge upon its sensory organs has aliefs.
- *Conceptually antecedent to other cognitive attitudes that the creature may go on to develop*: Aliefs are more primitive than beliefs or desires. While it may be possible to paraphrase the content of aliefs using the language of belief and desire, alief cannot be factorized into belief and desire.[8]
- *Action-generating*: Aliefs typically activate behavioral proclivities (though these may not translate into full-blown actions), and can do so directly, without the mediation of classic conative attitudes like desire.[9]
- *Affect-laden*: Aliefs typically include an affective component.

In short, adult humans have aliefs, but so do puppies and frogs. So do babies, and so do birds and bees.[10] Indeed, it's because of the birds and the bees and the babies—that is, because of sexual reproduction and the role that it plays in underpinning certain facts about evolution—that we shouldn't be surprised that, as human animals, we share a great deal of our cognitive apparatus with other, non-human animals. Rather, it would be surprising if the opposite were the case. Much of animal behavior—both human and non-human—is the result of innate or habitual propensities to respond to apparent stimuli in particular ways. What differentiates humans (and some non-human animals) from (other) non-human animals is that some of their behavior is the result of something else.

This completes my overview of the notion of alief. In the next subsection, I offer some clarification on how I propose to use the expression.

---

[8] As Paul Davies notes (p.c.): 'Given some very conservative assumptions about the evolution of the human brain, it is overwhelmingly plausible that mental capacities that produce alief-like states evolved prior to the relatively fussy capacities we have for belief and imagination, and these latter states should be conceptualized as probable effects or evolutionary byproducts of the former. So … the suggestion that alief might be assimilated to belief and imagination … is not merely implausible: It is naïve.'

[9] This may or may not differentiate alief from other attitudes. On many pictures of belief, belief motivates only in conjunction with desire. But there are other views of belief—Nagel's (1970) for example—according to which desire contributes to belief in motivating action 'only in the sense that *whatever* may be the motivation for someone's intentional pursuit of a goal, it becomes in virtue of his pursuit *ipso facto* appropriate to ascribe to him a desire for that goal' (Nagel, 1970, p. 29). (Thanks to Jessica Moss for discussion of these issues.)

[10] It appears, for example, that both birds and bees are subject to what is sometimes called the 'asymmetric dominance' or 'decoy' effect (cf. Ariely, 2008). 'Contrary to the theory of rational preference … honeybees (*Apis mellifera*) and gray jays (*Perisoreus canadensis*) are, … like humans, … influenced by the addition of an option to a rational choice set … Their relative preference between two original options change[s] with the introduction of a third … option' that is dominated by one but not the other of the original choices (Shafir *et al.* 2002, p. 180.) A similar effect has been reported in hummingbirds (*Selasphorus rufus*) and starlings (*Sturnus vulgaris*) (Bateson, 2002; Bateson *et al.*, 2002; Hurly and Oseen, 1999.) (For a dissenting interpretation of some of these results, see Schuck-Paim *et al.*, 2004.)

1.2 Examples and Usage

Traditional propositional and objectual attitudes are two-place affairs. A subject believes (that) *b* or desires (that) *d* or hopes (that) *h* or fears (that) *f*. But alief, as I propose to use the term, involves a relation between a subject and an entire associative repertoire, one that paradigmatically includes not only representational (or 'registered') content, but also affective states, behavioral propensities, patterns of attentiveness, and the like.[11] There is no natural way of articulating this, but—as a reasonable (if cumbersome) approximation—we can say that a subject in paradigmatic state of alief is in a mental state whose content is representational, affective and behavioral: she alieves *r, a, d*. Though this usage is approximate—and in that sense, misleading—it helps to emphasize the ways in which thinking in terms of alief differs from thinking in terms of the traditional cognitive and conative attitudes.

Examples will make this usage clearer. Consider again the frog going for the BB, the puppy batting at the mirror, and the suspended man trembling in the cage. In each of these cases the norm-discordant behavioral tendency can be explained by an alief with content that might be expressed, among other ways, as follows. The frog alieves (all at once, in a single alief): small round black object up ahead; appealing in a foody sort of way; move tongue in its direction. The puppy alieves (again, all at once): dog-shaped dog-motiony creature in front of me; attractive and threatening in a my-size-conspecific sort of way; engage in (play-)fighting. The suspended man alieves (all at once): high up above the ground right now, dangerous scary place to be, tremble. Likewise for each of the additional cases presented above: in each of them, the subject's behavioral tendencies can be explained by appeal to an alief with representational, affective and behavioral content.

Though paradigmatic alief is at least a four-place relation, it is tempting to slip into the more natural two-place usage. It is natural to say that the frog alieves that the BB is a fly, or that it is edible, or that it is worth jumping towards.[12] Or that the movie-goer alieves that there is an axe-murderer in front of him, or that he is in danger, or that there are good grounds for shrieking and cowering. Or that the sports fan alieves that his team is playing right now, or that the batter is in need of his support, or that his cheer will help his team win.

This 'loose' usage may be handy in some contexts. Its naturalness makes it easy to employ (an approximation of) the concept of alief, and employing (an

---

[11] For detailed discussion and defense of this decision, see Gendler (forthcoming). For the distinction between registering and representing content, see Prinz, 2004. The notion of representation here is a thin one: because they involve mechanisms that are wholly insensitive to the difference between seeming and being, or between appearance and reality, aliefs lack certain sorts of correctness conditions. Their representational content is akin to that of Aristotle's *phantasia* (being-appeared-to). (Thanks to Elliot Paul, Emily Esch, J. Brendan Ritchie and Jessica Moss for discussion here, though I suspect I have not fully appreciated all of their suggestions.)

[12] Note, however, that when we make this move, we open ourselves to the classic teleosemantic worries. (For discussion, see Millikan, 1995.)

approximation of) the concept of alief helps us to break free from the grip of the classic cognitivist picture. Moreover, although *paradigmatic* instances of alief involve the activation of associative repertoires that saliently include representational, affective and behavioral content, there may be cases where we wish to ascribe alief where the salient content falls primarily in only one or two of these domains. In the text that follows, I will be careful to use the expression in its canonical fashion when care is required. When care is not required, I will allow myself to slip into a more familiar two-place attitude structure.

2. Alternative Explanations

The time has come to explain why I think the family of alternative explanations offered by the classic cognitivist account provide an unhelpful way of mapping the cognitive territory. That is, the time has come to say what I think is unhelpful about the natural response that in the case of the frog and the puppy, the norm-discordant response is due to instinct; that in the case of the chef, it is due to habit; that in the case of the cinema-goer and the baseball fan, it is due to vivid imagining; that in the case of the man in the cage, it is due to false belief; and that in the case of the avowed anti-racist, it is due to hypocrisy—and why I instead want to say that in each of the cases, the response is (also or instead) helpfully understood as being due to alief.

Before continuing, two caveats that may forestall certain objections. First, it is clear that the disputed subject matter does not admit of easy classification: there is no periodic table of the attitudes, or Linnaean taxonomy of mental states. So some disputes that seem substantive may turn out to be merely terminological. Though the line is not as sharp as it may seem initially—after all, terminological habits may have cognitive effects—I will nonetheless try to be careful about this in the discussion that follows.

Second: I am not denying that habit, instinct, vivid imagining, false belief and hypocrisy can give rise to norm-discordant behavioral tendencies. Indeed, understood properly, I think that in *all* of the cases we are considering the relevant behavioral tendencies *do* arise from habit and instinct. And, understood properly, I think that in the case of the cinema-goer and the baseball fan it *is* vivid imagining that activates the relevant habitual propensity. Moreover, on certain spellings-out of the avowed anti-racist case—for example, a case where the subject is vividly aware of her discordant tendencies and makes no effort to extinguish them—the relevant state might indeed *be* one of hypocrisy. And there are even versions of the caged man story in which it *would* be correct to say that the subject has two sets of conflicting alief-driven behavioral tendencies, each norm-concordant with respect to one of his beliefs, and norm-discordant with respect to the other (though there is also a version where he does not.)

In short, to say that a behavioral response is due to instinct, or habit, or imagining, or false belief, or hypocrisy does not *preclude* its being due to a norm-discordant

alief. Moreover, there may be contexts where one of the other terms provides an especially useful description of one of the cases in question. My claim is simply that it is *also* useful to have recourse to the notion of alief, and that to describe the cases without recourse to such a notion will lead us to say things that are incomplete, or misleading, or false. In the remainder of this section, I will explain what I mean by this and why I think it is so.

2.1 Appeals to Belief

Let's begin with the example of the trembling man suspended safely in the cage.[13] Here is how Hume describes the case:[14]

> … consider the case of a man, who, being hung out from a high tower in a cage of iron cannot forbear trembling, when he surveys the precipice below him, tho' he knows himself to be perfectly secure from falling, by his experience of the solidity of the iron, which supports him (Hume, 1739/1978, p. 146).[15]

Hume prefaces his story by describing it as a 'familiar instance'. And, as I have learned from Saul Traiger, 'precipice thought experiments were common fare in a philosophical debate about reason and the passions in Hume's predecessors' (Traiger, 2005, p. 100.) Montaigne, for example, notes that 'if you place a sage on the edge

---

[13] Hume presents the case in his underappreciated chapter on 'unphilosophical probability' (1739/1978, I.iii.13), which is studded with striking psychological insights. In discussing the 'second unphilosophical species of probability' for instance, he foreshadows Kahneman and Tversky's *availability heuristic* (Tversky and Kahneman, 1973.) Hume writes: 'A lively impression produces more assurance than a faint one; because it has more original force to communicate to the related idea, which thereby acquires a greater force and vivacity. A recent observation has a like effect; because the custom and transition is there more entire, and preserves better the original force in the communication. Thus a drunkard, who has seen his companion die of a debauch, is struck with that instance for some time, and dreads a like accident for himself: But as the memory of it decays away by degrees, his former security returns, and the danger seems less certain and real' (Hume, 1739/1978, Book I.iii.13.2.). In discussing the 'fourth species'—a 'species of probability, deriv'd from analogy, where we transfer our experience in past instances to objects which are resembling, but are not exactly the same with those concerning which we have had experience'—he discusses not only the case of the man suspended in the cage, but also the example of racial stereotyping. In future work, I hope to devote an essay-length discussion to Hume's views on these matters.

[14] Interpreting the case in the Humean context is complicated, given Hume's associationism, and his correspondingly idiosyncratic conception of belief. I will set these issues aside in the discussion that follows. My humility in advancing any sort of interpretation has been reinforced by the insights gained from reading the following articles and books: Hearn, 1970; Falkenstein, 1997; Passmore, 1952; Loeb, 2002. (Many thanks to Ken Winkler for guidance concerning this literature.)

[15] Note that the explanation cannot be due to the actual danger of the situation. Exactly the same response would be evoked if one were actually on the ground, subjected to an optical illusion as of being suspended. Indeed, much the same response would be evoked even if one *knew* that one was on the ground, being subjected to such an optical illusion.

of a precipice he will shudder like a child' (Montaigne, 1958, p. 250).[16] And Pascal and Malebranche both consider such cases. (I discuss their versions below.)

Such cases have a common structure. As a follower of Plato might put it, the 'rational part of the soul' pulls in one direction; the 'spirited' or 'appetitive' part pulls in another. Or, as the early moderns would say, 'reason' inclines the subject one way, 'passion' inclines him another. While there are many ways of understanding such cases, here are two that I think are deeply misleading. As Traiger writes:

> Philosophers 'deployed precipice examples to support one of the following claims: (1) Affective mechanisms can *lead to beliefs* which we must embrace, but which are incompatible with the beliefs we are led to by causal reasoning. (2) Affective mechanisms *make it impossible to form beliefs* that would have been arrived at through reasoning in the absence of the affective response' (Traiger, 2005, p. 101, emphasis added).

On both interpretations, the precipice case is taken to be one that tells us something about the trembling man's *beliefs*. On the first view, the man's tendency to tremble shows that he believes (on one sort of ground) that he is in danger of falling, while his tendency to avow his safety shows that he believes (on another sort of ground) that he is not in danger of falling.[17] Call the assumption that lies behind this view: that *behavior reveals belief*. On the second view, the man's tendency to tremble shows that—even in the face of his avowals to the contrary—the man does not truly believe that he is safe.[18] Call the assumption that lies behind this view: that *hesitation precludes belief*.

---

[16] Or again: 'Put a philosopher in a cage of small bars of thin iron suspended at the top of the towers of Notre Dame de Paris, he will see for obvious reasons that it is impossible for him to fall, and yet (unless he is used to the roofer's trade) he will not be able to keep the vision of that height from frightening and astonishing him … Set a plank between those two towers, of a size such as is needed for us to walk on it: there is no philosophical wisdom of such firmness as to give us the courage to walk on it as we would do if it was on the ground (Montaigne, Essays II, ch. 12 (in Montaigne, 2003, p. 155).

[17] Cf. Louis Loeb: 'Suppose the man has observed that whenever he sees a precipice and has not been suspended, he has fallen … Suppose the man has observed that whenever he is suspended, he has not fallen. Suppose that for the first time the man both sees a precipice and is suspended. He will have the inclination to *believe* both that he will fall and that he will not fall … Here we have the presence of contrary *beliefs*' (Loeb, 2002, p. 107, emphasis added). Matters are a bit complicated here, since at times Loeb appears to be using 'belief' in a Humean sense (according to which beliefs differ from other (non-committal) mental states only in their degree of vivacity). But it seems clear from the context that at least the final sentence is *in propria persona*.

[18] Cf. Eric Schwitzgebel: 'Many Caucasians in academia sincerely profess that all races are of equal intelligence. Yet I suppose that many of these same people would also be less quick to credit the intelligence of a black student than a white or Asian student, feel some (perhaps suppressible) twinge of reluctance before hiring a black person for a managerial job requiring mental acuity, expect slightly less from a conversation with a black custodian than a white one—and, in short, reveal through their actions a pervasive if subtle racism. Such people, you will perhaps agree, don't fully and completely believe in the intellectual equality of the races, as genuine and unreserved as their rebukes of racism might be' (Schwitzgebel, manuscript).

It is easy to see how these two strategies would play out in our other cases. In the case of the cinema-goer, the first strategy would say that the viewer's tendency to shriek and cower shows that he believes (on one sort of ground) that he is in danger of being attacked by the creature on the screen, while his contrary tendency to remain in the room shows that he believes (on another sort of ground) that he is in no such danger. The second strategy would say that the cinema-goer lacks beliefs about his safety, because the tension between his reasoning and affective mechanisms make it impossible to form the relevant beliefs. In the case of Rozin's subjects, the first strategy say that the subject believes both that throwing a dart at a picture of a loved one will harm the loved one, and that it will not. The second would tell us that that the subject lacks a belief about the harmfulness of throwing a dart at a photograph. And so on. For ease of reference below, let's call the first of these two attitudes the subject's attitude towards the *real content* (e.g. safety) and the second his attitude towards the *merely apparent content* (e.g. danger.)

Both of these readings characterize the competing tendencies—the man's tendency to aver that he is safe or that throwing the dart is harmless (his response to the real content), and his tendency to tremble or hesitate (his response to the merely apparent content)—as being on a par. On the first view, each of the competing tendencies is seen as sufficient to credit the subject with the relevant belief; on the second, the latter tendency (trembling or hesitation) is seen as sufficient to undermine crediting the subject with the belief associated with the former (safety.) I will argue that the inclination to treat these tendencies as on a par reflects a picture of the relation between belief and behavior that is both deeply natural, and deeply mistaken. I will address the issue of naturalness first, and the issue of mistakenness second.

My discussion will proceed as follows. First, I will offer a diagnosis of why the behavioral account is so appealing, tracing its attractiveness to the 'feeling of naturalness' that attaches to it. I will argue that it is a mistake to take this feeling as an indicator of appropriate attribution, since there are many cases where we feel the pull to attribute belief but where such attribution is clearly mistaken. Second, I will argue that there is a distinct role that the notion of belief needs to play in our cognitive repertoire if it is to bear the relation to knowledge and rationality that philosophers require of it. In particular, in order for an attitude to count as a belief, the attitude needs to be responsive to changes in the world, and in our evidential relation to it. I will argue that the attitude present in the cases presented above does not satisfy these criteria.

2.1.1 The Attraction of the Behavioral Account. The tendency to 'infer' intention from action is deep-seated and automatic. We are all inclined to attribute intentions, beliefs, emotions, and personality traits to Heider and Simmel's (1944) moving triangles[19] ('The big triangle wants to get out the door' 'The little triangle keeps trying to block him')—even if on reflection we do not think that geometric

[19] Cf. Heider and Simmel, 1944. Note that the 'we' here includes infants and non-human primates: 'Numerous studies have since demonstrated this automatic attribution of high-level mental states to animate motion in adults in a wide range of cultures, young infants, and even chimpanzees' (Blakemore and Decety, 2001).

line-figures apparently moving across a screen could want or think or try. We are all inclined to experience ATMs and computers and cars as having mental states ('The machine won't believe that I don't want a receipt'; 'My car always wants to turn left when I leave the driveway')—even if on reflection we do not think that inanimate objects could be bearers of beliefs and desires. And the tendency emerges at higher levels as well, perhaps as a result of the same mechanisms, perhaps as the result of different ones. (Even with respect to ourselves, we are inclined to take an interpretative intentional stance, reading our own 'beliefs' off of our behavior. Post-hypnotic and left-brain/right-brain confabulation provide the most extreme examples; cognitive dissonance provides another.) (For related discussion, cf. Dennett, 1987; Carruthers 2006, manuscript.)

To put the point somewhat coyly and self-referentially: when subjects encounter patterns of motion that resemble genuine intentional actions, they have the habitual propensity to respond as if they were in the presence of an agent with beliefs and desires. (For those already convinced of the utility of the notion of alief, we might say: they come to have occurrent aliefs that they are in such a circumstance.) When the soda machine repeatedly returns the patron's dollar, the patron has the habitual propensity to respond as if the soda machine was an intentional agent who believes the dollar to be fake. (We might say: the patron alieves that the machine believes the dollar is fake.) When Heider and Simmel's subjects observe an animated triangle moving in a certain way, they have the habitual propensity to respond as if the triangle were an intentional agent looking for a door. (We might say: Heider and Simmel's subjects alieve that the triangle desires to find the door.) When a subject in a cognitive dissonance experiment observes that she has ended up with the red ball instead of the blue one, she finds herself in a situation normally associated with her preferring the red ball to the blue one. So her habitual propensities associated with her preferring the red ball to the blue one are activated. (We might say that she comes to occurrently alieve: nice red ball, appealing, happily retain possession.) And so on.

Let's get back to the larger dialectic. These examples are meant to discount one possible argument in favor of the parity accounts. What they show is that our natural inclination to treat something as indicative of belief regularly misfires, so that the presence of this natural inclination cannot be taken as decisive evidence for the correctness of the attribution.

Of course, there are also theoretical reasons that one might embrace such an equivalence. One might hold, for example, that there is a conceptual connection between belief and behavior, so that 'All that's necessary for an attitude to qualify as a belief is that it disposes the subject to behave in certain ways that would promote the satisfaction of his desires if its content were true. An attitude's tendency to cause behavioral output is thus conceived as sufficient to make it a belief' (Velleman, 2000, p. 255).[20]

[20] Velleman rejects this account, but goes on to give a long list of philosophers whom he contends have endorsed (some version of) it. These include: Braithwaite (1932-1933); Armstrong (1973); Quine and Ullian (1978); Stalnaker (1984); Baker (1995); and Dennett (1971, 1995).

As stated, this can't be right: belief-desire explanations are supposed to explain (or 'rationalize') *intentional* actions—not mere behaviors. But of course, that's precisely what is at issue in the cases we are considering. There is no question that the subjects' attitudes towards the merely apparent content dispose them 'to behave in certain ways that would promote the satisfaction of [their] desires if [that] content were true'. But are those behaviors intentional in the relevant sense? Presumably this is not something that can be read off the behaviors themselves. And to the extent that reflective verbal report can distinguish the cases, it tells against the intentional reading. (Did you really believe that there was an axe-murderer approaching you? that throwing the dart at the photograph would harm your loved one? that the metal bars were really not strong enough to hold you?)

2.1.2 The Mistake behind the Belief Interpretation. It remains to be established that the belief interpretation is misguided in cases such as precipice (where the attribution is most tempting.) I will base this argument on the role that belief needs to play in our cognitive repertoire. The defense can be made without offering a full-fledged account of belief.[21] All that is needed is to note that—whatever belief is—it is normatively governed by the following constraint: belief aims to 'track truth' in the sense that belief is subject to immediate revision in the face of changes in our all-things-considered evidence.[22] When we gain new all-things-considered evidence—either as the result of a change in our evidential relation to the world, or as a result of a change in the (wider) world itself—the norms of belief require that our beliefs change accordingly. I used to believe that stomach ulcers were caused primarily by stress and diet; but when Warren and Marshall's research on the *Helicobacter pylori* bacterium became widely known, I revised my belief to reflect this information.[23] Williamson's 'N.N'.—'who has not yet heard the news from the theatre where Lincoln has just been assassinated'—believes that Lincoln is

---

[21] Here are some features beliefs are supposed to have: Belief 'tracks truth'. It is 'responsive to evidence'. It is intimately connected with notions like knowledge and rationality. It gives rise to Moore's paradox, and its strength can be ascertained using Ramsey's methods. 'Believing *p* is, roughly, treating *p* as if one knew *p*' (Williamson, 2000, pp. 46-7.) Or, for those whose philosophical temperament inclines them towards Pittsburgh rather than Oxford. Belief falls 'within … the space of reasons:' 'A belief … is an actualization of capacities of a kind, the conceptual, whose paradigmatic mode of actualization is in the exercise of freedom that judging is. This freedom, exemplified in responsible acts of judging, is essentially a matter of being answerable to criticism in the light of rationally relevant considerations' (McDowell, 1998, p. 434).

[22] For classic discussion, cf. Williams, 1973. For recent discussion, see Velleman, 2000; Owens, 2003; Wedgwood, 2002; and Velleman and Shah, 2005. Cf. also Hieronymi, 2008, and Smith, 2005.

[23] It is an interesting fact—though one that will have to wait for another paper—that the medical community was apparently quite reluctant to accept these data as decisive. This introduces an important complication, namely, that numerous features may affect what a subject takes to be evidentially relevant, and that motivated attention to or ignoring of (apparent) evidence plays a major role in the formation even of belief in this narrower sense.

President; but as soon as he learns that Lincoln has been shot, he will make the corresponding adjustment in his belief (Willamson, 2000, p. 23).[24]

In each of the cases we have been considering, only *one* of the competing tendencies is evidence-sensitive in this way. The man suspended in the cage *believes* that he is safe because if he were to gain evidence to the contrary, his attitude would change accordingly. So too with Rozin's subjects, the baseball fan, the cinema-goer, and the rest. One—and only one—of the two behavior-generating attitudes can turn on a dime[25] in this way, even in the face of apparent sensory evidence to the contrary. This gives reason to treat the two as not being on a par.

Indeed, the argument can be made on the following simple grounds: Beliefs change in response to changes in evidence; aliefs change in response to changes in habit. If new evidence won't cause you to change your behavior in response to an apparent stimulus, then your reaction is due to alief rather than belief.[26] (Of course, there are strategies for changing aliefs as well—but these run through sub-rational mechanisms.)

2.1.3 Conclusion: Alief is not Belief. In conclusion, I think precipice examples show neither that 'affective mechanisms can lead to beliefs which we must embrace, but which are incompatible with the beliefs we are led to by causal reasoning' nor that 'affective mechanisms make it impossible to form beliefs that would have been arrived at through reasoning in the absence of the affective response' (Traiger, 2005, p. 101.) Rather, what precipice examples show is that affective mechanisms associated with habitual propensities to behave in particular ways may be predictably triggered by certain apparent stimuli. These alief-generated behaviors are naturally 'read' as indicative of belief (or as preclusive of belief to the contrary.) But this is a mistake. The behavioral account rests on an overextension of a heuristic: it depends on treating something that is a *general indicator* of belief as if it were a *necessary and sufficient correlate* of belief. The belief that behavior invariably indicates belief arises from aliefs that are mistaken for beliefs.

---

[24] Aliefs will be slower to change. As Hume notes, 'After the death of any one, 'tis a common remark of the whole family…that they can scarce believe him to be dead, but still imagine him to be in his chamber or in any other place, where they were accustomed to find him'. (Hume, 1739/1978, I.iii.9.) Cf. also Kübler-Ross's five stages of grief, where 'acceptance' may be understood as the stage at which one's beliefs that a loved one has died come to be matched by the relevant aliefs.

[25] The British equivalent, apparently, is 'turn on a sixpence' (http://en.wiktionary.org/wiki/turn_on_a_dime.)

[26] As stated, this principle is too strong, for there are certainly cases of subjects who hold evidence-recalcitrant *be*liefs. (Theists and atheists each consider the other to be an example. More mundanely, think about flat earthers, Roswellians, or your political opponents.) Cases of evidence-recalcitrant belief tend to be cases where the subject somehow *distorts* the evidence that is available to her through selective attention or sophisticated weighting. But this is a preliminary response at best: the matter requires further thought. (Thanks to Elliot Paul for pressing me on this issue.)

2.2 Appeals to Imagination and Pretense

Here is a natural response to many of the cases I have been discussing. The subject believes one thing (say, that affixing his signature to the piece of paper will have no practical consequences) but imagines or pretends another (say, that affixing his signature to the piece of paper will result in his soul belonging to the devil.) The content that he imagines or pretends conjoins with a (perhaps imaginary) desire to act in accord with that imagined content, resulting in the observed behavioral repertoire. (For recent discussion, cf. Currie, 2002 and responses thereto.)

Again, it is easy to see how this would go in other cases. The cinema-goer believes that she is safe, but imagines that she is in danger; Rozin's subjects believe that ingesting the food or throwing the dart is harmless, but imagine that doing so is harmful; the frequency/probability subject believes that she has a better chance of pulling out a red ball if she draws from the bag with the greater proportion of reds to whites, but imagines that she has a better chance if she draws from the bag with the greater frequency of reds. And so on. Here, for example, is Pascal on the precipice case: 'Put the world's greatest philosopher on a plank hanging over a precipice, but wider than it needs to be. Although his reason will convince him of his safety, his *imagination* will prevail' (Pascal, 1669/1966, S78/L44, 13, italics added).

As I indicated above, I do think imagining is (in some of the cases) part of the story. But not in the way traditional accounts suggest. This can be brought out by contrasting what goes on our cases with what goes on both in cases of voluntary pretense, and in cases of involuntary imagining as traditionally understood.

Suppose I am engaged in a classic game of pretense, where I pretend that a banana is a telephone.[27] I hold the banana to my ear, and say 'Hello: I'd like to order 100 large pizzas'. Here my action is the result of an imagined belief (this is a telephone) and an imagined desire (I'd like to order 100 pizzas): together these combine to produce the behavior in question. But here each of the pieces—the belief, the desire and the action—can plausibly be prefixed with 'make-': I make-believe that I am holding a telephone; I make-desire that I wish to order a pizza; and I make-behave that I am doing so.[28] My action in 'ordering' the 'pizza' is a controlled and deliberate, one that I can regulate at will. And it is one that takes place within a circumscribed realm of the merely pretend; my voice and body serve as props in a game of make-believe in much the same way that the banana does. So though it uses the same equipment as actual actions (my body, my voice, etc), the resulting behavior *represents* the content in question, rather than *manifesting* it; clearly this is not what is going on in our cases above. Moreover, in such cases the action in question is 'flagged'—both by the performer and by the perceiver—as merely symbolic. Both children and adults show a marked ability to distinguish

---

[27] Discussion in this paragraph draws on ideas from Gendler, 2006.

[28] Children are enormously adept at these sorts of games; for review, cf. Harris, 2000. For evidence regarding their ability to quarantine between games, cf. Skolnick (Weisberg) and Bloom, 2006.

such 'decoupled' actions from their ordinary counterparts. (For review, cf. Harris, 2000. I discuss these matters in more detail in Gendler, 2003 and 2006.)

So deliberate pretense is not a good model for our cases. But what about externally prompted involuntary imagining? Here the answer is more complicated. I *do* think that imagination plays an important role in (some cases of) alief—but when it does so, it does so by violating one of the norms of imagination. Imagination, like Las Vegas, is governed by a norm of *quarantining*: what happens in imagination stays in imagination. Our actual, real world, non-pretend actions aren't supposed to be guided by things that happen in *that* part of the mind.[29] But of course, this is exactly what happens in (some) cases of norm-discordant alief-generated action: a behavioral repertoire that is activated by merely imagined content manifests itself in observable actions or proclivities.[30]

So the behavioral response in precipice examples *can*, in an important sense, be traced to the imagination. But this does not mean that alief and imagination are the same. Some cases of imagining—at least in principle—do not give rise to these sorts of behavioral propensities. And some cases of alief-generated behavioral response—consider the frog and the puppy—are not the result of imagining. Imagination gives rise to behavior via alief. What happens in imagination may have (non-pretend) effects beyond imagination—but it does so when the process of imagining activates a subject's innate or habitual propensity to respond to an apparent stimulus in a particular way.

2.3 Appeals to Habit and Instinct

The time has come to address another natural objection. Even if the cases do all involve—as I have argued—an innate or habitual propensity to respond to an apparent stimulus in a particular way, what justification is there for treating these etiologically distinct cases as relevantly similar? The frog and the puppy and the vertigo-sufferers are responding as they are hard-wired to respond, whereas in many of the other cases, the source is merely habitual. Moreover, in some of the cases the response is highly impermeable to deliberate regulation, whereas in others, direct control seems possible.

Theses differences are indeed important. And they are ones to which philosophers worrying about (what I would call) alief were quite sensitive. Here, for example, is Malebranche, discussing the precipice case, invoking a sharp distinction between innate and merely habitual propensities (and offering a fine characterization of an alief-state as he does so):

> There are traces in our brains that are naturally tied to one another, and even to certain emotions of the spirits, because this is necessary to the preservation

---

[29] Cf. Nichols and Stich, 2000.

[30] Popular psychology, of course, is replete with strategies that exploit this leakage, as any self-help section will reveal: there is even a journal entitled th*e Journal of Imagery Research in Sport and Physical Activity*. For related discussion, see Gendler, 2006.

> of life; and their connection cannot be broken, or at least cannot be easily broken, because it is good that it always be the same. For example, the trace of a great elevation that one sees below oneself…is naturally tied to the one that represents death to us, and to an emotion of the spirit that disposes us to flight and to the desire to flee, The connection never changes, because it is necessary that it be always the same, and it consists in a disposition of the brain fibers that we have from birth (Malebranche, 1712/1997, p. 106).

By contrast, he continues:

> All the connections that are not natural can and should be broken, because different circumstances of time and place are bound to change them, so that they can be useful to the preservation of life…Thus, it is necessary for the conservation of all animals that there be certain connections of traces that can be easily formed and destroyed, and that there be others that can be broken only with difficulty, and finally, still others that can never be broken (Malebranche, 1712/1997, p. 106).

That is, 'connections of traces'—aliefs—differ in their etiology, and in their corresponding degree of malleability.

There is no doubt that etiology matters for some things: perhaps sunburn can only be produced by the sun. But it seems unmotivated in this case to distinguish propensities that result (directly) from the experiences and actions of a particular individual from those that result (indirectly) from the experiences and actions of her ancestors, merely on those grounds. Certainly we make no such distinction in the case of beliefs and other mental states, else the debate between Locke and Leibniz would have taken a very different form.[31]

A parallel response can be made to the argument that innate propensities are fixed while habitual propensities are malleable. For it is not so clear either that this difference obtains in a relevant sense, or that malleability is what really matters. Recent studies of plasticity suggest that in both cases, the development of alternative propensities proceeds through bypass, not erasure. As a recent *New York Times* article admonishes, 'don't bother trying to kill off old habits; once those ruts of procedure are worn into the hippocampus, they're there to stay'.[32] And, as the

---

[31] Interesting related discussion can be found in the writings of William James, George Herbert Mead and John Dewey, each of whom is highly sensitive to the important similarities between what Mead called 'inherited and acquired habits' (Mead, 1938, p. 68.)

[32] http://www.nytimes.com/2008/05/04/business/04unbox.html?emandex=1210910400anden=3259989c860445c2andei=5070
Cf. Tim Wilson *et al.*: 'When an attitude changes from A_1 to A_2, what happens to A_1? Most theories assume, at least implicitly, that the new attitude replaces the former one. The authors argue that a new attitude can override, but not replace, the old one, resulting in dual attitudes. Dual attitudes are defined as different evaluations of the same attitude object: an automatic, implicit attitude and an explicit attitude … Even if an explicit attitude changes, an implicit attitude can remain the same' (Wilson *et al.*, 2000).

discussion in section 4 will make clear, the processes by which we go about regulating unwanted discordant alief are the same, regardless of whether the aliefs are innate or acquired.[33]

That said, I remain open to the possibility that there are distinct subspecies of alief: innate and habitual, perhaps—or controllable and uncontrollable. All that matters to my argument is that these subspecies be more similar to one another than they are to other candidate states. And of this I remain convinced.

3. Norm-concordant and Norm-discordant Aliefs

Aliefs activate behavioral propensities. So (in conjunction with desire) do beliefs (and their teleofunctional analogues). Sometimes these behavioral propensities pull in opposite directions; sometimes they coincide. When they pull in opposite directions, the subject's belief-discordant behavioral tendencies are governed by what I have been calling *norm-discordant aliefs*. When they coincide, the subject's belief-concordant behavioral tendencies may be consciously regulated by her beliefs, or they may be governed by what I will call *norm-concordant aliefs*.

The main focus of this essay is on norm-discordant aliefs, and on the ways in which these sorts of aliefs are problematic. But it will also be worth saying a bit about cases where norm-discordant aliefs are desirable,[34] and also about cases—which in the well-lived life are the rule rather than the exception—in which behavior is governed by norm-concordant alief.

Given the nature of alief and belief, it is inevitable that there will be cases where alief-generated propensities and belief-generated propensities activate contrary behavioral repertoires. The reason is simple: Aliefs involve habitual responses to apparent actual stimuli, but things may not be as they seem, the world may change, and one's norms may demand that the way things are is not the way things ought to be. Aliefs by their nature are insensitive to the possibility that appearances may misrepresent reality, and are unable to keep pace with variation in the world or with norm-world discrepancies. By contrast, beliefs are (modulo error) responsive to the way things are: not merely to the way things tend to be or to the way things seem to be. Actions generated by beliefs are generated by a mental state that is proportioned to all-things-considered evidence and subject to rational and normative revision; actions generated by aliefs are generated by a mental state that

---

[33] Cf. Descartes: 'Although the movement of each gland seems to have been joined by nature to each of our thoughts from the beginning of our life, one can nevertheless join them to others by habituation … So when a dog sees a partridge, it is naturally inclined to run toward it, and when it hears a gun fired the noise naturally incites it to run away. But nevertheless setters are commonly trained so that the sight of a partridge makes them stop and the noise they hear … when the bird is fired on, makes them run up to it … [In such a way] even [men] who have the weakest souls could acquire a most absolute dominion over all their passions if one employed enough training and skill in guiding them' (Descartes, 1649/1989, para 50).

[34] Thanks to Daniel Bonavec for encouraging me to think about these sorts of cases.

is not. (See section 2.1.2 above.) So it should come as no surprise that human animals are rife with (the tendency to manifest) belief-discordant aliefs, and that our non-human counterparts are rife with (the tendency to manifest) teleofunctional-discordant aliefs.[35]

Since teleofunctions and beliefs (in conjunction with right desires) generally activate propensities to act and react in ways that we (think we) should, discordant aliefs must, by definition, generally activate propensities to act and react in ways that we (think we) shouldn't. To put things a bit too simply: except in certain exceptional cases (discussed in the next paragraph), teleofunctionally-discordant aliefs predispose (human and non-human) animals to behave in ways that violate their (local) self-interest, and belief-discordant aliefs predispose (human) animals to behave in ways that violate their intention to regulate their behavior according to some norm.[36]

That said, there are cases where local self-interest seems unharmed—even aided—and where freedom seems unimpeded—even enhanced—by the presence of norm-discordant alief. Theater, cinema, novel-reading, video games, board games, poetry, metaphor, circumlocution, daydreaming, therapy, roller-coasters and bungee jumping all exploit—in various ways—our tendency to respond to merely apparent stimuli in a habitual ways. Circumscribed indulgence of these associative chains is crucial to a richly-lived human life.[37] Further discussion of this issue will take us too far afield, but I hope to discuss these matters in greater detail in future work.[38]

---

[35] As Hume notes: 'In almost all kinds of causes there is a complication of circumstances, of which some are essential, and others superfluous; some are absolutely requisite to the production of the effect, and others are only conjoin'd by accident. Now we may observe, that when these superfluous circumstances are numerous, and remarkable, and frequently conjoin'd with the essential, they have such an influence on the imagination, that even in the absence of the latter they carry us on to the conception of the usual effect, and give to that conception a force and vivacity, which make it superior to the mere fictions of the fancy. We may correct this propensity by a reflection on the nature of those circumstances: but 'tis still certain, that custom takes the start, and gives a bias to the imagination' (Hume, 1739/1978, p. 147).

[36] There is an important strand of Western philosophical thought according to which action governed by norm-discordant alief is not just undesirable, but unfree. Here, for example, is Milton, anticipating Kant:

> Since thy original lapse, true Liberty
> Is lost, which always with right Reason dwells
> Twinn'd, and hath from her no divided being:
> Reason in man obscur'd, or not obeyed,
> Immediately inordinate desires
> And upstart Passions catch the Government
> From Reason, and to servitude reduce
> Man till then free.
>
> (Milton, 1667/1980 Book XII, 83-90)

'True liberty' on such a picture 'dwells Twinn'd … always with right Reason' and when 'upstart Passions catch the Government from Reason' man is 'reduce[d] … to servitude.' I hope to explore this theme further in another essay.

[37] Indeed, one of the deficits characteristic of those on the autistic spectrum is an inability or unwillingness to indulge in this way. This may be connected in interesting ways with a corresponding propensity not to engage in games of spontaneous pretense.

[38] Some of these topics are discussed under the rubric of what Paul Rozin has dubbed 'benign masochism'. See, for example, Rozin, 1999.

Typically, though, teleofunctionally-concordant aliefs predispose (human and non-human) animals to behave in ways that accord with their (local) self-interest, and belief-concordant aliefs predispose (human) animals to behave in ways that accord with their intention to regulate their behavior according to some norm. According to the ancient ideal, a well-functioning soul is one where, so to speak, alief and belief are in accord.[39] Plato writes:

> One who is just … regulates well what is really his own and rules himself. He puts himself in order, is his own friend, and harmonizes the three parts of himself … He binds together those parts and any others there may be in between, and from having been many things, he becomes entirely one, moderate and harmonious. Only then does he act' (*Republic*, 443de in Plato, 380BCE/1992).

The ideal that a person will act only after he has put 'himself in order, harmonize[d] the … parts of himself … and … become[] entirely one, moderate and harmonious' is what lies at the heart of the aspiration that belief, desire and action form a neat inter-derivable triangle. For only in such a case will belief be readable off of action and (presumed) desire. But, as the opening examples and subsequent discussion have brought out, this ideal is (inevitably) unrealized. In the final section, I consider some of the implications of this internal disharmony.

4. Regulating Unwanted Discordant Alief

As beings who are simultaneously embodied and capable of rational agency, the challenge of bringing our aliefs into line with our commitments is one that we face repeatedly. Given that we all have norm-discordant aliefs that we disavow, and whose influence on our actions we wish to reduce, what can be done? In this final section, I offer some preliminary remarks on this question.

4.1 Traditional Strategies

Both the ancient and early modern traditions are replete with strategies for bringing our aliefs into line with our considered commitments. The recommendations fall under two main rubrics: the first stresses the value of cultivating norm-concordant habits through actual rehearsal; the second brings out how (otherwise occurrent) norm-discordant aliefs can be regulated through the refocusing of attention

[39] I have been helped in my thinking about these questions by John Cooper (1999); G. R. F. Ferrari (2007); Hedrik Lorenz (2006); Jessica Moss (2005, forthcoming); A. W. Price (1994); and C. D. C. Reeve (1988). Cf. also the essays collected in Barney, Brittain, and Brennan (forthcoming).

(especially by directed imagination), thereby redrawing the lines of internal association. Here is Aristotle discussing the first:[40]

> [W]e learn a craft by producing the same product that we must produced when we have learned it, becoming builders, e.g. by building, and harpists by playing the harp; so, also, then we become just by doing just actions, temperate by doing temperate actions, brave by doing brave actions … a state [of character] arises from [the repetition of] similar activities … It is not unimportant, then, to acquire one sort of habit or another, right from our youth; rather, it is very important, indeed all-important (*Nicomachean Ethics*, 1103a30-1103b25 in Aristotle, 1999).[41]

One way, then, to cultivate aliefs in line with our reflective commitments is to make a conscious effort to behave in the ways that our commitments dictate, so that these patterns of behavior become familiar and habitual. (Of course, the discordant aliefs may also remain (hence the warning that it is 'not unimportant … indeed, all-important … to acquire one sort of habit or another, right from our youth'); but they will be so outweighed by the concordant aliefs that something close to harmony will be achieved.)

We can also make use of the resources of the imagination.[42] Descartes, for example, 'notes that an effective way of countering an undesirable passion is to imagine a new and different state of affairs, or response to the state of affairs'. Such 'voluntary, imaginative practice' may ultimately 'reshape our internal bodily 'dispositions' so that they produce specific passions under the appropriate, rationally endorsed circumstance' (Schmitter, 2006; *cf.* Descartes, 1649/1989). And Malebranche holds that '[a]ctively deploying the imagination … can generate the entire train of sensations and emotions typical of that passion we deem appropriate.' Using these techniques, we 'can … resist the pernicious influences of the imagination caused by recalcitrant passions.' By 'training ourselves to associate

---

[40] On one natural reading of Aristotle's view, moral virtue is a matter of alief: it is a *hexis*—'a tendency or disposition, induced by our habits, to have appropriate feelings'. Cf. Kraut, 2001/2007. (There are, however, passages in the *Ethics* that appear to give more weight to consciously-regulated mechanisms such as 'choice' (*prohairesis*) and 'action' (*praxis*) (for example, *NE* II.5 (1106a3) and *NE* II.6 (1106b26-36) in Aristotle, 1999 and Barnes, 1984. For related discussion, see Irwin, 1975 and Sorabji, 1980.) (Thanks to Ralph Wedgwood and Jessica Moss for this corrective.)

[41] Here is Descartes in a similar vein: 'We cannot continually pay attention to the same thing; and so, however clear and evident the reasons may have been that convinced us of some truth in the past, we can later be turned away from believing it by some false appearance unless we have so imprinted it on our mind by long and frequent meditation that it has become a settled disposition with us. In this sense the scholastics are right when they say that virtues are habits' (Letter to Elizabeth, September 15, 1645, CSM III 267/AT 295; in Descartes, 1645/1991, p. 267). (Thanks to Elliot Paul for this reference.)

[42] Detailed discussion of this enormously rich and exciting literature will need to wait for another context. For a tantalizing overview, see Schmitter, 2006, especially the Supplementary Document concerning individual philosophers.

some thought with whatever arouses our passion, we can redirect the accompanying bodily movements as we see fit. Doing so repeatedly produces a habituation that changes our dispositions for actions and passions' (Schmitter 2006; cf. Malebranche, 1712/1997).

In the final sections of this paper, I will examine analogues of these strategies in a particular contemporary context—that of reorienting unwanted racist alief.

4.2 Reorienting Racist Alief[43]

The literature on (what I would call) racist alief is large and highly consistent, at least in its broad outlines.[44] While it appears that there is a small portion of American Whites who 'do not experience the automatic activation of any negative evaluation from memory on encountering a Black person' and some portion who 'have no qualms about their experiencing such negativity or about expressing it', (Fazio *et al.*, 1995, p. 1025) many American Whites seem to be what Jack Dovidio calls 'aversive racists'—people who consciously endorse egalitarian values, but who have negative feelings towards the relevant racial group that are 'typically excluded from awareness' (Gaertner and Dovidio, 1986, p. 62; cf. Wilson *et al.*, 2000; Dovidio and Gaertner 2004.) Even among those who are explicitly and sincerely committed to anti-racism, then, the legacy of having lived in a society structured by hierarchical and hostile racial divisions retains its imprint. So, for example, White subjects primed with images of Black faces tend to be faster to identify an ambiguous image as a gun, and more likely to misidentify a (non-gun) tool as a gun (Payne, 2001.) Otherwise identical resumés bearing stereotypical black names (e.g. Jamal, Lakisha) are less likely to result in interviews than resumés bearing stereotypical White names (Emily, Greg) (Bertrand and Mullainathan, 2003.) In both Black and White Americans, fMRI scanning shows greater amygdale activity—associated with detection of threat—in subjects presented with images of outgroup (different race) as opposed to in-group (same race) members (Amodio *et al.*, 2003.) And so on (Devine *et al.*, 2002; cf. also Payne, 2006).

Some of the tendencies of aversive racism can be countered at the level of belief through deliberate control, or through indirect manipulation.[45] Conscious application

---

[43] Many thanks to Carolyn Caine for excellent research assistance on this section.

[44] For classic discussion, see Allport, 1954. For an overview of some of the philosophical issues involved, with useful bibliography, see Kelly and Roedder, 2008. Cf. also (among others) Alcoff, 2006; Blum, 2002; Levine and Pataki, 2004; and Sullivan, 2006. I focus here on the case of race; parallel literatures exist concerning other sorts of bias.

[45] As Descartes notes, it is often possible to control our passions only indirectly. 'If someone wills to dispose his eyes to look at an extremely distant object, this volition makes the pupils dilate … But if he thinks only of dilating the pupil, he may well have the volition but he will not thereby dilate it … Our passions cannot likewise be directly excited or displaced by the action of our will, but they can be indirectly by the representation of things which are usually joined with the passions we will to have and opposed to the ones we will to reject' (Descartes, 1649/1989, paras 44-45.)

of stereotype-generated information can be regulated, for example, by providing external motivation for subjects to act in nonprejudiced ways, by encouraging subjects to be aware of egalitarian norms and standards, or by setting goals for subjects that require them to acquire unique information about group members.[46] But what about the *unconscious* or *quasi-conscious activation* of stereotypical responses—that is, what about racist alief? How, if at all, can this be regulated?

Whatever techniques are available, they will need to be strong enough to balance the effects of enormously deep-seated habits.[47] For, as Patricia Devine writes:

> There is strong evidence that stereotypes are well established in children's memories before children develop the cognitive ability and flexibility to question or critically evaluate the stereotype's validity or acceptability (Allport, 1954; Katz, 1976; Porter, 1971; Proshansky, 1966).[48] As a result, personal beliefs (i.e. decisions about the appropriateness of stereotypic ascriptions) are necessarily newer cognitive structures (Higgins and King, 1981). An additional consequence of this developmental sequence is that stereotypes have a longer history of activation and are therefore likely to be more accessible than are personal beliefs … Inhibiting stereotype-congruent or prejudice-like responses and intentionally replacing them with nonprejudiced responses can be likened to the breaking of a bad habit … [E]limination of a bad habit requires essentially the same steps as the formation of a habit (Devine, 1989, p. 6).[49]

In the 20 years that have elapsed since Devine wrote these words, a great deal of research has been devoted to the topic of implicit prejudice, and to the question of whether—and if so, how—automatic activation of stereotypical responses can be controlled.[50] The literature on this topic is enormous and the examples that

---

[46] Cf. Kawakami *et al.*, 2000 citing on the first Devine, Monteith, Zuwerink and Elliot, 1991; Monteith, 1993; Monteith, 1996; Monteith, Devine, and Zuwerink, 1993; Monteith, Sherman and Devine, 1998); (on the second) (Macrae, Bodenhausen, and Milne, 1997); (and on the third) (Erber and Fiske, 1984; Fiske and Neuberg, 1990; Neuberg and Fiske, 1987.) Cf. also Kawakami *et al.*, 2005.

[47] Remember Aristotle's admonition that 'It is not unimportant … to acquire one sort of habit or another, right from our youth; rather, it is very important, indeed all-important'. (NE II 2, 1103b21-25 in Aristotle, 1999.)

[48] For fascinating recent work on this issue, see Baron and Banaji, 2006.

[49] Devine continues: '… An important assumption to keep in mind in the change process, however, is that neither the formation of an attitude from beliefs nor the formation of a decision from attitudes or beliefs entails the elimination of earlier established attitudinal or stereotype representations … [A]lthough low-prejudiced persons have changed their beliefs concerning stereotyped group members, the stereotype has not been eliminated from the memory system. In fact, it remains a well-organized, frequently activated knowledge structure' (Devine, 1989, p. 15).

[50] Research in this area is so lively that it would be pointless to attempt a survey here. But for those seeking a compact overview, three helpful starting points are Stangor, 2000; Dovidio *et al.*, 2005; and Devine, 2001. For additional bibliographies on particular topics, see http://www.understandingprejudice.org/readroom/.

follow are but two of (literally) hundreds that could have been selected. I invoke them here because I think this is an area ripe for philosophical reflection, one where ancient and early modern discussions of the regulation of the passions resonate profoundly.

Three studies by Kerry Kawakami *et al.* (2000) give a flavor of one line of response that echoes Aristotle's admonition that 'we learn a craft by producing the same product that we must produced when we have learned it' (*NE* 1103 in Aristotle, 1999). In each study, 'participants were presented with two types of tasks, one involving training and the other relating to the assessment of stereotypic activation. The goal of the training task was to allow participants to practice responding 'NO' to stereotypic traits following category representations and 'YES' to nonstereotypic associations'. The assumption behind this was that 'by repeatedly and consistently implementing this simple act of negating certain category–stereotype combinations while responding positively to other category–nonstereotype combinations, the presentation of the category [would] no longer automatically activate associated stereotypes'. Experimental evidence bore out this hypothesis: subjects who had undergone this training showed reduced stereotype activation—measured using two different sorts of standard psychological metrics—effects that lasted for at least 24 hours following the training (Kawamaki *et al.*, 2000).[51]

Related work by Irene Blair provides an example of a case involving mental imagery techniques akin to those suggested by Descartes:

> Prior research has shown that mental imagery increases the accessibility of the imagined event (e.g., Carroll, 1978; Gregory, Cialdini and Carpenter, 1982). By the same token, Blair *et al.* argued that counterstereotypic mental imagery ought to increase the accessibility of counterstereotypic associations, and thereby decrease automatic stereotypes. In four separate tests, the participants were asked to spend approximately 5 min creating a mental image of a (counterstereotypic) strong woman and then complete a measure of their automatic gender stereotypes. In each test, the participants who had engaged in the counterstereotypic mental imagery produced substantially weaker automatic stereotypes, compared to participants who, (a) engaged in neutral mental imagery, (b) did not engage in any imagery, (c) imagined a weak woman, (d) imagined a strong man, or (e) attempted to suppress their stereotypes during the task (Blair, 2002, p. 249, describing Blair, Ma and Lenton, 2001).

4.3 The Cost of Disharmony

What are the costs of disharmony in cases where our ideals and social reality come apart? In this final section, I offer a few sobering remarks on this matter.

---

[51] Similar effects can be seen to result from the contemplation of admired Black exemplars; cf. Dasgupta and Greenwald, 2001. For related discussion, see Pizarro and Bloom, 2003. Cf. also Rudman, Ashmore and Gary, 2001.

The Implicit Association Test (IAT) is a widely-used test in experimental social psychology. The test asks subjects to categorize a series of words and images presented on a computer screen into one of two disjunctively-specified categories, and measures the amount of time it takes for them to make these classifications. So, for example, subjects might be presented with sequence of Black and White faces and positive and negative words (e.g. 'happy' and 'harmful'), and asked to classify them either into the categories White-or-positive and Black-or-negative, or—alternatively—into the categories White-or-negative or Black-or-positive.[52] Hundreds of studies have shown that subjects—on aggregate—are faster to make classifications into the categories White-or-positive and Black-or-negative than to their converses, suggesting that the former categories are represented as more 'natural' or easily accessed.

There is some controversy about whether the relevant IAT measures anti-Black evaluative bias, or whether it merely measures the social knowledge of the cultural association between Blacks and a cluster of negative attributes (cf. Karpinski and Hilton, 2001; Olson and Fazio, 2004). But the notion of alief finesses this distinction. From the perspective of alief, it doesn't matter whether the IAT measures your degree of access to information that you endorse (as the 'evaluative bias' reading contends), or your degree of access to information that you may reject (as the 'mere social knowledge' reading contends).[53] What the IAT unquestionably reveals—as its name indicates—are *implicit associations*. And in the case in question, the social knowledge itself involves implicit associations between certain racial categories, and highly-valenced affective content.

So what about a subject who holds these associations in the form of an easily- and regularly-accessed alief, though she aims to be non-racist in her daily interactions? In what ways is her desire to regulate her actions in terms of her (non-racist) beliefs undermined by her very knowledge of the cultural categories of American race?

In an ingenious series of studies, psychologist Jennifer Richeson has demonstrated the cognitive cost of racist alief (cf. Richeson and Shelton, 2007; Trawalter and Richeson, 2006; Richeson, Trawalter and Shelton, 2005; Richeson and Shelton, 2003). In each experiment, subjects who had previously completed an IAT concerning racial attitudes interacted either with a same-race or different-race confederate. Following the interaction, subjects completed an ostensibly unrelated task—a Stroop color-naming task[54]—which is standardly used to measure executive control. Richeson reports her findings as follows:

---

[52] You can take the test yourself at: https://implicit.harvard.edu/implicit/ or http://www.understandingprejudice.org/iat/. For discussion, see Nosek, Greenwald, and Banaji, 2006.

[53] Cf. Jennifer Eberhardt (2005): '"Bias" calls forth a sense of moral condemnation in a matter that "social knowledge" does not … [P]eople may feel less urgently the need to remedy responses thought to reflect social knowledge as opposed to bias … [But] as researchers highlight the often unconscious and unintentional character of bias, they undermine the moral foundation of the dichotomy between bias and knowledge' (p. 184).

[54] On the unlikely chance that you are not familiar with this effect, see http://www.apa.org/science/stroop.html. Richeson provides a nice summary of the task in the opening paragraphs of Richeson and Shelton, 2007.

> Consistent with the prediction that interracial contact stress will undermine subsequent executive control, White individuals, on average, performed more poorly on the Stroop task after contact with a Black experimenter than they did after contact with a White experimenter. Furthermore, the greater the… relative ease with which [these subjects] associate[d] … negative words with … Black American racial categories…the poorer their Stroop performance after interracial interactions … [T]his … suggests that, like other stressors, interracial interactions can be cognitively costly (Richeson and Shelton, 2007, pp. 316-7, drawing from several paragraphs).

Subsequent neuroimaging traced the difference between the groups to differential activation of areas in prefrontal cortex associated with executive function and self-regulation. That is, subjects whose occurrent aliefs were out of line with their conscious goal of acting in a non-discriminatory fashion expended significant cognitive effort to suppress the response-tendencies activated through these associations.

This research suggests that living in a society that violates ones normative ideals has unavoidable cognitive consequences. For either you will need to deliberately restrict your attention or experiences so as not to encode certain sorts of genuine regularities[55] (for example, by deliberately preventing yourself from acquiring and attending to the fact that, in contemporary American society, certain racial categories are associated with certain sorts of highly-valenced affective content.) Or you will need to engage in alief-driven rationalization, changing your normative ideals to accord with the relevant sorts of experienced regularity (for example, by coming to endorse the legitimacy of these stereotypical associations.) Or you will experience the cognitive costs of disharmony, redeploying cognitive energy to suppress the pull of your belief-discordant aliefs (for example, by expending executive control in cases of interracial interaction to suppress your aliefs, thereby temporarily depleting your cognitive resources.) This is the trichotomy of norm-discordant alief.

Where ideals and reality come apart, reason and the passions will inevitably conflict. And the costs of this disharmony can be paid only through cognitive compromise. Such is our fate as embodied beings capable of rational reflection living in an imperfect world.

Philosophy Department
Yale University

---

[55] Cf. Kawakami *et al.*, 2002: 'It is possible that one of the reasons why people who are low in prejudice demonstrate lower levels of automatic stereotype activation associated with Blacks (Kawakami *et al.*, 1998; Lepore and Brown, 1997) is that these individuals have learned to automate through experience their explicit desires to be egalitarian (Moskowitz *et al.*, 1999)—specifically, because they have developed a strong associative link between this goal and specific target categories … [G]radually, by consistently and frequently inhibiting the activation of cultural stereotypes and possibly also concurrently developing and using new associations that are consistent with their egalitarian beliefs, their cognitive representations may actually change.'

References

Alcinous, 1993: *The Handbook of Platonism*, trans. and ed. J. Dillon. Oxford: Clarendon Press.

Alcoff, L.M. 2006: *Visible Identities: Race, Gender and the Self*. Oxford: Oxford University Press.

Allport, G. 1954: *The Nature of Prejudice*. Reading, MA: Addison Wesley.

Amodio, D., Harmon-Jones, E. and Devine, P. 2003: Individual differences in the activation and control of affective race bias as assessed by startle eyeblink response and self-report. *Journal of Personality and Social Psychology*, 84, 738–53.

Ariely, D. 2008: *Predictably Irrational: The Hidden Forces That Shape Our Decisions*. New York: Harper Collins.

Aristotle, 1999: *Nicomachean Ethics*, trans. T. Irwin. 2nd edn. Indianapolis, IN: Hackett.

Armstrong, D. 1973: *Belief, Truth and Knowledge*. Cambridge: Cambridge University Press.

Baier, A. 1991: *A Progress of Sentiments: Reflections on Hume's Treatise*. Cambridge, MA: Harvard University Press.

Baker, L. Rudder, 1995: *Explaining Attitudes: A Practical Approach to the Mind*. Cambridge: Cambridge University Press.

Barney, R., Brittain, C. and Brennan, T. (eds) Forthcoming: *Plato and the Divided Self*. Cambridge: Cambridge University Press.

Barnes, J. (ed.) 1984: *The Complete Works of Aristotle*. Princeton, NJ: Princeton University Press.

Baron, A.S. and Banaji, M.R. 2006: The development of implicit attitudes: evidence of race evaluations from ages 6 to 10 and adulthood. *Psychological Science*, 17, 53–58.

Bateson, M. 2002: Context-dependent foraging choices in risk-sensitive starlings. *Molecular Breeding*, 10, 119–29.

Bateson, M., Healy, S.D. and Hurly, T.A. 2002: Irrational choices in hummingbird foraging behaviour. *Animal Behavior*, 63, 587–96.

Bertrand, M. and Mullainathan, S. 2003: Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market. No. 9873, NBER Working Papers from National Bureau of Economic Research, Inc.

Blakemore, S.J. and Decety, J. 2001: From the perception of action to the understanding of intention. *Nature Reviews Neuroscience*, 2, 561–67.

Blair, I.V. 2002: The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review*, 6, 242–61.

Blair, I.V., Ma, J.E. and Lenton, A.P. 2001: Imagining stereotypes away: the moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, 81, 828–41.

Blum, L. 2002: *I'm Not a Racist, But…: The Moral Quandary of Race*. Ithaca, NY: Cornell University Press.

Braithwaite, R.B. 1932-1933: The nature of believing. *Proceedings of the Aristotelian Society*, 33, 129–146.

Carroll, J.S. 1978: The effect of imagining an event on expectations for the event: an interpretation in terms of the availability heuristic. *Journal of Experimental Social Psychology*, 14, 88–96.

Carruthers, P. 2006: *The Architecture of the Mind: massive modularity and the flexibility of thought*. Oxford: Oxford University Press.

Carruthers, P. manuscript: How we know our own minds: the relationship between mindreading and metacognition. At: http://www.philosophy.umd.edu/Faculty/pcarruthers/#. Retrieved 19 July 2008.

Cooper, J. ed. 1997: *Plato: Complete Works*. Indianapolis, IN: Hackett.

Cooper, J. 1999: *Reason and Emotion: Essays in Ancient Moral Psychology and Ethical Theory*. Princeton, NJ: Princeton University Press.

Currie, G. 2002: Imagination as motivation. *Proceedings of the Aristotelian Society*, 102, 201–16.

Darwin, C. 1898: *The Expression of the Emotions in Man and Animals*. New York: D. Appleton.

Dasgupta, N. and Greenwald, A.G. 2001: On the malleability of automatic attitudes: combating automatic prejudice with images of admired and disliked Individuals. *Journal of Personality and Social Psychology*, 81, 800–814.

Denes-Raj, V. and Epstein, S. 1994: Conflict between intuitive and rational processing: when people behave against their better judgment. *Journal of Personality and Social Psychology*, 66, 819–29.

Dennett, D. 1971: Intentional systems. *Journal of Philosophy*, 68, 87–106.

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

Dennett, D. 1995: Do animals have beliefs? In H. Roitblat (ed.), *Comparative Approaches to Cognitive Sciences*. Cambridge, MA: MIT Press.

Descartes, R. 1649/1989: *The Passions of the Soul*, trans. S.H. Voss. Indianapolis, IN: Hackett.

Descartes, R. 1645/1991: *The philosophical Writings of Descartes, Volume III: The Correspondence*, trans. J. Cottingham *et al*. Cambridge: Cambridge University Press.

Devine, P. 1989: Stereotypes and prejudice: their automatic and controlled components. *Attitudes and Social Cognition*, 56, 5–18.

Devine, P. (ed.) 2001: Special Section: Implicit Prejudice and Stereotyping: ?How Automatic Are They? *Journal of Personality and Social Psychology*, 81:5.

Devine, P., Monteith, M.J., Zuwerink, J.R. and Elliot, A.J. 1991: Prejudice with and without compunction. *Journal of Personality and Social Psychology*, 60, 817–30.

Devine, P., Plant, E.A., Amodio, D.M. and Harmon-Jones, E. 2002: The regulation of explicit and implicit race bias: the role of motivations to respond without prejudice. *Journal of Personality and Social Psychology*, 82. 835–48.

Dovidio, J.F., Glick, P.G. and Rudman, L. (eds) 2005: *On the Nature of Prejudice: Fifty Years After Allport*. Malden, MA: Blackwell.

Dovidio, J.F. and Gaertner, S.L. 2004: Aversive racism. In M.P. Zanna (ed.), *Advances in Experimental Social Psychology*. San Diego, CA: Academic Press, 1–52.

Eberhardt, J.L. 2005: Imaging race. *American Psychologist*, 60, 181–90.

Erber, R. and Fiske, S. 1984: Outcome dependency and attention to inconsistent information. *Journal of Personality and Social Psychology*, 47, 709–726.

Falkenstein, L. 1997: Naturalism, normativity, and skepticism in Hume's account of belief. *Hume Studies*, 23, 29–72.

Fazio, R.H., Jackson, J.R., Dunton, B.C., and Williams, C.J. 1995: Variability in automatic activation as an unobtrusive measure of racial attitudes: a bona fide pipeline? *Journal of Personality and Social Psychology*, 69, 1013–27.

Ferrari, G.R.F. 2007: The three-part soul. In G.R.F. Ferrari (ed.), *The Cambridge Companion to Plato's Republic*. Cambridge: Cambridge University Press.

Fiske, S. and Neuberg, S. 1990: A continuum of impression formation from category-based to individuating processes: influences of information and motivation on attention and interpretation. In M. Zanna (ed.), *Advances in Experimental Social Psychology*. San Diego, CA: Academic Press, Vol. 3, 1–74.

Fodor, J. 1999: A theory of content. Reprinted in W.G. Lycan (ed.), *Mind and Cognition: An Anthology*. Malden, MA: Blackwell.

Gaertner, S.L., and Dovidio, J.F. 1986: The aversive form of racism. In J.F. Dovidio and S.L. Gaertner (eds), *Prejudice, Discrimination, and Racism*. Orlando, FL: Academic Press.

Galen. 2005: *On the Doctrines of Hippocrates and Plato*, trans. and ed. P. de Lacy. 2nd edn., augmented and revised. Berlin: Akademie Verlag.

Gendler, T.S. 2003: On the relation between pretense and belief. In D. McIver Lopes and M. Kiernan (eds), *Imagination, Philosophy and the Arts*. New York: Routledge, 125–141.

Gendler, T.S. 2006: Imaginary contagion. *Metaphilosophy*, 37, 1–21.

Gendler, T.S. 2007: Philosophical thought experiments, intuitions and cognitive equilibrium. *Midwest Studies in Philosophy: Philosophy and the Empirical*, XXXI, 68–89.

Gendler, T.S. (Forthcoming): Alief and belief. *Journal of Philosophy*.

Gregory, W.L., Cialdini, R.B. and Carpenter, K.M. 1982: Self-relevant scenarios as mediators of likelihood estimates and compliance: does imagining make it so? *Journal of Personality and Social Psychology*, 43, 89–99.

Harris, P. 2000: *The Work of the Imagination*. London: Blackwell.

Hearn, T. 1970: General rules in Hume's treatise. *Journal of the History of Philosophy*, VIII, 405–422.

Heider, F. and Simmel, M. 1944: An experimental study of apparent behavior. *American Journal of Psychology*, 13, 243–59.

Hieronymi, P. 2008: Responsibility for believing. *Synthese*, 161, 357–373.

Higgins, E.T. and King, G. 1981: Accessibility of social constructs: information-processing consequences of individual and contextual variability. In N. Cantor and J.F. Kihlstrom (eds), *Personality and Social Interaction*. Hillsdale, NJ: Erlbaum, 69–121.

Hirschman A.O. 1977: *The Passions and the Interests*. Princeton, NJ: Princeton University Press.

Hume, D. 1739/1978: *A Treatise on Human Nature*, ed. L.A. Selby-Bigge. Oxford: Clarendon.

Hurly, T.A. and Oseen, M.D. 1999: Context-dependent, risk-sensitive foraging preferences in wild rufous hummingbirds. *Animal Behavior*, 58, 59–66.

Irwin, T. 1975: Aristotle on reason, desire and virtue. *The Journal of Philosophy*, 72, 567–578.

James S. 1999: *Passion and Action: The Emotions in Seventeenth Century Philosophy*. New York: Oxford.

Jenkins J., Whiting J. and Williams C. (eds) 2005: *Persons and Passions: Essays in Honor of Annette Baier*. Notre Dame, IN: University of Notre Dame Press.

Karpinski, A. and Hilton, J.L. 2001: Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology*, 81, 774–88.

Katz, P.A. 1976: The acquisition of racial attitudes in children. In P.A. Katz (ed.), *Towards the Elimination of Racism*. New York: Pergamon Press, 125–154.

Kawakami, K., Dion, K.L. and Dovidio, J.F. 1998: Racial prejudice and stereotype activation. *Personality and Social Psychology Bulletin*, 24, 407–416.

Kawakami, K., Dovidio, J.F., Moll, J., Hermsen, S. and Russin, A. 2000: Just say no (to stereotyping): effects of training in the negation of stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology*. 78, 871–888.

Kawakami, K., Dovidio, J.F. and van Kamp, S. 2005: Kicking the habit: effects of nonstereotypic association training on the application of stereotypes. *Journal of Experimental Social Psychology*, 41, 68–75.

Kelly, D. and Roedder, E. 2008: Racial cognition and the ethics of implicit bias. *Philosophy Compass*, 3, 522–540 doi:10.1111/j.1747-9991.2008.00138.x.

Kraut, R. 2001/2007: Aristotle's ethics. *Stanford Encyclopedia of Philosophy: (Spring 2008 Edition)*, E.N. Zalta (ed.), URL = < http://plato.stanford.edu/archives/spr2008/entries/aristotle-ethics/>.

Lepore, L. and Brown, R. 1997: Category and stereotype activation: is prejudice inevitable? *Journal of Personality and Social Psychology*, 72, 275–287.

Levine, M.P. and Pataki, T. (eds) 2004: *Racism in Mind*. Ithaca, NY and London: Cornell University Press.

Loeb, L.E. 2002: *Stability and Justification in Hume's Treatise*. New York: Oxford University Press.

Lorenz, H. 2006: *The Brute Within: Appetitive Desire in Plato and Aristotle*. New York: Oxford University Press.

Macrae, C., Bodenhausen, G. and Milne, A. 1997: Saying no to unwanted thoughts: the role of self-awareness in the regulation of mental life. *Journal of Personality and Social Psychology*, 74, 578–89.

Malebranche, N. 1712/1997: *The Search after Truth*, trans. and ed. T.M. Lennon and P. J. Olscamp. Cambridge: Cambridge University Press.

McDowell, J. 1998: Lecture I: Sellars on perceptual experience. *The Journal of Philosophy*, 59, 431–450.

Mead, G.H. 1938: *The Philosophy of the Act*. Chicago, IL: University of Chicago Press.

Millikan, R.G. 1995: Compare and contrast Dretske, Fodor, and Millikan on teleosemantics. In R.G. Millikan (ed.), *White Queen Psychology and Other Essays for Alice*. Cambridge, MA: MIT Press.

Milton, J. 1667/1980: *Paradise Lost*. New York: W.W. Norton.

Montaigne, M.E. 1958. *The Complete Essays of Montaigne*, trans. D. Frame. Palo Alto, CA: Stanford University Press.

Montaigne, M.E. 2003. *Apology for Raymond Sebond*, trans. R. Ariew and M. Grene. Indianapolis, IN: Hackett.

Monteith, M. 1993: Self-regulation of prejudiced responses: implications for progress in prejudice-related discrepancies. *Journal of Personality and Social Psychology*, 65, 469–85.

Monteith, M. 1996: Contemporary forms of prejudice-related conflict: in search of a nutshell. *Personality and Social Psychology Bulletin*, 22, 461–473.

Monteith, M., Devine, P. and Zuwerink, J. 1993: Self-directed versus other-directed affect as a consequence of prejudice-related discrepancies. *Journal of Personality and Social Psychology*, 64, 198–210.

Monteith, M., Sherman, J. and Devine, P. 1998: Suppression as a stereotype control strategy. *Personality and Social Psychology Review*, 1, 63–82.

Moskowitz, G.B., Gollwitzer, P.M., Wasel, W. and Schaal, B. 1999. Preconscious control of stereotype activation through chronic egalitarian goals. *Journal of Personality and Social Psychology*, 77, 167–184.

Moss, J. 2005: Shame, pleasure and the divided soul. *Oxford Studies in Ancient Philosophy*, 29, 137–70.

Moss, J. forthcoming: Appearances and Calculations: Plato's Division of the Soul. *Oxford Studies in Ancient Philosophy*.

Nagel, T. 1970: *The Possibility of Altruism*. Princeton, NJ: Princeton University Press.

Neuberg, S. and Fiske, S. 1987: Motivational influences on impression formation: outcome dependency, accuracy-driven attention, and individuating processes. *Journal of Personality and Social Psychology*, 53, 431–44.

Nichols, S. and Stich, S. 2000: A Cognitive Theory of Pretense'. *Cognition*, 74, 115–147.

Nosek, B.A., Greenwald, A.G. and Banaji, M.R. 2006: The Implicit Association Test at age 7: a methodological and conceptual review. In J.A. Bargh (ed.), *Social Psychology and the Unconscious: The Automaticity of Higher Mental Processes*. Philadelphia, PA: Psychology Press, 265–292.

Olson, M.A. and Fazio, R.H. 2004: Reducing the influence of extra-personal associations on the Implicit Association Test. *Journal of Personality and Social Psychology*, 86, 653–67.

Owens, D. 2003: Does belief have an aim? *Philosophical Studies*, 115, 283–305.

Pascal, B. 1669/2005: *Pensées*, trans. R. Ariew. Indianapolis, IN: Hackett.

Passmore, J. 1952: *Hume's Intentions*. Cambridge: Cambridge University Press.

Payne, K. 2001: Prejudice and perception: the role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, 81, 181–92.

Payne, K. 2006: Weapon bias: Split-second decisions and unintended stereotyping. *Current Directions in Psychological Science*, 15, 287–291.

Pizarro, D. and Bloom, P. 2003: The intelligence of the moral intuitions: comment on Haidt. *Psychological Review*, 110, 193–6; discussion 197.

Plato, 380BCE/1992: *The Republic*, trans. G.M.A. Grube and C.D.C. Reeve. 2nd edn. Indianapolis, IN: Hackett.

Plotinus, 1984: *The Enneads*, trans. A.H. Armstrong. Cambridge, MA: Loeb Classical Library.

Plutarch, 1976: *Moralia, XIII, Part 1. Platonic Essays*, trans. H. Cherniss. Cambridge, MA: Loeb Classical Library.

Porter, J.D.R. 1971: *Black Child, White Child: The Development of Racial Attitudes*. Cambridge, MA: Harvard University Press.

Price, A.W. 1994: *Mental Conflict* (Issues in Ancient Philosophy). London: Routledge.

Prinz, J.J. 2004: *Gut Reactions: A Perceptual Theory of Emotion*. Oxford: Oxford University Press.

Proshansky, H.M. 1966: The development of intergroup attitudes. In L.W. Hoffman and M.L Hoffman (eds), *Review of Child Development Research*. New York: Russell Sage Foundation, Vol. 2, 311–371.

Quine, W.V. and Ullian, J.S. 1978: *The Web of Belief*, 2nd edn. New York: Random House.

Reeve, C.D.C. 1988: *Philosopher-Kings: The Argument of Plato's Republic*. Princeton, NJ: Princeton University Press.

Richeson, J.A. and Shelton, J.N. 2003: When prejudice does not pay: effects of interracial contact on executive function. *Psychological Science*, 14, 287–290.

Richeson, J.A., Trawalter, S. and Shelton, J.N. 2005: African Americans' racial attitudes and the depletion of executive function after interracial interactions. *Social Cognition*, 23, 336–352

Richeson, J.A., and Shelton, J.N. 2007: Negotiating interracial interactions: costs, consequences, and possibilities. *Current Directions in Psychological Science*, 16, 316–320.

Rozin, P., Millman, L. and Nemeroff, C. 1986: Operation of the laws of sympathetic magin in disgust and other domains. *Journal of Personality and Social Psychology*, 50, 703–712.

Rozin, P. 1999: Preadaptation and the puzzles and properties of pleasure. In D. Kahneman, E. Diener and N. Schwartz (eds), *Well Being: The Foundations of Hedonic Psychology*. New York: Russell Sage, 109–133.

Rudman, L.A., Ashmore, R.D. and Gary, M.L. 2001: 'Unlearning' automatic biases: the malleability of implicit prejudice and stereotypes. *Journal of Personality and Social Psychology*, 81, 856–868.

Schmitter, A.M. 2006: 17th and 18th Century Theories of Emotions. *The Stanford Encyclopedia of Philosophy (Summer 2006 Edition)*, E.N. Zalta (ed.), URL = http://plato.stanford.edu/archives/sum2006/entries/emotions-17th18th/.

Schuck-Paim, C., Pompilio, L. and Kacelnik, A. 2004: State-dependent decisions cause apparent violations of rationality in animal choice. *PLoS Biology*, 2, 2305–2315.

Schwitzgebel, E. (manuscript): Acting Contrary to Our Professed Beliefs. At: http://www.faculty.ucr.edu/~eschwitz/. Retrieved 12 February 2008.

Shafir, S., Waite, T.A. and Smith, B.H. 2002: Context-dependent violations of rational choice in honeybees (Apis mellifera) and gray jays (Perisoreus canadensis). *Behavioral Ecology and Sociobiology*, 51, 180–87.

Skolnick (Weisberg), D. and Bloom, P. 2006: The intuitive cosmology of fictional worlds. In S. Nichols (ed.), *The Architecture of the Imagination: New Essays on Pretense, Possibility, and Fiction*. Oxford: Oxford University Press.

Smith, A.M. 2005: Responsibility for attitudes: activity and passivity in mental life. *Ethics*, 115, 236–271.

Sorabji, R. 1980: The role of intellect in Aristotle's ethics. In A.O. Rorty (ed.), *Essays on Aristotle's Ethics*. Berkeley: University of California Press.

Stalnaker, R. 1984: *Inquiry*. Cambridge, MA: MIT Press.

Stangor, C. (ed.) 2000: *Stereotypes and Prejudice: Key Readings (Key Readings in Social Psychology)*. Philadelphia, PA: Psychology Press.

Sullivan, S. 2006: *Revealing Whiteness: The Unconscious Habits of Racial Privilege*. Bloomington and Indianapolis, IN: University of Indiana Press.

Traiger, S. 2005. Reason Unhinged: Passion and Precipice from Montaigne to Hume. In Jenkins J., Whiting J. and Williams, C. (eds.), 2005, 100–115.

Trawalter, S. and Richeson, J.A. 2006: Regulatory focus and executive function after interracial interactions. *Journal of Experimental Social Psychology*, 42, 406–412.

Tversky, A. and Kahneman, D. 1973: Availability: a heuristic for judging frequency and availability. *Cognitive Psychology*, 50, 207–32.

Velleman, D. 2000: The aim of belief. In *The Possibility of Practical Reason*. New York: Oxford University Press.

Velleman, D. and Shah, N. 2005: Doxastic deliberation. *Philosophical Review*, 114, 497.

Wedgwood, R. 2002: The aim of belief. *Philosophical Perspectives*, 16, 267–297.

Williams, B. 1973: Deciding to believe. In *Problems of the Self*. New York: Cambridge University Press.

Williamson, T. 2000: *Knowledge and its Limits*. New York: Oxford University Press.

Wilson, T.D., Lindsey, S. and Schooler, T.Y. 2000: A model of dual attitudes. *Psychological Review*, 107, 101–126.